SHIKEN

Volume 24 • Number 2 • December 2020

Contents

Testing and Evaluation SIG Newsletter

ISSN 1881-5537

Shiken

Volume 24 No. 2
December 2020

The validation of an L2 English listening self-efficacy instrument using Rasch analysis

Eric Shepherd Martin
ericshepherdmartin@gmail.com
Department of Education, Junior and Senior High School English Education Course, Shitennoji University

Abstract

This paper details the development and validation of a listening self-efficacy instrument for EFL/ESL learners with beginner-to-intermediate-level English language proficiency. Self-efficacy, or the belief in one's ability to perform a task successfully, is believed to determine how likely individuals will be to cope with difficulties relating to the task domain (e.g., listening, speaking, reading, or writing), and to sustain their effort in spite of obstacles (Bandura, 1997). To date, few instruments have been developed to evaluate English L2 listening self-efficacy. The instrument presented here was distributed among a sample of first- and second-year Japanese university students (*N* = 121), and, unlike most previously developed questionnaires, was validated through the use of Rasch analysis. The results of the administration of the questionnaire showed that learners' responses differed predictably and considerably, thereby suggesting the utility of the instrument for future use by EFL/ESL practitioners.

Keywords: Rasch analysis, listening, self-efficacy, motivation, university education, assessment

The construct of self-efficacy, developed by Bandura (1977), can be defined as the degree to which people judge their capabilities to complete a specific task with the skills that they possess, and the degree to which they believe that the performance will have positive consequences. High self-efficacy can "determine whether coping behavior will be initiated, how much effort will be expended, and how long it will be sustained in the face of obstacles and aversive experience" (Bandura, 1977, p. 191). According to Bandura (1997), self-efficacy is developed via four sources. The first is through experiences of success, which are said to be the most influential source of efficacy information. Second, self-efficacy can also be developed when individuals assess their performance by comparing it to the performance of others (e.g., comparing one's test score with a peer's). Third, positive feedback and, finally, affective arousal, have also been demonstrated to influence people's sense of self-efficacy.

Other self-referent constructs, such as self-esteem and self-concept, share similarities with self-efficacy, yet self-efficacy is distinguishable from them. Although self-esteem and self-efficacy are positively correlated, self-esteem is specifically related to a person's perception of their own self-worth (e.g., "I am a good person."), while self-concept refers to people's beliefs about how well they can perform in a domain in general (e.g., "I am good at learning languages."). Self-efficacy, on the other hand, relates to how well a person believes they are capable of performing tasks in a specific domain (e.g., "I can order a pizza on the phone in English" as a task of English communication) (Wang et al., 2014).

Over the past 37 years there has been a steady increase in studies linking self-efficacy to academic achievement (Mills, 2014). In their 1996 review of motivational research (as cited in Mills, 2014), Graham and Weiner wrote that studies consistently indicated that students with high academic task self-efficacy exhibited lower levels of anxiety, greater persistence in the face of obstacles, a willingness to exert greater effort, a greater use of learning strategies, and higher levels of intrinsic academic motivation than students with low academic task self-efficacy. Studies have also linked self-efficacy with second language achievement. Domains of interest have included reading (Burrows, 2013; Leung et al., 2019), speaking (Busse & Walter, 2013), writing (Ruegg, 2014), and listening (Graham, 2007; Graham & Macaro, 2008;

Mills et al., 2006, 2007; Yan, 2012; Yang, 1999). In each domain self-efficacy universally has been found to be positively correlated with and, for EFL reading, to have a causal effect on, achievement (Burrows, 2013).

Researchers who have conducted investigations of L2 listening self-efficacy have used several instruments to do so. For example, Yan (2012) employed 16 items using an 11-point Likert scale that asked participants to rate their predicted ability to understand main points, details, the meanings of unknown words, and keywords on four kinds of listening tasks on the Chinese College Entrance Test 4 (CET4). A more general instrument was created by Mills, Pajares, and Herron (2006; 2007), who used a 14-item, eight-point Likert survey in their study of university intermediate-to-advanced L2 French learners in the United States.

However, the results of these questionnaires were validated through the use of traditional statistical methods (e.g., correlations and factor analyses). Rasch analysis offers several advantages over other traditional analytical methods, such as Cronbach's alpha reliability estimates, factor analyses, and correlation to data from other questionnaires (Apple, 2013). First, Rasch analysis can determine how difficult individual items are to agree with (endorse), whereas other measurements assume that all items are equally endorsable. This is especially useful for allowing test creators to identify items that potentially ask the same question in different wording. Second, Rasch analysis can identify misfitting people and items that might not be contributing productively to the measurement of the construct. Third, although Rasch reliability is considered akin to Cronbach's alpha reliability, Rasch analysis provides reliability estimates for both persons and items, while Cronbach's reliability estimates only show the consistency of person responses. Finally, Rasch principal components analysis (PCA) of item residuals can demonstrate the degree to which items cohere to a single construct, while other measurements cannot.

Recognizing these advantages, on at least two occasions researchers have used Rasch analysis to validate instruments that were created to measure L2 self-efficacy. Burrows (2013) used Rasch analysis to validate his *Reading Self-Efficacy Questionnaire*, which was piloted among 200 Japanese university students. Lake (2013) also created self-efficacy questionnaires to measure L2 English speaking self-efficacy (nine items), reading self-efficacy (seven items), and listening self-efficacy (10 items) among 539 all-female Japanese L2 English learners at two universities. The present study adds to the literature by providing a detailed account of the development and validation of an L2 English listening self-efficacy questionnaire. The analysis provided here is intended to guide researchers in the development of future questionnaires that investigate self-efficacy and other psychological variables related to L2 education.

Purpose of This Study

The purpose of this study was to create a Likert-type questionnaire to evaluate EFL/ESL learners' English listening self-efficacy. The three research questions were as follows:

1. Does the order of item endorsability present a coherent picture of greater and lesser levels of listening self-efficacy, as predicted by theory?
2. Do the questionnaire items fit the Rasch model sufficiently to indicate that they are measuring a coherent, unidimensional construct?
3. What task features tend to make a listening self-efficacy item more difficult to endorse?

Materials and Methods

Participants

Initially there were 121 participants in this study. The participants (*N* = 121) were education majors at a private university in western Japan. Of these participants, 36 were first-year elementary school education majors in a four-skills English class (21 males, 15 females); 46 were second-year junior high school English education majors in reading and writing English classes (27 males, 19 females); and 39 were third-year elementary school education majors in an intensive reading English class (23 females, 16 males). Their TOEIC Reading and Listening scores ranged from 300 to 600, with an average score of just over 400 points.

Prior to data analysis, questionnaires were examined for obvious patterns of irregularity (e.g., tests in which the participant circled the same number for every item). As a result, seven participants were removed, leaving data from 114 participants for analysis.

Instrument

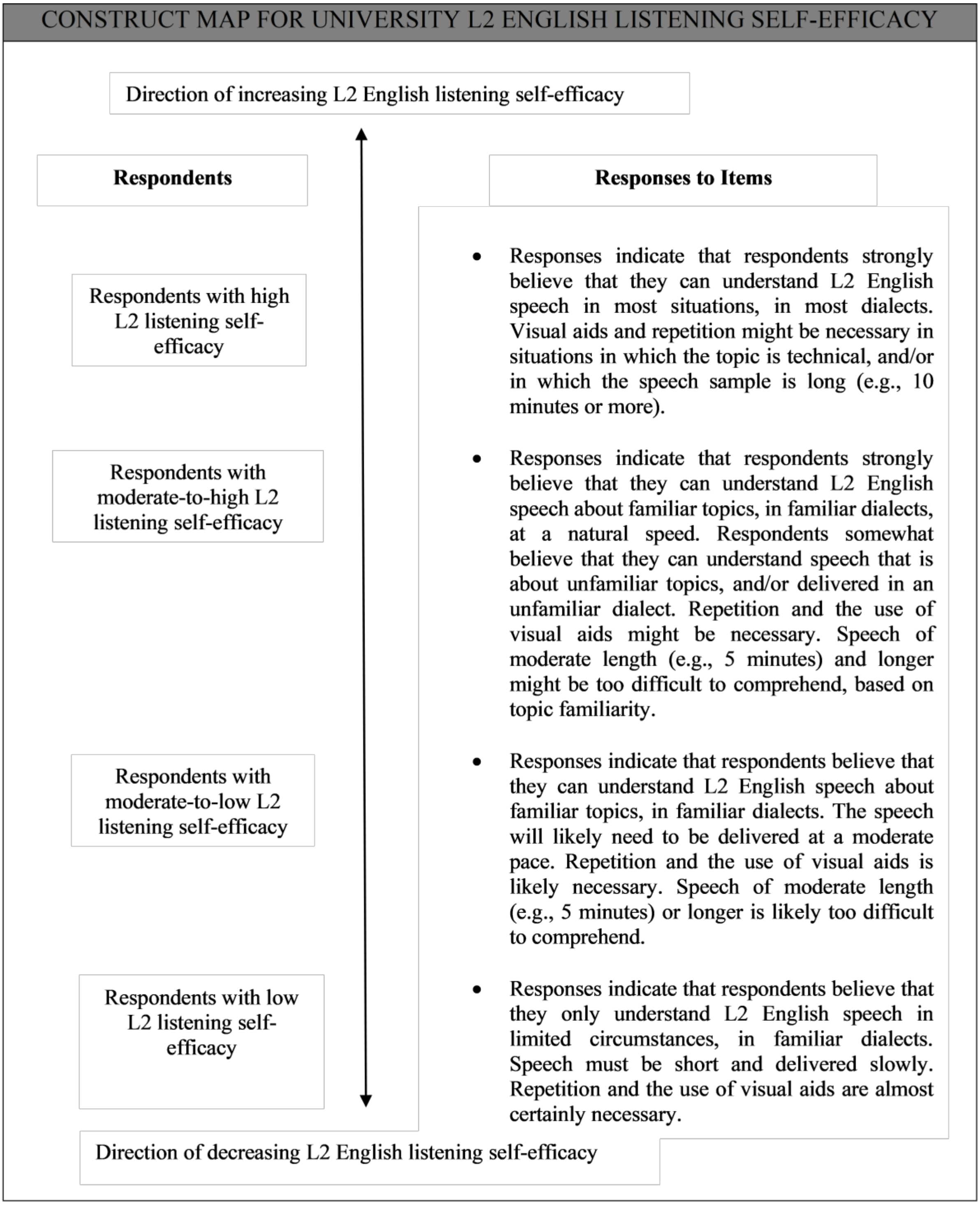
The *L2 English Listening Self-Efficacy Questionnaire* was developed as a six-point Likert-type questionnaire (see Appendices A and B). It contains 16 items that describe concrete listening scenarios. The instructions indicate that participants should imagine that English is used in each scenario, and that they should endorse their likelihood of accomplishing the task described by each item, on a scale of 1 (*I most likely cannot do it*) to 6 (*I most likely can do it*). The questionnaire was initially written in English and then translated into a Japanese version, which was answered by the participants in this study.

Items for the *L2 English Listening Self-Efficacy Questionnaire* were created based on descriptions of listening ability as described by the Common European Framework of Reference for Languages (Council of Europe, 2001), and from the American Council of the Teaching of Foreign Languages (ACTFL, 1986). The items in this instrument are worded similarly to, and contain task features similar to, the one used by Mills, Pajares, and Herron (2006, 2007), an instrument which was also based on ACTFL descriptions. These descriptions were examined and the following features were found to influence how easy or difficult a task was to endorse: (a) task familiarity (i.e., how much previous experience that a learner has had with a task), (b) topic familiarity, (c) amount of time listening, (d) the use or absence of visual aids, (e) the need to understand main points versus details, and (f) the ability to listen more than once (see Table 1). Using the instrument by Mills et al. (2006, 2007) as an example, items were created that contained variations of the identified task features, and that provided concrete descriptions of situations that were applicable to the sample group (e.g., discussions of "life in Kansai").

Table 1
Effects of task features on listening task endorsement difficulty

Feature	Easier-to-Endorse	More Difficult-to-Endorse
Task Familiarity	More familiar to the listener	Less familiar to the listener
Topic Familiarity	More familiar to the listener	Less familiar to the listener
Speaker Familiarity	More familiar speaker or dialect	Less familiar speaker or dialect
Length of Speech	Shorter	Longer
Use of Visual Aids	Greater use of visual aids	Less-or-no use of visual aids
Degree of Understanding	Listening for main points	Listening for details
Repetition	Listening more than once	Listening only once

The above features were expected to account for a large degree of the variance in CEFR and ACTFL item endorsability. However, the impact of individual features on item endorsability remained unclear. Therefore, a list of items was created and items were ordered from *most endorsable* to *least endorsable* based on the CEFR and ACTFL listening proficiency descriptors. This ordering was used in the creation of an a priori *construct map* for this questionnaire, prior to administering the questionnaire (see Figure 1). A construct map is a visual representation of the relationship between expressions of a construct (often latent, or hypothesized) and rater, item, and test-taker performance data (Wilson, 2005). Wilson wrote that a construct map must include two features: (a) a well-defined explanation of the content of the construct, and (b) evidence that an underlying continuum represents the construct, and that respondents or items should be ordered upon it. A test specifications table was also created for this questionnaire (see Table 2). It describes the guidelines for the development of the *L2 English Listening Self-Efficacy Questionnaire* and for its implementation. This table was modeled on the test specification table used for the Test of English for Academic Purposes (TEAP; Taylor, 2014).


CONSTRUCT MAP FOR UNIVERSITY L2 ENGLISH LISTENING SELF-EFFICACY
Direction of increasing L2 English listening self-efficacy
Respondents
Responses to Items
Respondents with high L2 listening self-efficacy
Respondents with moderate-to-high L2 listening self-efficacy
Respondents with moderate-to-low L2 listening self-efficacy
Respondents with low L2 listening self-efficacy
Direction of decreasing L2 English listening self-efficacy

- Responses indicate that respondents strongly believe that they can understand L2 English speech in most situations, in most dialects. Visual aids and repetition might be necessary in situations in which the topic is technical, and/or in which the speech sample is long (e.g., 10 minutes or more).
- Responses indicate that respondents strongly believe that they can understand L2 English speech about familiar topics, in familiar dialects, at a natural speed. Respondents somewhat believe that they can understand speech that is about unfamiliar topics, and/or delivered in an unfamiliar dialect. Repetition and the use of visual aids might be necessary. Speech of moderate length (e.g., 5 minutes) and longer might be too difficult to comprehend, based on topic familiarity.
- Responses indicate that respondents believe that they can understand L2 English speech about familiar topics, in familiar dialects. The speech will likely need to be delivered at a moderate pace. Repetition and the use of visual aids is likely necessary. Speech of moderate length (e.g., 5 minutes) or longer is likely too difficult to comprehend.
- Responses indicate that respondents believe that they only understand L2 English speech in limited circumstances, in familiar dialects. Speech must be short and delivered slowly. Repetition and the use of visual aids are almost certainly necessary.

Figure 1. The construct map for L2 English self-efficacy.

Table 2
L2 English Listening Self-Efficacy Questionnaire test specifications

Construct	L2 listening self-efficacy, defined by the listeners' belief in their ability to understand the main points and/or details of L2 speech.
Theory	Self-efficacy, defined by Bandura (1997) as "beliefs in one's abilities to organize and execute the courses of action required to produce given attainments."
Purpose of this test	This test should diagnose the English L2 listening self-efficacy of Japanese EFL university learners.
Target population	Japanese university non-EFL majors with low-to-intermediate English language proficiency.
Time given	10 minutes to complete the questionnaire. Additional time can be provided if necessary.
Instructions to participants **(English)**	For each item, circle the answer that best describes how sure you are that you can understand English in each of the situations described. All of the items refer to listening in English. 1 - I very likely can't do it. 2 - I probably can't do it. 3 - Maybe I can't do it. 4 - Maybe I can do it. 5 - I probably can do it. 6 - I very likely can do it.
Instructions to participants (Japanese)	以下の項目は英語のリスニング技能に関する内容です。客項目につき、どの程度できるかを自己評価し、１～６の数字で答えてください。なお、１～６の数字については、以下の基準を参考にしてください。
Format	Likert-type questionnaire with six possible choices for each item. The choices are identical between items.
Task description	Participants respond to 16 statements written in their first language, circling the response which best reflects their beliefs.
Administration	Testing should be conducted in a quiet, spacious environment. Test takers should not be able to see the responses of other participants. The questionnaire should be printed on A4-size paper with font large enough for all test takers to read it comfortably. Participants should be asked to complete a questionnaire about their belief in their ability to understand L2 English speech. They should be told that the results will not impact their coursework grades, and that participation is not mandatory. After participants have agreed, they should receive the questionnaire. Participants should be given time to read the instructions, which can be read aloud by the administrator. The administrator should answer any questions about the purpose and procedure of the test. Once all questions have been answered, the administrator should inform participants that they have 10 minutes to complete the questionnaire.

Scale attribute	Each item should be rated on a 1-to-6 scale. To reflect the construct of self-efficacy, each scale item should be worded in degrees of "can" and "can't do" endorsement labels. An even number of choices should be provided. The most extreme choices should contain adverbs such as "very likely," rather than absolute terms such as "definitely."
Prompt attributes **(PA)**	All items should be written in Japanese. Each item should be no longer than 30 Japanese characters. Items should describe situations that reflect varying degrees of the following attributes: Task familiarity (More or less) Topic familiarity (More or less) Speaker familiarity (Classroom/Japanese speaker vs. non-Japanese speaker) Length of speech (Longer or shorter) Use of visual aids (Greater use and less-or-no use) Degree of understanding (Listening for details or for main points) Repetition (More or fewer opportunities to listen) Items should be worded positively, containing verbs such as "understand" and "comprehend," and should not include negative verbs (e.g., can't, unable to, etc.). Four items should be created to represent each of the four descriptions of deceasing-to-increasing self-efficacy levels, resulting in a total of 16 items.
Example items	Understand when a teacher asks me to stand up or sit down. Understand a recorded dialogue in English about two people going to the supermarket. Understand the main points of an English TV news broadcast about Japan. Understand the main points of an English lecture about Inuit.
Response attributes **(RA)**	Participants consider each item. They reread the statement as needed to try to connect it to their perceived level of self-efficacy. They then circle the answer which they believe best reflects their own beliefs. Ideally, reponses should include the numbers 1 through 6, which are defined in the test instructions. A high score indicates a strong agreement with the statement.
Scoring parameters	Scores can range from 16 (all items answered as "1") to 96 (all items answered as "6"). Participants who score between 0 and 23 should be rated as having "low L2 listening self-efficacy." Respondents with scores between 24 and 47 should be rated as having "moderate-to-low L2 listening self-efficacy." Respondents with scores between 48 and 71 should be rated as having "moderate-to-high L2 listening self-efficacy." Respondents with scores between 72 and 96 should be rated as having "high L2 listening self-efficacy." Unanswered items should be prorated.

The questionnaire was further developed based on self-efficacy theory and the guidelines for the development of self-efficacy-measuring instruments, as described by Bandura (2006). Notably, Bandura wrote that test makers should word items or scale descriptions in terms of *can do* statements to reflect the perceived ability, rather than *will do* statements, which measure intention. The guidelines for survey instrument construction described by Nemoto and Beglar (2014) were also adhered to. Their suggestions include the use of items that represent concrete aspects of the construct, even-numbered scales so that test takers fall either positively or negatively on the scale, and the avoidance of negatively worded items.

Analysis Procedure

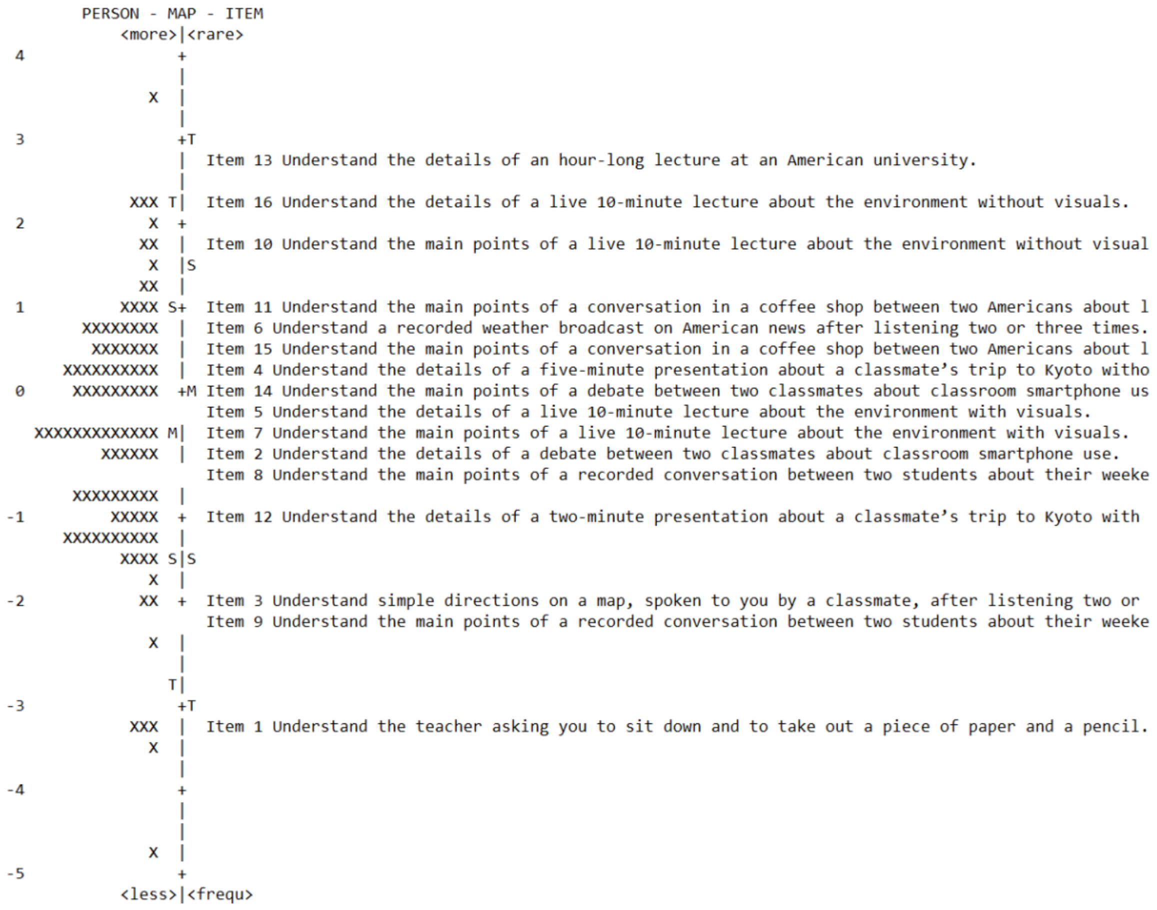
Winsteps version 3.73 (Linacre, 2011) was used to analyze the data, using the Rasch Rating Scale Model for categorical data (Andrich, 1978). The Rasch analysis consisted of person and item fit analysis, item-person (Wright) maps, and a Rasch PCA of item residuals.

Results

Wright Map

An item-person map, also called a Wright map, was created first and examined (see Figure 2). The Wright map locates persons and items side by side on a single, logit scale. A logit measure is an indication of the probability that an item will be endorsed positively by a participant, and participants are placed on the scale based on their overall level of the construct, listening self-efficacy in this case. Items are ordered according to their difficulty of endorsement, and participants opposite an item on the map are modeled to be 50% likely to endorse an item at that level. By convention, the zero point on the scale is set as the mean item difficulty. The Wright map produced for this instrument showed that Item 13 ("Understand the details of an hour-long lecture at an American university," Rasch item difficulty measure = 2.63) was the most difficult to endorse. The item easiest to endorse was Item 1 ("Understand the teacher asking you to sit down and take out a piece of paper and a pencil," Rasch item difficulty measure = -3.31). The map shows that mean person scores fell on Item 7 ("Understand the main points of a live 10-minute lecture about the environment with visuals," Rasch item difficulty measure = -.29). This indicates that Item 7 could be used to distinguish between participants with higher and lower self-efficacy.

The results closely matched the *a priori* prediction of item difficulties. In general, and as predicted, item difficulty was largely shown to be a factor of task familiarity (i.e., the more easily endorsable items were the ones that described situations that the learners had successfully engaged in). The term "American" also made items more difficult to endorse. Again, this was predictable, as it was likely that most of the participants had comparatively fewer experiences engaging in English activities with native English speakers than with their peers.



```
        PERSON - MAP - ITEM
            <more>|<rare>
 4                +
                  |
               X  |
                  |
 3                +T
                  |  Item 13 Understand the details of an hour-long lecture at an American university.
                  |
             XXX T|  Item 16 Understand the details of a live 10-minute lecture about the environment without visuals.
 2             X  +
              XX  |  Item 10 Understand the main points of a live 10-minute lecture about the environment without visual
               X  |S
              XX  |
 1          XXXX S+  Item 11 Understand the main points of a conversation in a coffee shop between two Americans about l
        XXXXXXXX  |  Item 6 Understand a recorded weather broadcast on American news after listening two or three times.
         XXXXXXX  |  Item 15 Understand the main points of a conversation in a coffee shop between two Americans about l
      XXXXXXXXXX  |  Item 4 Understand the details of a five-minute presentation about a classmate’s trip to Kyoto witho
 0     XXXXXXXXX  +M Item 14 Understand the main points of a debate between two classmates about classroom smartphone us
                     Item 5 Understand the details of a live 10-minute lecture about the environment with visuals.
   XXXXXXXXXXXXX M|  Item 7 Understand the main points of a live 10-minute lecture about the environment with visuals.
          XXXXXX  |  Item 2 Understand the details of a debate between two classmates about classroom smartphone use.
                     Item 8 Understand the main points of a recorded conversation between two students about their weeke
       XXXXXXXXX  |
-1         XXXXX  +  Item 12 Understand the details of a two-minute presentation about a classmate’s trip to Kyoto with
      XXXXXXXXXX  |
            XXXX S|S
               X  |
-2            XX  +  Item 3 Understand simple directions on a map, spoken to you by a classmate, after listening two or
                     Item 9 Understand the main points of a recorded conversation between two students about their weeke
               X  |
                  |
                 T|
-3                +T
             XXX  |  Item 1 Understand the teacher asking you to sit down and to take out a piece of paper and a pencil.
               X  |
                  |
-4                +
                  |
                  |
               X  |
-5                +
            <less>|<frequ>
```

Note: Each X equals 1 person. M = Mean; S = one standard deviation from the mean; T = two standard deviations from the mean.

Figure 2. The Wright map produced for the L2 English Listening Self-Efficacy Questionnaire.

Person Fit Analysis

Both unstandardized (*mean squares*) and standardized (*z-scores*) *infit* and *outfit* statistics for person and item fit were analyzed. According to Bond and Fox (2015), infit statistics are calculated by giving more weight to performances of persons whose responses were closer to the item’s value of endorsement difficulty (i.e., participants whose likelihood of item endorsement was similar to the item difficulty). Outfit statistics are unweighted, and are more sensitive to the scores of participants whose answers were far removed from the item difficulty. Researchers are generally advised to pay more attention to infit to determine the quality of items (Bond & Fox, 2015). From these values, a mean-square statistic of 1.0 means that there is perfect fit. Linacre (2007) recommended treating scores below 0.5 mean-squares or -2.0 *z*-scores, or above 1.5 mean-squares or 2.0 *z*-scores as misfit and investigating them further. He also wrote that any persons or items with mean-square statistics greater than 2.0 distort the measurement system and should be removed from the analysis.

To evaluate the reliability of the instrument, person reliability (used to determine how consistent person responses are) and person separation (used to estimate the instrument's ability to separate participants into different levels of the construct) were examined (Apple, 2013). The Rasch person reliability estimate of responses was estimated at .90, with a Rasch person separation value of 3.03 (see Table 3).

Table 3
Descriptive coefficients for 114 participants

	Total Score	Count	Measure	Real *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	52.90	15.90	-0.08	0.33	1.00	-0.20	1.04	-0.10
SD	11.50	0.40	1.08	0.06	0.60	1.50	0.71	1.60
Max	86.00	16.00	3.38	0.57	3.67	4.80	4.67	5.60
Min	21.00	14.00	-3.80	0.29	0.23	-3.20	0.25	-3.10

REAL *RSME* 0.34 TRUE SD 1.03 SEPARATION 3.03 PERSON RELIA. 0.90
MODEL *RSME* 0.31 TRUE SD 1.04 SEPARATION 3.37 PERSON RELIA. 0.92
SE OF PERSON MEAN = 0.10

PERSON RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (*KR-20*) PERSON RAW SCORE "TEST" RELIABILITY = .91

Note. SD = standard deviation; Max. = maximum value; Min = minimum value; RSME = square-root of the average error variance; SD = Standard deviation; RELIA. = reliability; SE = standard error.

Both infit and outfit scores were then examined. Based on fit criteria, 10 participants were found to be misfitting (resulting in a mean-square value of greater than 2.0) (see Table 4). Upon examination, consistent extreme scoring and patterning was found among these participants' responses. Furthermore, the possibility that these participants had accidently reverse-scored the items (i.e., wrongly understood "1" to mean that items were easy to endorse, and a "6" to mean that items were difficult to endorse) was ruled out. Consequently, because such responses can have an adverse impact on the construct unidimensionality and item fit measures, these participants' scores were removed from the analysis.

Table 4
Person fit statistics for the 10 most misfitting participants

Entry	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Person
16	1.32	0.40	1.75	1.90	4.67	5.60	116
66	-0.26	0.57	3.67	4.80	4.10	5.30	232
45	-1.85	0.49	1.82	1.80	3.30	3.40	209
85	-0.09	0.52	3.12	4.10	3.07	4.00	308
95	-0.52	0.49	2.69	3.50	2.73	3.50	318
99	-0.88	0.46	2.39	3.00	2.19	2.70	322
43	-0.17	0.45	2.37	3.00	2.27	2.80	207
68	1.81	0.49	2.32	2.80	1.98	2.00	234
26	-0.09	0.43	2.18	2.70	2.11	2.60	126
30	-0.35	0.43	2.11	2.60	2.07	2.50	130

Note. MNSQ = mean-squared; ZSTD = standard z-scores.

A second Rasch analysis was run on the data from the remaining 104 participants, which resulted in an estimated person reliability of .92, and a person separation value of 3.37 (see Table 5). The separation value of 3.37 indicates that the instrument could be used to reliably separate this sample into three groups, based on how willing the participants were to endorse the items.

Table 5
Descriptive coefficients for 104 participants

	Total Score	Count	Measure	Real *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	53.00	15.90	-0.25	0.35	0.98	-0.10	0.99	-0.10
SD	11.50	0.30	1.25	0.05	0.42	1.20	0.41	1.20
Max	86.00	16.00	3.45	0.58	2.00	2.40	1.94	2.30
Min	21.00	14.00	-4.65	0.32	0.28	-2.80	0.30	-2.80

REAL *RSME* 0.36	TRUE SD 1.20	SEPARATION 3.37	PERSON RELIA. 0.92
MODEL *RSME* 0.33	TRUE SD 1.21	SEPARATION 3.66	PERSON RELIA. 0.93
SE OF PERSON MEAN = 0.12			
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99			
CRONBACH ALPHA (*KR-20*) PERSON RAW SCORE "TEST" RELIABILITY = .92			

Note. SD = standard deviation; Max. = maximum value; Min = minimum value; RSME = square-root of the average error variance; SD = Standard deviation; RELIA. = reliability; SE = standard error.

Item Fit Analysis

To evaluate the reliability of the item difficulty estimates, item reliability (used to estimate the variance of item endorsement difficulty) and item separation (used to estimate how well participants were able to distinguish between items measuring different levels of the construct) values were examined. The Rasch item reliability was .99 which indicates a wide range of endorsement among the items (see Table 6).

Table 6
Descriptive coefficients for items

	Total Score	Count	Measure	Real *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	344.30	103.50	0.00	0.13	1.00	-0.10	0.99	-0.20
SD	95.60	0.70	1.53	0.02	0.20	1.40	0.20	1.30
Max	545.00	104.00	2.63	0.17	1.39	2.60	1.37	2.50
Min	188.00	102.00	-3.31	0.12	0.67	-2.70	0.64	-2.70
REAL *RSME*	0.13	TRUE SD 1.53		SEPARATION 11.35		PERSON RELIA. 0.99		
MODEL *RSME*	0.13	TRUE SD 1.53		SEPARATION 11.85		PERSON RELIA. 0.99		
SE OF ITEM MEAN = 0.40								

Note. SD = standard deviation; Max = maximum value; Min = minimum value; RSME = square-root of the average error variance; SD = Standard deviation; RELIA. = reliability; SE = Standard error.

The Rasch item separation was 11.35, which indicates that participants were able to distinguish between 11 different levels of the construct. The high reliability indicates that a very similar hierarchy of item endorsement difficulty would be obtained if the questionnaire were administered to a different, similar sample of persons. An analysis of *z*-score and mean-square values showed that there were no misfitting items.

Principal Components Analysis of Item Residuals

A Rasch principal components analysis (PCA) of item residuals was conducted for the 16 items to examine construct unidimensionality (see Table 7). Researchers have proposed different percentage thresholds for the amount of raw variance that suggests that a data set is unidimensional. However, Linacre (2018) has since suggested that the evaluation of raw variance is less important than an evaluation of unexplained variance and contrast values. According to Linacre, contrasts—clusters of survey items that produce unexplained variance and which might suggest the existence of an additional construct—that contain eigenvalues of less than 3.0 and that account for less than 10% of the variance can likely be ignored. A greater eigenvalue and variance might indicate that an additional, unwanted construct exists in the data.

Table 7
L2 English listening self-efficacy instrument standard residuals in eigenvalues

	Eigenvalue	Observed
Total raw variance in observations	54.20	100.00%
Raw variance explained by measures	38.20	70.50%
Raw variance explained by persons	12.20	22.50%
Raw variance explained by items	26.00	48.00%
Raw unexplained variance (total)	16.00	29.50%
Unexplained variance in 1st contrast	3.10	5.80%
Unexplained variance in 2nd contrast	1.90	3.50%
Unexplained variance in 3rd contrast	1.70	3.10%
Unexplained variance in 4th contrast	1.40	2.70%
Unexplained variance in 5th contrast	1.30	2.40%

Note: Values are expressed in eigenvalue units.

The table of standardized residuals for this questionnaire showed that 70.5% of the variance was explained by the person and item measures, and that all of the observed values were within .04% of the expected (model) values. This suggested that the data was a strong fit to the model of the data as produced by Winsteps (Linacre, 2011). Five contrasts were found in the data of unexplained variance. The first principal contrast accounted for 5.8% (eigenvalue 3.1) of the variance. Because the eigenvalue of this contrast was greater than 3.0, the contrast was further investigated. Standardized residual loadings for Items 16 (“Understand the details of a live 10-minute lecture about the environment without visuals”), 13 (“Understand the details of an hour-long lecture at an American university”), 11 (“Understand the main points of a conversation in a coffee shop between two Americans about life in America”), and 10 (“Understand the main points of a live 10-minute lecture about the environment without visuals”) were above .40, which can be considered high. Items 9 (“Understand the main points of a recorded conversation between two students about their weekends, after listening two or three times”), 3 (“Understand simple directions on a map, spoken to you by a classmate, after listening two or three times”), and 12 (“Understand the details of a two-minute presentation about a classmate’s trip to Kyoto with visuals”), had low loadings under -.40 (see Table 8). These items appeared to account for the high eigenvalue found in the principal contrast.

Table 8

Rasch component analysis of item residuals for the principal contrast

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
16	**0.74**	**2.32**	**1.18**	**1.12**
13	**0.69**	**2.63**	**1.20**	**1.30**
11	**0.57**	**1.12**	**1.07**	**1.05**
10	0.52	1.71	0.67	0.64
15	0.40	0.43	1.39	1.37
6	0.14	0.80	0.70	0.69
2	0.10	-0.30	1.09	1.11
9	**-0.63**	**-2.01**	**0.99**	**0.96**
3	**-0.52**	**-1.96**	**0.97**	**0.94**
12	**-0.49**	**-0.92**	**0.95**	**0.93**
8	-0.33	-0.38	0.86	0.85
1	-0.32	3.31	1.33	1.22
7	-0.29	-0.29	0.98	1.00
14	-0.23	0.10	0.86	0.85
5	-0.20	-0.03	0.83	0.85
4	-0.14	0.37	0.91	0.89

Note. Measure is in Rasch logits. Items above the dotted line were positively loading. Items below the dotted line were negatively loading. Loading values above .40 and below -.40 are labeled in bold.

Grouped together, the items with positive loadings nearly all (five out of seven) described the speaker as "American," whereas the items with negative loadings include terms such as "student," "classmate," and "teacher" on seven out of the nine items. This separation suggests that there is a contrast between items that describe listening situations involving unfamiliar native English speakers, and familiar speakers, such as students and teachers. In addition, the higher loading items described situations involving extended lectures, something also likely to be unfamiliar to the participants in this sample.

The ranking of item difficulties, as shown in the Wright map, is consistent with the *a priori* prediction that less familiar tasks will be perceived as more difficult to endorse. However, the existence of the above contrast in the residuals indicates that familiarity, in particular those tasks involving unknown native speakers, might affect perceptions of self-efficacy differentially. That is, some participants might consistently feel less daunted by encounters with unknown native speakers. Alternatively, perhaps they are better able to evaluate their ability to succeed at an unfamiliar task based on their prior experiences.

Importantly, although the eigenvalue for this contrast was just above the recommended 3.0 value, it represented less than 10% of the variance of the instrument and thus did not suggest a substantive secondary construct that would be great enough to distort the measurement of the primary construct. All other contrasts were insignificant, with eigenvalues below 3.0.

Discussion

Regarding Research Question 1, the item and person reliability and separation indices, as well as the Wright map, both suggested that the order of item endorsement difficulties produced evidence of greater and lesser levels of L2 English listening self-efficacy among the participants. The Wright map showed that the person and item means were nearly matched, and that the spread of items appeared to be a close match to the range of participants' likelihood to endorse. The item separation values suggested that the items represented listening scenarios that varied in their difficulty in a reasonably uniform manner, with at least 11 levels of difficulty identified. The person separation value (3.37) suggested that participants could be separated into three groups, based on the results of the questionnaire. These could represent groups of participants with low, medium, and high listening self-efficacy. Finally, the Wright map also showed that only one participant approached the highest score, and only five approached the lowest score. This suggested that no "ceiling" or "floor" effect existed, and that the instrument was able to measure all participants on the continuum of low-to-high listening self-efficacy, as hoped.

The second research question was whether the questionnaire items fit the Rasch model sufficiently to indicate that they are measuring a unidimensional construct. The PCA of item residuals indicated that a single, coherent construct was measured. One significant contrast was also found, between items that included "American" speakers and other, classroom-based speakers (e.g., "teacher" and "classmate"). However, that contrast, although interesting, accounted for only a small amount of unexplained variance (5.8%), and therefore arguably did not disrupt the measurement of the main construct (self-efficacy).

Finally, the third research question asked which task features were found to make items more difficult to endorse. Several factors were found to make items more or less difficult to endorse, most of which arguably reflected the amount of task familiarity, or mastery experience, that the participants had in relation to each item. The term "American" was found in five of the six most difficult-to-endorse items, and appeared to have the greatest impact on item endorsability. I used the term "American" to describe a native-English speaking stranger. If the participants also interpreted the "American" speaker to be a stranger, then these items represent situations in which participants likely had little experience (low task familiarity), and that could explain why these items were more shown to be more difficult to endorse. Broadly, length of tasks also appeared to be a determining factor, with two-minute scenarios shown to be more easily endorsable than the five-minute scenarios, and those easier to endorse than the 10-minute scenarios. Beyond that, item difficulty generally was found to increase as predicted by the by task feature table. Items that described listening for "main points" were easier to endorse than items that described listening for "details," items that described listening scenarios that included "visual aids" were also easier to endorse than those without visual aids, and items that described shorter listening tasks were easier to endorse than those with longer tasks.

Future Directions

The Rasch analysis of the questionnaire suggests that it measured a unidimensional construct, which, based on the evidence described previously, represents L2 English listening self-efficacy. However, the instrument could be improved in at least two ways. First, as previously mentioned, the term "American" appeared to have some impact on the perceived endorsement difficulty of the items in the instrument. I used the term "American" to indicate a "native English" or "non-Japanese" speaker of English. I chose this term to make the scenarios concrete for the participants, because the participants had experience listening to their class teacher, who spoke English with an American accent. However, this word arguably represents a cultural bias. Another term, such as "native English speaker" or "non-Japanese speaker of

English" would help to eliminate this potential bias, and might therefore have an impact on the degree of item difficulty.

Second, although the items appear to represent a wide range of endorsement difficulty, a wider range might be desirable in some circumstances, such as a group of participants with generally lower listening self-efficacy. In such contexts, researchers might require more items in the lower range with which to better differentiate among participants. The construct map provided previously (see Figure 1) could be used to guide the development of such items.

Third, I did not include a qualitative element to this study. In the future, a deeper insight into L2 listening self-efficacy could be gained by asking participants which item features they believed made items easier or more difficult to endorse.

Finally, the results suggested that the items produced 11 levels of endorsability. This suggested that several levels only contained one or two items. Future instruments might be made more accurate if more items are developed for the various levels.

Conclusion

Results from the validation of the instrument using Rasch analysis indicated that the instrument reliably measured several levels of the construct. This analysis adds further support to similar instruments used in previous studies of L2 self-efficacy (e.g., Burrows, 2013; Mills et al., 2006; 2007). These results, as well as the description of the theoretical basis for the creation of the test items, can hopefully be used as a basis for future investigations into L2 self-efficacy in a range of contexts and among learners of differing proficiency levels.

Acknowledgments

I would like to thank Dr. James Sick for introducing me to Rasch analysis, and for his feedback on earlier drafts of this manuscript.

References

American Council for the Teaching of Foreign Languages. (1986). *ACTFL Proficiency Guidelines*. Revised 1986. ACTFL Materials Center.

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38*, 665–680. https://doi.org/10.1177/001316447803800308

Apple, M. (2013). Using Rasch analysis to create and evaluate a measurement instrument for foreign language classroom speaking anxiety. *JALT Journal, 35*(1), 5–28. Retrieved from https://jalt-publications.org/sites/default/files/pdf/jaltjournal/jj2013a.pdf

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Cambridge University Press.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In T. Urdan & F. Pajares (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Information Age Publishing.

Bond, T. G., & Fox., C. M. (2015). *Applying the Rasch model: Fundamental measurement in human sciences* (3rd ed). Routledge.

Burrows, L. (2013). *The effects of extensive reading and reading strategies on reading self-efficacy* (Unpublished doctoral dissertation). Temple University.

Busse, V., & Walter, C. (2013). Language motivation in higher education: A longitudinal study of motivational changes and their causes. *The Modern Language Journal, 97*(2), 435–456. https://doi.org/10.1111/j.1540-4781.2013.12004.x

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Press Syndicate of the University of Cambridge.

Graham, S. (2007). Learner strategies and self-efficacy: Making the connection. *Language Learning Journal, 35*(1), 81–93. https://doi.org/10.1080/09571730701315832

Graham, S., & Macaro, E. (2008). Strategy instruction in listening for lower-intermediate learners of French. *Language Learning, 58*, 747–783. https://doi.org/10.1111/j.1467-9922.2008.00478.x

Lake, J. (2013). Positive L2 self: Linking positive psychology with L2 motivation. In M. T. Apple, D. Da Silva & T. Fellner (Eds.). *Language Learning Motivation in Japan*. (pp. 225-244). Multilingual Matters.

Leung, C.-Y., Mikami, H., & Yoshikawa, L. (2019). Positive psychology broadens readers' attentional scope during L2 reading: Evidence from eye movements. *Frontiers in Psychology, 10*, 1–12. https://doi.org/10.3389/fpsyg.2019.02245

Linacre, J. M. (2007). *A user's guide to WINSTEPS: Rasch-model computer program.*

MESA.

Linacre, J. M. (2011). Winsteps (Version 3.73). [Computer software]. Winsteps.com

Linacre, J. M. (2018, September 2). *Detecting multidimensionality in Rasch data using Winsteps Table 23* [Video]. YouTube. https://www.youtube.com/watch?v=sna19QemE50

Mills, N. (2014). Self-efficacy in second language acquisition. In S. Mercer & M. Williams (Eds.), *Multiple perspectives on the self in SLA* (pp. 6–19). Multilingual Matters.

Mills, N., Pajares, F., & Herron, C. (2006). A re-evaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency. *Foreign Language Annals, 39*(2), 276–295. https://doi.org/10.1111/j.1944-9720.2006.tb02266.x

Mills, N., Pajares, F., & Herron, C. (2007). Students: Relation to achievement and motivation. *Language Learning, 57*(3), 417–442. https://doi.org/10.1111/j.1467-9922.2007.00421.x

Nemoto, T., & Beglar, D. (2014). Developing Likert-scale questionnaires. In N. Sonda & A. Krause (Eds.), *JALT2013 Conference Proceedings*. JALT. Retrieved from https://jalt-publications.org/files/pdf-article/jalt2013_001.pdf

Rahimi, M., & Abedi, S. (2014). The relationship between listening self-efficacy and metacognitive awareness of listening strategies. *Procedia: Social and Behavior Sciences, 98*, 1454–1460. https://doi.org/10.1016/j.sbspro.2014.03.565

Ruegg, R. (2014). The effect of peer and teacher feedback on changes in EFL students' writing self-efficacy. *The Language Learning Journal, 46*(2), 87–102. https://doi.org/10.1080/09571736.2014.958190

Taylor, L. (2014). *A report on the review of test specifications for the reading and listening papers of the Test of English for Academic Purposes (TEAP) for Japanese university entrants.* Retrieved from https://www.eiken.or.jp/teap/group/pdf/teap_rlspecreview_report.pdf

Wang, C., Kim, D.-H., Bai, R., & Hu, J. (2014). Psychometric properties of a self-efficacy scale for English language learners in China. *System, 44*, 24–33. https://doi.org/10.1016/j.system.2014.01.015

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Lawrence Erlbaum Associates.

Yan, R. (2012). *Improving English listening self-efficacy of Chinese university students: Influences of learning strategy training with feedback on strategy use and performance* (Unpublished doctoral dissertation). Retrieved from https://core.ac.uk/download/pdf/6116686.pdf

Yang, N.-D. (1999). The relationship between EFL learners' beliefs and learning strategy use. *System, 27*(4), 515–535. https://doi.org/10.1016/S0346-251X(99)00048-2

Appendix A

L2 ENGLISH LISTENING SELF-EFFICACY QUESTIONNAIRE (ENGLISH VERSION)

Name: ____

Directions

Please use the following scale (1-6) to answer the questions. Choose the number that best describes how sure you are that you can perform each of the English listening tasks below. All of the items refer to listening in English.

1	2	3	4	5	6
I most likely cannot do it.	**I probably cannot do it.**	**Maybe I cannot do it.**	**Maybe I can do it.**	**I probably can do it.**	**I most likely can do it.**

1. Understand the teacher asking you to sit down and to take out a piece of paper and a pencil. **1 2 3 4 5 6**

2. Understand the details of a debate between two classmates about classroom smartphone use. **1 2 3 4 5 6**

3. Understand simple directions on a map, spoken to you by a classmate, after listening two or three times. **1 2 3 4 5 6**

4. Understand the details of a five-minute presentation about a classmate's trip to Kyoto without visuals. **1 2 3 4 5 6**

5. Understand the details of a live 10-minute lecture about the environment with visuals. **1 2 3 4 5 6**

6. Understand a recorded weather broadcast on American news after listening two or three times. **1 2 3 4 5 6**

7. Understand the main points of a live 10-minute lecture about the environment with visuals. **1 2 3 4 5 6**

8. Understand the main points of a recorded conversation between two students about their weekends, after listening once. **1 2 3 4 5 6**

9. Understand the main points of a recorded conversation between two students about their weekends, after listening two or three times. **1 2 3 4 5 6**

10. Understand the main points of a live 10-minute lecture about the environment without visuals. **1 2 3 4 5 6**

11. Understand the main points of a conversation in a coffee shop between two Americans about life in America. **1 2 3 4 5 6**

12. Understand the details of a two-minute presentation about a classmate's trip to Kyoto with visuals. **1 2 3 4 5 6**

13. Understand the details of an hour-long lecture at an American university. **1 2 3 4 5 6**

14. Understand the main points of a debate between two classmates about classroom smartphone use. **1 2 3 4 5 6**

15. Understand the main points of a conversation in a coffee shop between two Americans about life in Kansai. **1 2 3 4 5 6**

16. Understand the details of a live 10-minute lecture about the environment without visuals. **1 2 3 4 5 6**

Appendix B

L2 ENGLISH LISTENING SELF-EFFICACY QUESTIONNAIRE (JAPANESE VERSION)

Name: ______________________________

Directions

以下の項目は英語のリスニング技能に関する内容です。客項目につき、どの程度できるかを自己評価し、１～６の数字で答えてください。なお、１～６の数字については、以下の基準を参考にしてください。

1	2	3	4	5	6
非常にそう思わない	**比較的にそう思わない**	**あまり思わない**	**あまり思う**	**比較的にそう思う**	**非常にそう思う**

1. 講師が座席に座って、用紙と鉛筆を出すように指示しているのを理解できる。 **1 2 3 4 5 6**

2. ２人のクラスメートが、授業中のスマートフォンの使用に関して話し合っているディベートの内容を細部まで理解できる。 **1 2 3 4 5 6**

3. クラスメートが説明している地図上の簡単な道案内を、2・3回聴けば理解できる。 **1 2 3 4 5 6**

4. ビジュアルの資料なしでも、クラスメートの京都旅行に関する5分間のプレゼンの内容を細部まで理解できる。 **1 2 3 4 5 6**

5. ビジュアルの資料があれば、環境に関する10分間の生の講義の内容を細部理解できる。 **1 2 3 4 5 6**

6. アメリカのニュースで放送された録画されたニュースの内容を、2・3回聴けば理解できる。 **1 2 3 4 5 6**

7. ビジュアルの資料があれば、環境に関する10分間の生の講義の主な内容を理解できる。 **1 2 3 4 5 6**

8. 二人の生徒が週末について話している会話の主な内容を、1回聴けば理解できる。 **1 2 3 4 5 6**

9. 二人の生徒が週末について話している会話の主な内容を、2・3回聴けば理解できる。 **1 2 3 4 5 6**

10. ビジュアルの資料なしでも、環境に関する10分間の生の講義の主な内容を理解できる。 **1 2 3 4 5 6**

11. 二人のアメリカ人がアメリカでの生活についてカフェで話している会話の主な内容を理解できる。 **1 2 3 4 5 6**

12. ビジュアルの資料があれば、クラスメートの京都旅行に関する２分間のプレゼンの内容を細部まで理解できる。 **1 2 3 4 5 6**

13. アメリカの大学に関する１時間の講義を細部まで理解できる。 **1 2 3 4 5 6**

14. ２人のクラスメートが、授業中のスマートフォンの使用に関して話し合っているディベートの主な内容を理解できる。 **1 2 3 4 5 6**

15. 二人のアメリカ人が関西での生活についてカフェで話している会話の主な内容を理解できる。 **1 2 3 4 5 6**

16. ビジュアルの資料なしでも、環境に関する10分間の生の講義の内容を細部理解できる。 **1 2 3 4 5 6**

Proposing change in university entrance examinations: A tale of two metaphors

David Allen
allen.david@ocha.ac.jp
Ochanomizu University

Abstract

This article describes a recent education reform initiative concerning English education in Japan, specifically the proposed introduction of four-skills tests as part of the university entrance admissions process. The first aim is to summarize, in English, some of the key issues and events concerning the reform. To this end, background information and a timeline of key events since 2016 is provided. The second aim is to contrast proposals made by two academic organizations, the Japan Language Testing Association (JLTA) and the Science Council of Japan: Language and Literature Committee (SCJ). It is shown that, while agreeing on a number of specific issues related to the reform, these two organizations take starkly different positions in terms of their general orientation, which, it is argued, reflects the background of the organizational members and their views on foreign language education in Japan. These contrasting positions are discussed with reference to the metaphor, to throw the baby out with the bathwater. Finally, it is argued that a number of criticisms levelled at the proposed use of private four-skills tests illustrate a reluctance to engage with issues related to the currently used university entrance exams; in other words, these criticisms are made while ignoring the elephant in the room.

Keywords: university entrance exams, four-skills tests, MEXT, communicative language education reform

Educational systems are complex. They involve, at one level of abstraction, a *curriculum* (what is to be taught and learned), the *delivery* of the curriculum (teaching, materials, learning environments), and *assessment* of learning (tests, evaluation methods and materials; see Bunch, 2012; O'Sullivan, 2020). These three elements (curriculum, delivery, and assessment) must work together for a system to function optimally. When they do not, there is a problem.

In Japan, a key issue concerns the assessment in the English education system. The assessment comes at the end of the national course of study (NCS), which for many students functions as a stepping stone to further education[1]. The problem is that the content of the curriculum and that of the assessment do not align. The Ministry of Education, Culture, Sports, Science, and Technology (hereafter, MEXT) curricula through elementary, junior high and senior high school are designed to help students develop knowledge of the English language and the ability to *both* comprehend and produce language in *both* spoken and written modes (MEXT, 2017a, 2017b, 2018a). The final assessment, which functions as an achievement test, is the Common Examination for University Admission (*daigaku nyūgaku kyōtsū tesuto*; hereafter, the Common Test). The Common Test was introduced in 2020 with the first English test section to be administered in January 2021. Prior to the Common Test, English was assessed in the National Center Test (NCT, *sentā shiken*; see Watanabe, 2013). The Common Test has a reading and a listening section, and is essentially a continuation of the NCT. It is in multiple-choice format and the reading and listening texts include both written and spoken discourse genres. While many, including myself, see the Common Test as a significant improvement on the NCT[2], the fact remains that neither test directly assesses productive abilities in English, as they do not require the candidate to write or speak. In other words, there is a fracture in the system created by the difference in curriculum and assessment.

An additional issue is that of the university entrance exams (UEEs)[3]. These UEEs include the second-stage exams administered by national and public universities (i.e., *nijishiken*), which applicants must take in addition to the Common Test (or prior to 2020, the NCT). The UEEs also include the in-house developed entrance exams of private universities, which are typically taken by applicants in addition to, or in place of, the Common Test/NCT, though some universities require only the Common Test/NCT scores (see Kuramoto & Koizumi, 2018). Content analyses of UEEs have shown that they typically test reading ability, with a focus on grammatical and lexical knowledge, and regular use of translation tasks, especially in national/public university exams (Brown & Yamashita, 1995a, 1995b; Kikuchi, 2006; Watanabe, 1997). Given that these tests are typically perceived as high-stakes tests that are taken by candidates immediately after completion of the high school curriculum, they can impact what is taught and learned in schools (e.g., see Green, 2014) and test preparation contexts, such as cram schools (e.g., Allen, 2016b). Consequently, not only is the assessment of productive skills lacking in the Common Test, it is also lacking, either completely or in terms of balance, in the UEEs. This situation creates a conflict for high school teachers and learners: Should they follow MEXT's balanced curriculum, or should they instead focus on the knowledge and abilities needed to navigate the assessments?

In August 2016, MEXT outlined a reform in which the English section of the Common Test administered by MEXT is to be abolished in 2024, at which point universities should adopt one of a number of four-skill tests administered by private companies to assess the English ability of candidates (MEXT, 2016). During the transitional period of 2020 through 2023, universities have three options: They can utilize the common test only, a four-skills test only, or a combination of tests. This upheaval of the articulation process between pre-tertiary and tertiary education has created considerable anxiety and confusion among stakeholders. Commentators have taken to mainstream and social media, generating a storm of criticism and opinion. Consequently, in late 2019, MEXT postponed the initiative (MEXT, 2019) and convened a new committee to find a solution. This committee has been in session throughout 2020 (MEXT, 2020c) and MEXT has been soliciting opinions from a wide range of stakeholders (MEXT, 2020b).

Below is a non-exhaustive annotated timeline of some key events beginning in 2016 that are particularly relevant to this reform initiative. More extensive background and historical information on the various reforms and proposals concerning English language education in Japan can be found elsewhere (e.g., Butler & Iino, 2005; Kuramoto & Koizumi, 2018; Otsu et al., 2013; Sasaki, 2008)[4]. Importantly, the majority of the events in the timeline below were held, and books and articles published, in Japanese. Therefore, for non-Japanese speakers, it is difficult to keep up-to-date and informed on the issue. Hence, one aim of this paper is to fill this gap by presenting the views of academics in Japan who are experts in language assessment or a related field (e.g., linguistics, literature, education, educational measurement)[5]. By doing so, English-language readers can hopefully better understand not only the problems and their proposed solutions, but also the apparently conflicting views on the direction of language education in Japan.

Timeline of events related to the reform

Aug 2016: MEXT outlines proposal to introduce four-skills private tests (MEXT, 2016)

Sep 2016: Japan Language Testing Association (JLTA) Annual Conference, themed ‘Between Validity and Practicality of University Entrance Exams Based on Communication Skills: The Possibility of Reform’ http://jlta2016.sakura.ne.jp/?page_id=18

Jan 2017: JLTA (2017) proposal submitted to MEXT

Dec 2017: Abe (2017) ‘*English education in chaos…*’ is published

Mar 2018: The University of Tokyo, Center for Research and Development on Transition from Secondary to Higher Education (*Tōkyōdaigaku kōdai setsuzoku kenkyū kaihatsu sentā*) Symposium https://www.ct.u-tokyo.ac.jp/news/20180210-symposium/

Apr 2018: Koizumi (2018) ‘*How to choose and use four-skills tests …*’ is published

Jun 2018: Haebara (2018) edited collection is published

Aug 2018: The Japan Association for Language Learning and Technology (LET) Symposium, themed ‘Reconsidering four-skill assessment of foreign language proficiency’ http://www.j-let.org/let2018/page_20180222024053

Symposium at the Japan Association of College English Teachers (JACET) international convention, themed ‘Current and future assessment of four skills in an entrance examination’ http://www.jacet.org/convention/2018-2/

Mar 2019: The University of Tokyo Center for Research and Development on Transition from Secondary to Higher Education (*Tōkyōdaigaku kōdai setsuzoku kenkyū kaihatsu sentā*) Symposium (2) https://www.ct.u-tokyo.ac.jp/news/20190110-symposium2019/

Aug 2019: Symposium sponsored by the National English Education Society (JASELE, *Zenkoku eigo kyōiku gakkai shusai no shinpojiumu*), themed ‘Symposium on the significance and issues of the English four-skills assessment in the university entrance examination’ http://www.jasele.jp/symposium2019/

Sep 2019: Symposium ‘Evaluating fairness and justice of university entrance English examinations in Japan’ at JLTA Annual Conference http://jlta2016.sakura.ne.jp/?page_id=606

Oct 2019: Education Minister Koichi Hagiuda creates uproar with his gaffe about university applicants competing ‘in accordance with their standing’ (The Mainichi, 2019)

Nov 2019: Official proposal of using four-skill tests for university entrance purposes from 2020 postponed (MEXT, 2019)

Dec 2019: Committee formed to discuss future direction (MEXT, 2020c)

British Council New Directions Conference, themed ‘Realising Potential: Policy, Engagement and Impact’ https://www.britishcouncil.jp/en/new-directions/about/theme

Aug 2020: Science Council of Japan (2019) proposal is published

MEXT collects public opinions (669 submissions received in one-month period) (MEXT, 2020b)

…

Apr 2024: Planned implementation of the new system

Introducing two proposals

MEXT's proposal to use four-skills English tests as part of general route university admissions has met with resistance from many stakeholders; even those in favor have typically held some reservations about its implementation.[6] What is perhaps most interesting, however, is the emergence of a clear difference in position that can be seen by analyzing two formal proposals made by academic organizations in Japan in response to the reform. These two proposals are the primary focus of this paper.

The first is by the Japan Language Testing Association (2017, henceforth, JLTA), released in Japanese and English, and entitled *Proposal for handling English testing within the 'Prospective university entrance scholastic abilities evaluation test [provisional name]'.* The contents were submitted to MEXT as a position statement in January, 2017. The nine authors include Yoshinori Watanabe (president of JLTA), Rie Koizumi, and other well-known language assessment experts based in Japan. The position of the JLTA committee is made clear from the outset: 'While endorsing their [MEXT's] general orientation, the JLTA herewith voices our opinions on the range of issues regarding the policy based on our expertise.' (p. 1). In other words, the JLTA committee took the position that MEXT's aim of improving language education in Japan through improved alignment of the curriculum, teaching, and assessment, is a commendable one. Nevertheless, given the number of concerns raised by the JLTA committee, there are several serious problems concerning its conception and implementation. Therefore, the JLTA committee proposed a series of strategies and requirements for arriving at evidence-based solutions to the issues of test quality and implementation feasibility. Specifically, the committee recommended mobilizing language testing researchers to assist in these evaluations, the results of which should be made transparent and in accordance with state-of-the-art research and practice in the field. They also called on the test agencies to ensure certain requirements are met and evidence for meeting them is to be provided. Moreover, they suggested that individual universities conduct their own systematic studies to determine which, if any, test is appropriate for their purposes. In sum, this position agrees with the initiative to move towards more comprehensive assessment of language skills, both receptive and productive, in line with the NCS, while it also details serious issues with the innovation. In other words, *let's keep the baby, but throw out the bathwater.*

The second proposal was published by the Science Council of Japan: Language and Literature Committee (2020, henceforth, SCJ), released in Japanese and entitled '*Recommendations for English exams of university entrance exams*.' This committee consists of thirteen academics in Japan, including Takane Ito (chair) and Yoshifumi Saito (vice chair), both at the University of Tokyo. The committee is comprised of many regular commentators on the issue, including Kumiko Torikai and Yukio Otsu. In addition, Masahiko Abe and Tomokazu Haebara, both regular commentators, were present in an advisory capacity. The position taken by the SCJ committee is clearly more critical not only of the details of the specific proposal but also the general move towards assessment of four skills. Firstly, the proposal states that given the limited input available in the EFL context of Japan, learners must acquire explicit (not implicit) knowledge of the English vocabulary and grammar system, and that it is crucial to use the first language (Japanese) and the written English language in doing so. This 'basic knowledge' (*kiso chishiki*), which underlies all language use, according to the SCJ, cannot be developed through activities focused on the four skills. Secondly, it argues that given that language learning proceeds from comprehension (i.e., via receptive skills) to production, aiming for a balance of four skills is inappropriate. Therefore, learners must acquire receptive knowledge sufficiently before developing productive skills, which entails that the curriculum, and assessment, should focus primarily on the former. They argue that attempting to focus on productive skills to achieve a balanced proficiency will, in fact, impede learners' overall growth in the future. All in all, these statements reveal that the SCJ's position runs contrary to the four-skills approach adopted in the NCS, and in fact, represents what could be termed a more traditional view of language education. In addition, the SCJ committee lists numerous problems with the proposed innovation, which make it impossible in practice to assess four skills at the national level as part of the Common Test. They suggest instead that ultimately the decision, that is, whether to use a four-skills test for entrance purposes or not, should be up to each university. In sum, the SCJ committee's proposal outlines reservations concerning both the communicative approach to curriculum and assessment, suggesting that the aim of communicative reform is unnecessary, and indeed, unwanted. In other words, *throw out the baby along with the bathwater.*

Specific issues raised in the proposals

The main problems identified in MEXT's innovation are described below. Most have been taken up in some detail in books, symposia, and committee meetings noted in the above timeline, and so are only discussed briefly here. It should also be noted that many of the problems were identified relatively early and test agencies were requested to submit information that was made available online (4skills.jp), though this website is no longer in operation.

Both the JLTA and the SCJ proposals point out that MEXT's recommendation to utilize a wide range of tests (i.e., Cambridge English Assessment, EIKEN, GTEC, IELTS, TEAP, and TOEFL; a total of 23 tests when all levels are included) creates serious problems for test users in terms of determining which test is appropriate. This is because each test differs in its intended purpose, its intended target test-taker and the target proficiency. With the aim of facilitating appropriate selection and use of tests, the JLTA recommended disseminating information on these facts to ensure all test users can select and use tests appropriately.

Both proposals note the limitations of the Common European Framework of Reference (CEFR; Council of Europe, 2001) conversion table (MEXT, 2018b) for comparing scores of tests that have different purposes, content and score ranges. The SCJ also registers the concern that the CEFR levels are determined by the test companies themselves, not by third-parties; moreover, some tests have changed CEFR benchmarks, revealing instability that would lead to confusion (p. 6). The SCJ also note the can-do descriptors are not detailed criteria that can be used for assessing candidate ability in language assessments (see Weir, 2005, for early criticisms on this point) and they were not designed for the purpose MEXT is intending (i.e., to select students based on score comparisons across tests). They also criticize MEXT's claim that the CEFR is an 'international indicator' (p. 7).

Both proposals discuss the alignment (or lack thereof) of the tests with the NCS. The JLTA request that evidence be provided demonstrating the extent of alignment between each test and the NCS. Moreover, both proposals note the suitability of the four-skills tests for individual universities' needs. The JLTA proposal states that 'it is desirable that each university engages in systematic and specific studies, and selects and uses tests found to be appropriate.' It also refers to 'needs analysis in language education and test creation.' (p. 3), which is noteworthy because the needs analysis is a cornerstone of communicative language teaching syllabus design (e.g., Munby, 1978). In contrast, the SCJ proposal suggests that university exams should measure candidate abilities with respect to the curriculum in the university, its admission policy and its educational philosophy; and therefore, each university should determine whether or not to utilize a specific, or indeed any, four-skills test.

Both proposals stress the likelihood of unequal opportunities to take private four-skills tests due to economic and regional disparities in terms of location of test centers and the frequency of testing. That is, not only are the tests themselves expensive, but also candidates who live in rural areas may need to pay for travel and accommodation, further increasing their financial burden. In addition, students with disabilities may be disadvantaged, especially when it comes to speaking tests. Although MEXT has collected information on test agencies' abilities to address these concerns, they remain perhaps the most controversial aspect of the innovation.

Both proposals registered concern about the information that is made publicly available about each test, though the wording differs notably: The JLTA proposal reads, 'Testing agencies must publish the kind of detailed information that is essential for selecting tests… test's purpose and targeted proficiencies, scoring criteria and methods, as well as how the test was drafted and implemented and methods for its appropriate use.' (p. 2). The SCJ proposal states that 'the quality of questions and other issues regarding fairness are left to test companies and the actual situation is unknown,' and that 'grading criteria for speaking and writing are sometimes vague and unspecified' (p. 6).

Only the JLTA specifically refers to washback (i.e., the effect of a test on teaching and learning; see pp. 3-4). It notes 'the introduction of comprehensive four-skills testing *by itself* does not *necessarily* lead to improvements in English education at the high-school level' (italics added). It also recognizes that while appropriate test use can generate a positive effect, inappropriate use can generate a negative effect, such as narrowing of the curriculum. It notes the importance of 'engineering a positive effect of motivating test-takers to work towards improvement in four skills.' Furthermore, it recommends that to support positive washback from a test, test agencies must provide a score report and feedback that is instructive for guidance and learning. In contrast, the SCJ document does not once mention washback (*hakyū kōka*). It does state, however, that because four-skills tests differ in purpose from the NCS, they will end up replacing the current curriculum (p. 5), which appears to be an expectation of negative washback. Moreover, only the JLTA proposal stresses the importance of teacher training, which must be strengthened in terms of programs and content, 'so it may fully reflect the results of language testing research' (p. 2). Essentially, this emphasizes the importance of developing teachers' assessment literacy. The SCJ proposal does not mention teacher training directly, though states the need to improve school education and contents, and to strengthen English education at universities.

Only the SCJ registers concern that maintaining confidentiality and responding to unforeseen circumstances are overseen solely by the test agencies. Moreover, only the SCJ argues that language cannot simply be divided into four skills. Further, because neither the new companion volume to the CEFR (Council of Europe, 2018) nor the NCS does this in practice, it is not necessary to assess these skills individually, and doing so would be inconsistent with the NCS.

A general observation can be made at this point: In line with the general orientations of the two proposals, while the JLTA's concerns are accompanied by concrete suggestions for mitigating them, the SCJ's concerns are made with the aim of supporting rejection of the proposal. This is perhaps partly due to the time difference: The JLTA made its proposal within six months of MEXT's announcement, while the SCJ proposal came four years after it. Moreover, the SCJ proposal was released when MEXT had already formally postponed the plan and was holding a committee to determine the future course of action. Nevertheless, the difference in approach is also indicative of underlying theoretical positions on the nature of language learning in the Japanese context, which dictate the general orientation towards, or away from, four-skills assessment. This issue is discussed in the following section.

Discussion

Here, my intention is to point out what I see as a fundamental difference in orientation of the two proposals, and by extension the different perspectives of academics in Japan who have commented on the entrance exam issue. To this end, I refer more broadly to the literature but try to stay focused on views of authors and researchers who are connected to the proposals, primarily by their affiliation, but also, by extension, their academic background (i.e., language assessment researchers and applied linguists, or those in related disciplines, such as linguistics, education, or literature). Firstly, I overview relevant research in language assessment, then I examine the commentary of the SCJ authors. Finally, I take up the second metaphor in this article's title, the elephant in the room.

As noted above, the JLTA proposal adopts a solution-oriented, research-informed approach to the problem. A similar approach is evident in a number of recent publications by those involved in language assessment. Perhaps most notably is Koizumi (2018), which is a 263-page book entitled '*How to choose and use four-skills tests: From the viewpoint of validity*.' The aim of the book is clear from the title; however, Koizumi's aim is also to develop the reader's assessment literacy by overviewing test validity frameworks and validation studies. As noted above, this is in line with the principles of the JLTA, and other language testing organizations in Japan, such as the JALT TEVAL SIG. In another study, Baba (2019) sought to find solutions to the problem of how to generate positive washback in the high school context through the application of theory in educational psychology. He notes the lack of a solution-oriented approach in the wider discussion: 'With the exception of very few cases (e.g., Koizumi, 2018), the academic discussion has completely ignored the issues of how to reduce negative washback and whether, and how, positive washback can be generated' (p. 45). It is pertinent to note that Baba's observation is also true of the SCJ proposal, which does not once refer to washback. Most recently, Allen (2020) argued that research into the use of four-skills tests in Japan is necessary to determine, on the basis of evidence, whether specific uses of tests in specific contexts can be deemed appropriate. Concrete suggestions for future studies were provided, such as domain and needs analyses at universities, investigations into curriculum alignment, and impact studies. In sum, these studies all share a common aim, which is to resolve the issues at hand while maintaining a focus on improving the educational system through learning, teaching and assessment of all four skills. Again, they seek to *keep the baby, throw out the bathwater*.

In addition to the above studies are those that directly investigate test washback. Although some early washback studies were conducted in Japan, such as Watanabe (1996, 1997), over the last few years, there has been a marked increase in the number of studies that have sought to understand the impact that introducing four-skills tests can have in the Japanese educational context (e.g., Allen, 2016a, 2017; Allen & Nagatomo, 2019; Green, 2014; Sato, 2018, 2019; Nagatomo & Allen, 2019). One aim of these studies has been to see how a specific test influences learning and/or teaching in a specific context. Another aim is to understand the washback mechanism at a general level so that it can be better predicted in specific educational contexts. From an assessment perspective, washback and impact studies can be considered paramount to predicting and understanding the effect of tests in educational contexts.

Now, we turn to the work of a number of authors present in the SCJ committee who continue to be vocal critics of MEXT's various communication-oriented reforms. A good example of this can be found in Otsu, Erikawa, Saito and Torikai's (2013) book entitled *English education in danger*. All of the authors argue from different perspectives for the importance of the 'basic abilities' (*kiso-ryoku*) of grammar, vocabulary, and reading. For instance, Otsu (2013) makes his view clear: 'The problem with emphasizing "communication" in English education policy is the deemphasizing of grammar' (p. 65). These authors view MEXT's various reforms as threats to English education in Japan and view themselves as its protectors. These threats include the use of four-skills tests for entrance purposes, conducting English classes in English, and introducing English into primary schools. To hammer home the message, this is depicted visually on the front cover of the 2013 text as four characters (the authors) attempt to deflect these incoming attacks.

In other work, Torikai (2018, and see Fujiwara, 2018, who summarized panelists' arguments at a regional JACET symposium) asserts that reading ability is the foundation of the four skills and should be the focus of teaching in schools. Consistent with this belief, she is also highly critical of the proposal to assess four skills. Her stance is that, 'in the university entrance examination, test-takers' basic English ability should be assessed through their reading skill' (Fujiwara, 2018, p. 2) and that, 'each university can develop comprehensive English ability after admission' (Torikai, 2018, p. 142). In other words, pre-tertiary education should focus on developing 'basic English ability' (*eigo no kiso-ryoku*), which is characterized as the ability to read and knowledge of sentence structure, grammar and vocabulary, while university education should develop students' productive abilities.

A related but somewhat more aggressive argument is that of Abe (2017). He singles out a number of proponents of four-skills tests (the so-called 'Neo-four-skills group', which includes two education ministers, a cram school teacher and author, and an academic advisor for the MEXT curriculum) and argues that they are not trying to reform the curriculum and assessment system for the benefit of the students, but for financial gain. That is, they are promoting an ideology of an oral approach to language education, which includes, for instance, the recommendation to teach English in English, and by privatizing the exam system, money will inevitably flow to test companies, materials providers and cram schools. Abe's own views on language teaching are distinctly vague and impressionistic yet converge with the idea that there must be a focus on the basics: Reading and listening should be primary; writing and speaking should be left until later. Consistent with this view, he sees no need to drastically reform the exam system.

The claims of the SCJ members above are broadly in line with the key points emphasized in the SCJ position statement: Learners must acquire explicit (not implicit) knowledge of the English vocabulary and grammar system (the basics); it is crucial to use Japanese in class (not English) and to focus on the written (not oral) language; learners must acquire receptive knowledge sufficiently before developing productive skills; and focusing on productive skills will impede learners' future language development. In sum, there is a belief in the centrality of reading (i.e., the written mode, receptive ability) in pre-tertiary education, which contradicts the balanced communicative approach of the curriculum.[7] The committee thus argues against the use of balanced four-skills tests; in other words, *throw out the baby and the bathwater.*

The elephant in the room

In this section, we turn to the second metaphor in the title. When combing through the criticisms levelled at private tests and test agencies, the discerning reader will note that many are often equally applicable, if not more so, to the entrance exams that are currently in use. This inconvenient truth, or the elephant in the room, is often omitted entirely from the discussion. In other words, critiquing private four-skill tests without applying the same critique to the current exams is conveniently, but inappropriately, ignoring the major problem (the elephant!). Some of these criticisms are discussed below.

The private four-skills tests may not match the NCS (noted in both proposals). Although it is clear that some tests (e.g., TOEFL) are less aligned with the NCS guidelines than others (e.g., EIKEN), this must be demonstrated through research as recommended by the JLTA. But what about the current university entrance exams? The NCT and Common Test are based directly on the NCS and therefore should align well. However, these tests do not test productive knowledge and ability, which is, in fact, the primary reason for reform. In other words, these tests do not align with the curriculum because they only assess a subset of it. Considering the UEEs, they also suffer from the same problem of imbalance in the skills assessed. Moreover, early content analyses of these exams have shown that they diverged considerably from the NCS in terms of text complexity and vocabulary range (e.g., Brown & Yamashita, 1995a, 1995b; Hasegawa, Chujo, & Nishigaki, 2006; Kikuchi, 2006). Their reliance on word re-ordering and translation tasks likely also renders them ineffective at identifying candidates with writing abilities specified in the NCS (Moore, 2015). In sum, even Torikai has noted, 'there is also a view that individual entrance examinations at some universities are already outside the curriculum guidelines' (Fujiwara, 2018, p. 2).

Assessing four skills individually is inconsistent with the NCS, which aims to nurture integrated language abilities (SCJ, p. 4). This claim insists that the assessment of four skills individually is theoretically problematic and is indicative of a general belief that MEXT's reform proposal is fundamentally flawed and should be rejected. However, putting aside the numerous possible counter-arguments to this claim, the same question must be asked of the current exams: How theoretically consistent are these tests with the NCS? According to the preceding discussion, it is unlikely that a strong argument could be made in their support.

It is impossible to measure productive skills ... because it is difficult to guarantee fairness (SCJ, p. 5); *the quality of questions and other issues regarding fairness are left to test companies and the actual situation is unknown* (SCJ, p. 6); *grading criteria for speaking and writing assessments are sometimes vague and unspecified* (SCJ, p. 6). These criticisms concern the reliability of subjective grading and the transparency of test procedures, and again, they are likely to apply to some tests/test agencies more than others. Considering the case of the Cambridge Assessment English and IELTS, for instance, numerous empirical research papers are available online that have examined the content and procedures of these tests (see www.cambridgeassessment.org.uk and www.ielts.org), and public versions of grading descriptors are similarly available. Moreover, a glance through the accompanying reports for a relatively new test, the TEAP, reveals state-of-the-art test development procedures (see www.eiken.or.jp). In contrast, it is well-known that very little information is publicly available concerning the design and evaluation of the UEEs[8]. Many questions have been asked by researchers over 25 years ago (Brown & Yamashita, 1995c; O'Sullivan, 1995), though responses have not been forthcoming. For example, what is the intended purpose of the exam? What are the specific abilities targeted by items? What is the reliability? What grading criteria are used for scoring written responses (in English or Japanese)? How are raters trained? From a 'measurement' perspective (e.g., Brown, 1996; Kuramoto & Koizumi, 2018), these questions are legitimate and fundamental, certainly not taboo, and they should be addressed by test developers in a transparent fashion. It is true that Japan's so-called 'test culture' (Yoshida, 1996) and the 'principle of education' (Kuramoto & Koizumi, 2018) may explain some of the idiosyncrasies in approaches to testing in Japan. However, when tests are being proposed for the same test use situation, that is, for general admissions at universities in Japan, it is unacceptable to expect certain standards from one kind of test (i.e., four-skills tests) but not another (i.e., UEEs).

Introducing four-skills tests will turn secondary schools into test preparation schools (SCJ, p. 5). This criticism concerns the predicted washback of four-skills tests on school education. Implicitly, however, it assumes that the current exam system has no, or at least less, negative impact. Although determining what is negative impact is, to some extent, a matter of perspective, it is clear that the current entrance exams do have an impact on teaching and learning in schools both within the mainstream and shadow education sectors (e.g., Allen, 2016b; Green, 2014; Watanabe, 1996, 1997). Despite the fact that washback is undoubtedly a complex phenomenon, and the way teachers teach and the way students learn will never be completely dictated by a test, the content of current exams is perceived by teachers and learners to influence the content of their English teaching and learning. For example, see the self-report data of 190 university students in Allen (2017) and of 3,766 high school students and 390 teachers in Green (2014). The impact of UEEs is not a myth, or an unfounded belief, as claimed by some (e.g., Torikai, 2013, p. 97). Therefore, when critiquing the potential impact of newly introduced tests, academics must also be cognizant of the impact of those in current use.

Conclusions

There exists a problem with the Japanese educational system in terms of the English curriculum, its delivery and assessment. The proposal to introduce four-skills tests was intended to address the problem by making all aspects of the system fall into line. However, this proposal has been postponed and is now under review.

By comparing two proposals, this study has shown that while the JLTA and SCJ overlap in many of the specific criticisms of MEXT's proposal, they diverge in terms of their position on the general aim and hence their approach to the problem. The JLTA was shown to be supportive of the general aim of assessing four skills, presenting numerous concrete suggestions, including the deployment of language assessment researchers and professionals. In contrast, the SCJ was shown to be critical not only of the aim to assess four skills but also of the aim to teach four skills in a communicative, balanced approach adopted in the NCS.

The SCJ proposal appears to distance itself from the concept of test washback, perhaps because the impact of the current exams is more in line with their vision of what language study should involve (i.e., reinforcing a focus on primarily the written mode and receptive skills). Meanwhile, perhaps they are concerned about the potential impact of four-skills assessment because it is contrary to this vision. Conversely, the JLTA proposal indicates that engineering positive washback and mitigating negative washback is a core concern, in line with current approaches to language assessment, in which the consequences of using an assessment are (or should be) a test user's first concern (e.g., Koizumi, 2018).

It is these points of conflict which highlight the starkly different viewpoints of the two proposals. If academics have fundamentally different views on the curriculum, naturally they will have fundamentally different views on the assessment. In other words, if the SCJ position is taken that a balanced four-skills curriculum is undesirable, it is impossible to argue for the use of balanced four-skills tests. The alternative position, which appears to be adopted by the JLTA, is that it is not essential to change the curriculum. However, if there is to be a change in the assessment so that it better aligns with the curriculum, there needs to be even greater strengthening of the delivery, particularly regarding teacher training and teaching materials.

Unfortunately, the stark difference in opinion here is deep-rooted and reflects what Henrichsen (1989) observed: "The debate between those who advocate the 'practical' objectives of oral English study and those who defend the traditional 'cultural' purpose of language study, which is more closely allied with the nature of the examinations, is a long standing one and has not yet been resolved" (p. 178; cited in Watanabe, 2004, p. 132). This debate continues to affect both discussions of exams and the national curriculum, as demonstrated here in the two proposals.

At the time of writing, the future of the proposed innovation is still undetermined. It is unknown whether the baby will indeed be thrown out with the bathwater and whether the elephant will continue to go unacknowledged. For the sake of the millions of stakeholders in the Japanese English education system, especially the learners and the teachers, let us hope that those making the decisions will make fair and just ones, and that the English education system in Japan is remedied so that it functions as intended.

Acknowledgements

I am indebted to the numerous expert readers who have reviewed and commented upon earlier drafts of this article. Any errors or omissions remain entirely my own.

Notes

[1] Over recent decades there has been a rise in the number of admissions to university via alternative routes, such as recommendation-based, Admissions Office (AO) and special examinations; this is particularly the case for private universities, where around half of incoming students enter via these 'alternative' admissions (MEXT, 2020a). Consequently, not all students in Japan are required to take the final assessment in order to enter university and thus the discussion presented in this paper is most relevant to applicants who sit English examinations as part of their university entrance admission procedure.

[2] One example of a much-needed and now implemented revision is the removal of items that required test-takers to select the correct stressed syllable in a word (see Buck, 1989). Another is the inclusion of

listening texts that are heard only once, rather than twice, and the inclusion of more varieties of English (see Yanagawa, 2007).

[3] In this paper, the Common Test and NCT are not included within the definition of UEEs, though these tests are indeed used for university admissions. The reason for the distinction is that the Common Test/NCT is developed centrally by MEXT as an achievement test and as an entrance test for any/all universities, while the UEEs are developed by individual institutions for their specific admissions purposes.

[4] For brevity the time line begins in 2016; however, the background to the initiative stretches back many decades and includes countless relevant events that cannot be documented here. Reviewers, however, have noted that 2013 was a particularly important year as this was when the Japan Association of Corporate Executives (*keizai dōyūkai*) proposed the use of TOEFL iBT as the national UEE; the 'Kyoto Appeal' was announced (see English summary in Mizumoto, 2016; in Japanese, http://www.jasele.jp/wp-content/uploads/kyoto_appeal_2013.pdf); and Otsu et al. (2013) began their counter-assault.

[5] All quotations from Japanese-language texts are indicated with single quotation marks and are my translations.

[6] It must be noted that four-skills tests have been much more widely adopted by institutions for alternative admission routes (e.g., the current use of EIKEN, TEAP, and IELTS by universities: https://search.eiken.or.jp/qualification/exam/). The present reform seeks to expand the use of four-skills tests to general admissions.

[7] It is not the purpose of this paper to critique the rationale and empirical support for the approach to language education and assessment outlined in the SCJ proposal. Nor is the purpose to critique each of the authors' individual views, which are, in fact, quite diverse, while still converging on the principles outlined above. Nevertheless, it should be clear that the numerous claims made in the SCJ are highly controversial (for discussion on the intersection between pedagogical approaches and second language acquisition, see e.g., Ellis & Shintani, 2014). Indeed, all indications are that the position is essentially a hark back to the grammar-translation approach that was widespread in the pre-communicative era.

[8] As stated previously, the definition of UEE here does not include the Common Test and NCT. Detailed information of test specifications and validation is provided about these centrally developed tests (see Watanabe, 2013; National Centre for University Entrance Exams, https://www.dnc.ac.jp/kyotsu/index.html).

References

Abe, M. (2017). *Shijō saiaku no eigo seisaku – uso-darake no '4 ginō' kanban* [English education in chaos: Confusion and dishonesty in Japanese government policy]. Hitsuji Shobō.

Abe, M. (2018). Naze supīkingu nyūshi de, supīkingu-ryoku ga ochiru no ka. [Why will speaking ability decrease through use of a speaking entrance exam?] In T. Haebara (Ed.), *Kenshō meisō suru eigo nyūshi — supīkingu dōnyū to minkan itaku —* (pp. 69–88). Iwanami Shoten.

Allen, D. (2016a). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language Testing in Asia, 6*(7), 1–20. https://doi.org/10.1186/s40468-016-0030-z

Allen, D. (2016b). Japanese cram schools and entrance exam washback. *Asian Journal of Applied Linguistics, 3(1),* 54–67. https://www.caes.hku.hk/ajal/index.php/ajal/article/download/338/412

Allen, D. (2017). Investigating Japanese undergraduates' English language proficiency with IELTS: Predicting factors and washback. *IELTS Partnership Research Papers 2*. IELTS Partners. https://www.ielts.org/teaching-and-research/research-reports

Allen, D. (2020). *Shiken*: Past and future. *Shiken, 24*(1), 22-36. http://teval.jalt.org/sites/teval.jalt.org/files/24_01_22_Allen_Shiken_past_and_future.pdf

Allen, D. & Nagatomo, D. H. (2019). Investigating the consequential validity of TEAP: Washback to high school learners of English. *Eiken Research Report.* Eiken Foundation of Japan. https://www.eiken.or.jp/center_for_research/pdf/bulletin/vol99/vol_99_21.pdf

Baba, S. (2019). Dōsureba haisuteikusu tesuto no nozomashī hakyū kōka o mo tarasu koto ga dekiru no ka? Kyōiku shinri-gaku kenkyū kara no teian [How to produce beneficial washback effect by using high-stakes testing? Proposal from educational psychology]. *JLTA Journal, 22,* 44–64. https://doi.org/10.20622/jltajournal.22.0_44

Bunch, M. B. (2012). *Aligning curriculum, instruction, and assessment*. Measurement Inc. https://measurementinc.com/white-papers

Brown, J. D. (1996, February 5). Japanese entrance exams: A measurement problem? *The Daily Yomiuri* (Educational Supplement), 15.

Brown, J. D., & Yamashita, S. O. (1995a). English language entrance examinations at Japanese universities: What do we know about them? *JALT Journal, 17*(1), 7–30. https://jalt-publications.org/sites/default/files/pdf-article/jj-17.1-art1.pdf

Brown, J. D. & Yamashita, S. O. (1995b). English language entrance examinations at Japanese universities: 1993 and 1994. In J. D. Brown and S. O. Yamashita (Eds.) *Language Testing in Japan.* Japanese Association for Language Teaching.

Brown, J. D., & Yamashita, S. O. (1995c). The authors respond to O’Sullivan’s letter to JALT Journal: Out of criticism comes knowledge. *JALT Journal, 17*(2), 257-260. https://jalt-publications.org/sites/default/files/pdf-article/jj-17.2-art8.pdf

Buck, G. (1989). Written tests of pronunciation: Do they work? *ELT Journal, 43*(1), 50-56. https://doi.org/10.1093/elt/43.1.50

Butler, Y., & Iino, M. (2005). Current Japanese reforms in English language education: The 2003 “action plan”. *Language Policy, 4(1)*, 25-45. https://doi.org/10.1007/s10993-004-6563-5

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/16802fc1bf

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment (Companion volume with new descriptors).* https://rm.coe.int/cefr-companion -volume-with-new-descriptors-2018/1680787989

Ellis, R., & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition research.* Routledge.

Fujiwara, Y. (2018). Daigaku eigo nyūshi de nani o hakarubeki ka [What to measure in college English entrance exams]. *JACET Chūbu shibu kiyō*, *16*, 1–32. https://researchmap.jp/yasuhiro008/published_papers/18506664

Green, A. (2014). *The Test of English for Academic Purposes (TEAP) impact study: Report 1 - preliminary questionnaires to Japanese high school students and teachers*. Eiken Foundation of Japan. https://www.eiken.or.jp/teap/group/pdf/teap_washback_study.pdf

Haebara, T. (Ed.) (2018). *Kenshō — meisō suru eigo nyūshi — supīkingu dōnyū to minkan itaku —* [Inspection of stray English entrance exams – Introduction of private company speaking exams]. Iwanami Shoten.

Hasegawa, S., Chujo, K., & Nishigaki, C. (2006). Daigaku nyūshi eigomondai goi no gaido to yūyō-sei no jidai-teki henka [A chronological study of the level of difficulty and the usability of the English vocabulary used in university entrance examinations]. *JALT Journal, 28*(2), 115–134. https://doi.org/10.37546/JALTJJ28.2-1

Henrichsen, L. E. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956–1968*. Greenwood Press.

Japan Language Testing Association (JLTA). (2017). *Proposal for handling English testing within the 'Prospective university entrance scholastic abilities evaluation test [provisional name]'*. http://jlta2016.sakura.ne.jp/?page_id=865

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities after a decade. *JALT Journal, 27*(1), 77–96. https://doi.org/10.37546/JALTJJ28.1-5

Koizumi, R. (2018). *Eigo 4 ginō tesuto no erabikata to tsukaikata: Datōsei no kanten kara* [How to choose and use four-skills tests: From the viewpoint of validity]. ALC.

Kuramoto, N. & Koizumi, R. (2018). Current issues in large-scale educational assessment in Japan: Focus on national assessment of academic ability and university entrance examinations, *Assessment in Education: Principles, Policy & Practice*, *25*(4), 415-433, https://doi.org/10.1080/0969594X.2016.12256677

The Mainichi. (2019, October 28). *Japan minister apologizes for comments downplaying inequality among university test takers.* https://mainichi.jp/english/articles/20191028/p2a/00m/0na/013000c

MEXT. (2016). *Kōdai setsuzoku shisutemu kaikaku kaigi: Saishū hōkoku* [The final announcement of reports on discussions regarding the improvement of the upper secondary school-university articulation]. https://www.mext.go.jp/b_menu/shingi/chousa/shougai/033/toushin/1369233.htm

MEXT. (2017a). *Gaikokugokatsudō gaikoku-go-hen: Shōgakkō gakushū shidō yōryō (Heisei 29-nen kokuji) kaisetsu*. [Foreign language activities / foreign languages: Explanation of elementary school curriculum guidelines (2017 notification)]. https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/18/1387017_011.pdf

MEXT. (2017b). *Gaikoku-go-hen: Chūgakkō gakushū shidō yōryō (Heisei 29-nen kokuji) kaisetsu* [Foreign language edition: Explanation of junior high school curriculum guidelines (2017 notification)]. https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/18/1387018_010.pdf

MEXT. (2018a). *Gaikoku-go-hen eigo-hen: Kōtōgakkō gakushū shidō yōryō (Heisei 30-nen kokuji) kaisetsu* [Foreign languages (English): Explanation of high school curriculum guidelines (2018 Notification)].

https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/28/1407073_09_1_1.pdf

MEXT. (2018b). *Kaku shikaku kentei shiken to CEFR to no taishō-hyō* [Comparison table between each qualification / certification test and CEFR]. https://www.mext.go.jp/b_menu/houdou/30/03/__icsFiles/afieldfile/2019/01/15/1402610_1.pdf

MEXT. (2019). *Reiwa 3-nendo daigaku nyūgaku-sha senbatsu ni kakaru daigaku nyūshi eigo seiseki teikyō shisutemu un'ei taikō no haishi ni tsuite (tsūchi)* [Regarding the abolition of the university entrance examination English grade provision system management outline for the selection of university enrollees in the 3rd year of Reiwa (Notice)] https://www.mext.go.jp/a_menu/koutou/koudai/detail/1420229.htm

MEXT. (2020a). *Daigaku nyūgaku-sha senbatsu kanren kiso shiryōshū* [Collection of basic materials related to selection of university enrollees]. https://www.mext.go.jp/content/20200318-mxt_daigakuc02-000005103_8.pdf

MEXT. (2020b). *Daigaku nyūshi ni kansuru web iken boshū ni tsuite* [About soliciting web opinions about university entrance exams]. https://www.mext.go.jp/content/20200929-mxt_daigakuc02-000009870_3.pdf

MEXT. (2020c). *Daigaku nyūshi no arikata ni kansuru kentōkai* [Examination meeting regarding the ideal way of university entrance examination]. https://www.mext.go.jp/b_menu/shingi/chousa/koutou/103/index.htm

Mizumoto, A. (2016). Introducing Kyoto Appeal: Issues in and implications of using four-skills proficiency tests as entrance examinations in Japan. *British Council new directions in language assessment: JASELE journal special edition*, 59-68. http://hdl.handle.net/10112/13018

Moore, Y. (2015). An evaluation of English writing assessment in Japanese university entrance examinations. *Writing & Pedagogy, 7*(2), 233–260. https://doi.org/10.1558/wap.v7i2-3.26227

Munby, J. (1978). *Communicative syllabus design.* Cambridge.

Nagatomo, D. H. & Allen, D. (2019). Investing in their futures: Highly-motivated students' perceptions of TEAP and university entrance exam. *The Language Teacher*, *45*(5), 3-7. https://jalt-publications.org/sites/default/files/pdf-article/43.5-tlt-art1.pdf

O'Sullivan, B. (1995). A reaction to Brown and Yamashita "English language entrance exams at Japanese universities: What do we know about them?" *JALT Journal, 17*(2), 255-257. https://jalt-publications.org/sites/default/files/pdf-article/jj-17.2-art7.pdf

O'Sullivan, B. (2020). *The Comprehensive Learning System.* British Council White Papers on English Language Policy & Education. British Council.

Otsu, Y. (2013). *Eigo kyōiku seisaku wa naze machigau no ka: Ninchi kagaku gakushū kagaku no shiten kara* [Why is the English education policy wrong? From the perspective of cognitive science and learning science]. In Y. Otsu, H. Erikawa, Y. Saito, & K. Torikai (Eds.), *Eigo kyōiku, semari kuru hatan* (pp. 51-72). Hitsuji Shobō.

Otsu, Y., Erikawa, H., Saito, Y., & Torikai, K. (Eds.) (2013). *Eigo kyōiku semari kuru hatan* [*English education in danger*]. Hitsuji Shobō.

Sasaki, M. (2008). The 150-year history of English language assessment in Japanese education. *Language Testing, 25*(1), 63–83. https://doi.org/10.1177/0265532207083745

Sato, T. (2018). The impact of the Test of English for Academic Purposes (TEAP) on Japanese students' English learning. *JACET Journal, 62*, 89–107. https://doi.org/10.32234/jacetjournal.62.0_89

Sato, T. (2019). An investigation of factors involved in Japanese students' English learning behavior during test preparation. *Papers in Language Testing and Assessment, 8*(1), 69-95. http://www.altaanz.org/uploads/5/9/0/8/5908292/8_1_s4_sato.pdf

Science Council of Japan, Language and Literature Committee (2020). *Daigaku nyūshi ni okeru eigo shiken no arikata ni tsuite no teigen* [Recommendations for the English test for university entrance exam]. http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-24-t292-6.pdf

Torikai, K. (2013). *Eigo komyunikēshon nōryoku wa hakareru ka* [Can English communication skills be measured?] In Y. Otsu et al. (Eds.), *Eigo kyōiku, semari kuru hatan* (pp. 83-116). Hitsuji Shobō.

Torikai, K. (2018). *Eigo kyōiku no kiki* [Crisis of English education]. Chikuma Shobō.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*(3), 318-333. https://doi.org/10.1177/026553229601300306

Watanabe, Y. (1997). *The washback effects of the Japanese university entrance examinations of English - classroom-based research* [Unpublished doctoral dissertation]. Lancaster University.

Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Context and method in washback research: The influence of language testing on teaching and learning*, (pp. 129–146). Lawrence Erlbaum.

Watanabe, Y. (2013). The National Center Test for University Admissions. *Language Testing, 30*, 565–573. https://doi.org/10.1177/0265532213483095

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 2*(3), 281–300. https://doi.org/10.1191/0265532205lt309oa

Yanagawa, K. (2012). *A partial validation of the contextual validity of the Centre listening test in Japan* [Unpublished doctoral dissertation]. University of Bedfordshire. https://uobrep.openrepository.com/handle/10547/267493

Yoshida, K. (1996, January 15). Language testing in Japan: A cultural problem? *The Daily Yomiuri* (Educational Supplement), 15.

An investigation into the use of Rasch analysis to aid L2 writers in anonymized peer-assisted learning

Jeffrey Martin
jeffmjp@gmail.com
J. F. Oberlin University

Abstract

This study critically evaluated an anonymized peer feedback and assessment design for L2 writers enhanced by the use of Rasch analysis. This approach centered on the acts of giving assessment (Topping, 1998) and feedback (Lundstrom & Baker, 2009) in exchange with multiple peers. Each participant received feedback comments and class-wide statistical measures summarized for students without the need for their peers to rate all papers. Anonymity was maintained to bring unencumbered attention to the role of the reader (Booth et al., 2008) and to provide a space for interpretation and reflection on the potentially contrasting data and experiences that emerge. This process is argued to drive cognitive development and improve L2 writing skills. An initial trial with 15 high-proficiency EFL learners indicated that the design facilitated an effective exchange for each participant. The effects of anonymously including teacher comments also brought informative insights about the perception of feedback and its sources. Issues were found regarding overly narrow use of the ratings scales by some participants. A 6-point rating scale is proposed for more differentiated scoring. Overall, positive engagement and reception by the participants suggests that this peer assisted learning approach holds promise for L2 writers.

Keywords: peer feedback, peer assessment, L2 writing, anonymous feedback/assessment, Rasch analysis, judging plan

Peer feedback and peer assessment activities invite L2 student writers to develop their ability to better self-evaluate their work and practice the skill of discernment when reading the work of others. Giving feedback (Lundstrom & Baker, 2009) and *learning by assessing* (Topping, 1998, p. 254) may be critical drivers of these developments. Additionally, Booth et al. (2008) argued that effective writers embody the role of their "distant readers" in the world (p. 280). Recognizing the variety of possible reader responses, this distance can be simulated in a learning environment through varied, anonymous exchange. Through the giving and receiving of comments and ratings by classroom peers, potentially contrasting information becomes available to the L2 writing student, an opportunity not always forthcoming in real-world communication. Fundamentally, the student-centered and formative benefits of peer feedback and peer assessment activities have been theorized and empirically tested in SLA research (Hyland & Hyland, 2006; Liu & Hansen, 2002; Yu & Lee, 2016). While there are distinctions between feedback and assessment (Falchikov, 2001), these approaches are viewed here as complimentary. To achieve the different aims proposed by Topping (1998) and Booth et al. (2008), the current study investigated the potential of incorporating anonymity and Rasch analysis within a judging plan (Linacre, 1989; Rasch, 1961). These were used to capture the characteristics described above and place them within a peer feedback and assessment process that can be feasible for both L2 student writers and instructors.

Rasch analysis is a statistical tool that provides class-wide and individual measures calibrated for both writing ability and rating behavior, all of which can be provided back to students anonymously together with qualitative feedback. The output from Rasch analysis is achieved with relatively limited rater data from each student and the instructor prepares this using Rasch analysis software. To further emphasize the student-centered environment, teacher feedback and ratings can be intermixed anonymously within the data. Such conditions provide opportunities for L2 learners to engage with the levels of Bloom's taxonomy (Bloom, 1956; Krathwohl, 2002) in a number of ways as they make judgements at several stages along the peer-assisted learning process. Importantly, outcomes of the Rasch analysis invite students to individually revisit the papers they previously reviewed and self-assess their judgements. This paper critically evaluated an application of this approach in a semester-long academic course on business communication and leadership with 15 high-proficiency learners of English.

Literature review

Peer feedback and peer assessment

Peer feedback and peer assessment are associated terms that are viewed differently by researchers working under distinct theoretical frameworks. Research on peer-exchange often cites Vygotsky's zone of proximal development (ZPD, 1978) while other influential work builds on the cognitive developmental theory of Piaget (1971). The differences between these theories underpin the separation of peer feedback and peer assessment as applied by some in the field. This paper maintains that these activities are complementary and that the distinction between "giver" and "receiver" in peer-exchange may offer informative insights for both.

Differing views on peer feedback and peer assessment stem from different views on the learning process (Falchikov, 2001). Liu and Carless (2006) asserted that "peer feedback is primarily about rich detailed comments but without formal grades, whilst peer assessment denotes grading (irrespective of whether comments are also included)" (p. 280). Liu and Hansen (2002) outlined a Vygotskian perspective on peer feedback by defining it as a form of *peer response* where students take on reciprocal "roles and responsibilities normally taken on by a formally trained teacher, tutor, or editor" (p. 1). The resulting social interaction is what builds an individual's "current competence through the guidance of a more experienced individual" (p. 5). In this view, L2 student writers provide each other feedback and form a mutual ZPD within a sociocultural framework (Guerrero & Villamil, 1994). Student peers become "individually novices and collectively experts" for each other (Donato, 1994, p. 46), which allows for collective scaffolding that benefits the individuals within the group.

Topping and Ehly (1998) placed peer feedback and peer assessment together within a wider paradigm of *peer-assisted learning*, with Topping (1998) defining peer assessment as an arrangement for peers of equal status who consider the level, value, or worth of each other's products or outcomes of learning. In this view, peer-assisted learning finds grounding in the theories of Vygotsky (1978) but centers more on cognitive developmental theory (Piaget, 1971). The emphasis of the guiding expert(s) within a ZPD is lessened in Topping's paradigm. Rather, Topping argues that each learner is invited to reconcile differences between prior and new experiences through interactions with peers of relatively equal ability but with varied competencies. Peer assessment together with peer feedback can facilitate this symmetrical and reciprocal process where students may learn as much or more through the experience of engaging in the work of others than from the receiving of guidance. Foot and Howe (1998, p. 33) highlighted the Piagetian model of *cognitive conflict* as aptly describing this mechanism for peer-assisted learning. Topping (1998) additionally noted that the process emphasizes assessment to be a formative way to "maximize success rather than merely determine success or failure only after the event" (p. 249). In this Piagetian sense (1971), the learner enters a process of adaptation by reciprocally engaging with peers and their work. The cognitive conflicts that likely occur are points for the learner's cognitive development.

Anonymity and the role of the reader

As Zamel (1982) emphasized, the process of peer-exchange can develop in students "the crucial ability of re-viewing their writing with the eyes of another" (p. 206). Yu and Lee's (2016) overview of many studies, mostly featuring openly paired dyads, outlined how peer feedback activities benefit writing development. However, the social distractions of open peer-review can complicate a learner's perspectives on the sources of feedback. Intermixing peer feedback and teacher feedback, Xu and Liu

(2010) found that students took up anonymized feedback from both peers and teachers at similar rates. This is in contrast to the preference of some students who only value teacher feedback. Studies have also shown that anonymity can further influence the quality of peer-exchange. Rotsaert et al. (2018) found that the quality of peer feedback improved when it was given anonymously, suggesting that familiarity with the writer can inhibit genuine responses by peers. Affective factors were also found by Cheng and Tsai (2012) by showing how learners preferred anonymity to avoid social pressures. Additionally, it was essentially an anonymized peer review methodology in Lundstrom and Baker (2009) that demonstrated that the giving of feedback is potentially more effective for improving student writing than receiving it.

An important aspect for effective writing is the student's conceptualization of unknown readers outside of a classroom setting. This holds true for any writing domain. For example, Booth et al. (2008) emphasized that researchers and professionals write to achieve objectives with their readers, who are distant in the sense that they are often not known personally to the writer. Likewise in everyday business writing, Garner (2012) argued that writers more likely succeed by anticipating the "goals and priorities" of their readers (p. 7). This suggests that objectives and strategies vary widely by the style, purpose, and domain of writing, but the importance of the reader is constant. Logistically, the anonymous exchange in a writing course can be facilitated via digital file sharing techniques. Such exchange allows for data collection from many "readers" in a class, but the output of the students' rating scores would need to be calibrated for rater effects. To present scores adjusted for rater severity, the instructor can run a Rasch analysis on the collected data. This also formulates a summarized class-wide result for comparison. Accompanied by the feedback data, this information can then be shared with each student writer to create further opportunities for learning as an audience of readers. Sharing of class-wide scoring results can be done anonymously by using a coding system.

Rasch analysis

Rasch analysis is a psychometric tool widely used by researchers in SLA to measure constructs such as performance and motivation. It is also used in reliability assurance of language proficiency testing. The Rasch model (Rasch, 1961) and many-facet Rasch measurement (Linacre, 1989) allow for the facets of a construct to be simultaneously calculated and calibrated along a common interval scale, probabilistically accounting for the variance in human judgments. For example, when a student rates a peer's essay on a set of criteria, the thresholds between points on the rating scale, in each criteria, may change in meaning between criteria and between members of the rater group (Eckes, 2015). A score of eight on one criterion by a rater may be of a different severity than an eight by the same rater on another criteria. The Rasch model predicts this variability to a degree and estimates a *fair score* in comparison to an averaged rater severity (see Linacre, 2020b, p. 130). In the present design, the facets are writing performance, rater severity, and criteria.

An advantage of Rasch analysis is that values for 'rater severity' and 'good writing' can be calculated even if not all raters assess all essays, so long as the structure, or judging plan, underlining the peer exchange ensures sufficient interconnection between raters and essays. The design of the judging plan and the consistency of ratings highly affect the accuracy of overall measurement. Research has shown that resources applied to rater training can improve aspects of rating consistency, but that the consistency between raters is difficult to fully attain (Farrokhi et al., 2012; Weigle, 1998). It is an important aim of a Rasch analysis to achieve an accuracy of scoring appropriate to its purposes. The related fit statistics and the judging plan design are addressed further in the next section. Applying Rasch analysis within a peer-assisted learning process can place L2 writers in a position to

- assess and give feedback on a wider selection of essays written by peers,
- interpret fair scores from Rasch in conjunction with qualitative feedback comments,

- self-assess their own rater consistency and revisit their assessments of essays, and
- evaluate and reflect on the experiences of writing, reading, and reviewing.

The resulting cognitive conflict experienced and the assimilation of possibly conflicting information is thought to encourage learning.

The present study

This exploratory study critically evaluated an application of this approach in a semester-long academic course with high-proficiency L2 learners of English. The aim of the approach is to benefit student learning. To assess its success, the study seeks to answer the following questions:

1. Will this peer feedback and assessment design reliably meet its theoretical aims?
2. How will the participants perceive and engage in this process featuring anonymity and multiple roles taken at different stages?
3. Based on these outcomes, how might the design be improved upon for future applications?

Method

Participants

A class of 15 L2 speakers of English, all in their second year of university, participated in the 10-week process. The participants' proficiency level was CEFR B2 and above based on the general requirement of having a TOEIC score of 750 points to enter the course. Their proficiency was adequate for the set course book: *HBR's 10 Must Reads on Leadership* (Harvard Business Review, 2011), a book containing a selection of pragmatic and readable articles from thought leaders in the world of business. Two thirds of the participants were international participants coming from a variety of countries in Asia. Of the Japanese participants, all had educational experiences of at least 2 years outside of Japan.

Materials

Judging plan of an incomplete block design

Each participant rated and gave written feedback to essays assigned according to a judging plan. The judging plan of an incomplete block design allows for a peer-exchange procedure to be economically executed without each participant needing to rate every paper. 'Incomplete' means that each participant's essay is rated by varied subsets of classmates within the group (Eckes, 2015). As long as the raters and samples within the block sufficiently overlap, Rasch analysis will be able to produce group-wide measures (Linacre & Wright, 2002). Without this technique, group measures would require each rater to rate all essays. Such rating would be laborious and could raise concerns over feasibility and rater fatigue. Table 1 demonstrates this particular incomplete block design, which uses a repeating pattern to assign connections between participants and papers within the data set. Additionally, the exemplar paper was assigned to all raters and the instructor was assigned to rate each participant's work. Feedback comments and ratings were collected accordingly.

Table 1

Demonstration of Judging Plan

Student paper code	Student rater code	s-1	s-2	s-3	s-4	s-5	s-6	s-7	s-8	s-9	s-10	s-11	s-12	s-13	s-14	s-15	t-1
		r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code
s-1	p-code						7				6			8		9	7
s-2	p-code							8	7						8	7	7
s-3	p-code	8	9					7		7							7
s-4	p-code	8		8				9			9						8
s-5	p-code	7			8				7	8							8
s-6	p-code	7				7			8		9						7
s-7	p-code		9	5						6	8						7
s-8	p-code		7		9							7	9				8
s-9	p-code		7			6						4		10			8
s-10	p-code			7	10							7			9		9
s-11	p-code			8		7							7			8	7
s-12	p-code				8	4						3		9			8
s-13	p-code						3	5					7		7		6
s-14	p-code						1		5				7			8	6
s-15	p-code						3			5				8	7		6
ex-1	p-code	8	9	7	8	10	10	8	8	8	7	5	7	9	9	7	8

Note. *Incomplete block pattern (6 x 10, boxed with thick lines) repeated to create the plan. Same incomplete block design used for each criterion.*

Criteria for assessment

The participants rated on three criteria: Value to Reader, Clarity of Concept, and Language Mechanics, each on a 10-point Likert scale. The first criterion of Value to Reader is reflected in Topping's (1998) definition of peer-assessment. The categories, or levels, of each criterion were not given detailed descriptors. The simple design of three criteria was to help the participants remember the criteria and transfer learning onto their own writing. Measuring value to the reader introduced a degree of subjectivity because readers can hold different senses of value. Nonetheless, the set of criteria as a whole was meant to be straightforward to complete and to help facilitate accompanying feedback comments. The criteria and scale were discussed in an orientation session where participants were encouraged to use the full range of the scale as they deemed appropriate. The participants were not given specialized assessment training. Unpredictable ratings by participants were anticipated as becoming material which peers could later critically evaluate. Both narrow and unpredictably wide use of the rating scales beyond a certain range was understood to possibly lessen the precision of the Rasch analysis (Eckes, 2015).

Procedure

The workflow had participants write, evaluate, and make interpretations. The instructor gathered data, conducted the Rasch analysis, and prepared a summarized report on the analysis and feedback comments for the participants. The scheme took place over 16 sessions, spanning 10 weeks. In brief, each participant

1. covered course materials;
2. completed the writing task;
3. assessed and gave feedback on a selection of writings from peers (quantitatively and qualitatively, assigned via the judging plan);
4. received an analysis and feedback report prepared by the instructor, with the instructor's feedback added anonymously to the report;
5. made interpretations from the experience and from the results of the report; and
6. completed a reflection paper.

The writing task was a single 1000-word argumentative essay centering on a topic raised by articles in the course book focusing on leadership in business. Participants had to critically address conflicting aspects of the articles and introduce counterarguments using sources as appropriate. They were to make a claim and support it, that is, not to merely summarize the content of articles. The participants kept their essays private in the early stages to maintain anonymity.

One exemplar paper, coded p-E, was included anonymously from a previous course. It was chosen for having above average language mechanics but with a strongly worded, controversial argument. Each participant was assigned this exemplar paper in addition to a subset of four anonymized essays written by peers according to the judging plan demonstrated in Table 1. Therefore, the participants each had a total of five essays to evaluate along three criteria. Each participant received PDF files of their assigned essays and completed an online form from home. No minimum wordcount or other requirements were requested of the participants. The evaluations were collected, anonymized and kept in a password-protected file offline. Teacher assessment and written feedback for all papers was included anonymously into the data at this point.

The analysis and feedback report were prepared by the instructor for the participants. Calculations and the output of data were done with Facets software, version 3.83.2 (Linacre, 2020a). The data produced by this software is typically meant for research purposes, so the instructor simplified this output for the participants. The report consisted of the following:

- vertical rulers (seen in the results section)
- simplified fit statistics
- tables of unexpected responses
- simple explanation and instructions
- the feedback the participant's essay received
- all feedback received by the exemplar essay

The Rasch analysis included four sets of results: one for all criteria combined and one each for the three criteria. Participants were given a unique ID code to locate their results in relation to the anonymized essay and rater data. The identity of each participant was protected within the report. Participants also received anonymized feedback comments for their paper. Participants did not revise their original papers; rather, each participant completed a reflection paper on their experiences at the different stages during the process.

Reflection

The participants wrote a 500-word reflection paper to bring together and reinforce their learning, as underpinned by Bloom's taxonomy (Anderson & Krathwohl, 2000; Bloom, 1956). Writing the reflection paper invited the participants to once again learn by assessing (Topping, 1998). This self-assessment gave the participants an opportunity to review the ratings and feedback they had given and received, evaluate conflicting experiences, and reflect. The reflection paper also served as data for the researcher to determine the efficacy of the procedure.

A final step was to anonymously complete an exit survey with the following five statements, each rated on a 6-point agreement scale:

1. Giving feedback and ratings to papers written by my peers required deep thinking.
2. Giving feedback and ratings helped me better self-assess my own writing.
3. It was good that the peer-exchange was anonymized (names were kept secret).
4. I can better see how responses by readers can be unpredictable or unexpected.
5. The final analysis and feedback report were helpful.

Analysis

The analysis focused on aspects of the feedback comments, the class-wide measures, and the participants' perceptions about the procedure. The volume and variety of feedback comments were analyzed in relation to the effects of using the judging plan. The model statement was set before running Facets software, version 3.83.2 (Linacre, 2020a; also see Linacre, 2020b, for further examples and explanation of model statements). The raters and the essays were set to the rating scale model. The facet of criteria was set to the partial credit model, because it was assumed that use of rating scales would vary by criteria (Value to Reader, Clarity of Concept, and Language Mechanics). Attention was paid to underfitting (unpredictable) ratings where raters erroneously rate less-effective papers leniently or rate highly rated papers harshly. Attention was also paid to overfitting raters, those who either used a narrow range on the Likert scale or those who potentially exhibited a halo effect and scored the criteria holistically. General impressions of the participants' perceptions about the procedure were analyzed by examining the results of the anonymous exit survey and by highlighting selected excerpts from the reflection papers.

Results

Descriptive statistics on feedback

A summary of the amount and distribution of feedback comments indicated that the judging plan ensured that each paper received a reasonable amount of feedback despite the wide variation in feedback length given by participants. To recap, each participant wrote a 1000-word argumentative essay and a 500-word reflection paper. No word requirements were imposed on feedback. Even so, most participants voluntarily wrote a sizable amount in relation to the writing assignments.

Each participant wrote on average 292.6 words (*SD* = 202.2 words) of feedback in total for their five assigned papers. The high variation of feedback given by the participants was balanced out via the structure of the judging plan. As a result, the standard deviation for the feedback each paper received was halved (*SD* = 102.7 words). Each paper received a mean amount of 241.2 words in total feedback from the four classmates assigned as reviewers. This was in addition to teacher comments, which were written anonymously and roughly matched in length. Reasons for variation in participant feedback are unclear, but the data indicates that the judging plan effectively served to balance out the effects of participant variation and provide a base level of feedback for each participant's essay.

A portion of the variance in feedback for each essay appears to be accounted for by a moderate negative relationship between the evaluation of essays and the word counts of feedback ($r = -.529, p = .047$), which would indicate that essays with lower evaluations tended to receive more written feedback than more highly rated essays. The evaluation of each paper was estimated using its adjusted fair score via Rasch measurement. The exemplar essay, reviewed by each participant, received a total of 771 words of feedback.

However, there was a range of feedback for this essay, with an average of 55 words produced by each participant (n = 14 participants giving feedback, *SD* = 53.5 words) and one participant giving no feedback. Put all together, however, this collection of feedback provided all participants a common reference to review and re-evaluate within the analysis and feedback summary report.

Rasch measures and fit statistics

The "vertical ruler" in Figure 1 is a graphic display that shows logit measures for the three criteria combined and it is indicative of the criterion-specific vertical rulers also in the participants' feedback and assessment report. For value, concept, language, and all criteria combined, the rating data is transformed along a common equal-interval measure of logits, or "log-odds units" (see Bond & Fox, 2015, p. 46). The left side of Figure 1 shows participant ability (writing performance) ranked from high to low (top to bottom), with argumentative essays (papers) P, C, and N receiving the highest adjusted scores. The next column represents rater severity (with leniency at the bottom), where raters #12, #11, and #2 appeared to be the strictest raters. Comparing laterally, the vertical rulers line up participant ability with rater severity. While each rater rated only five papers, the class-wide approximation in Figure 1 tells us that Rater #14, for example, would have probably rated half of all essays more positively and the other half more negatively. Papers J and B would have a 50/50 chance of being rated high or low by Rater #14. Such comparisons imply a rough match at this point between performance and severity.

However, the logit spread observed for rater severity (3.39 logits) exceeds the logit spread for writing performance (2.82 logits). The severity and lenience of Raters #7 and #12 may have been exaggerated due to too many raters using a narrow range of the scales for each criterion. The category threshold measure on the right side of the vertical rule demonstrates that many papers received scores of 7 and 8 across, particularly for the criteria of Value. This narrow use of the scale reduces the accuracy of measurement. Use of the whole scale was discussed with participants in class, but in the end, the rating patterns seemed to mimic the typical school grading pattern A-D and F. This instinctive use of a 10-point scale is natural and should have been better foreseen in the design. Research suggests that a scale of six categories or less can ensure better measurement by minding potential limits in raters' motivation and working memory (see Nemoto & Beglar, 2014).

```
+-----------------------------------------------------------------------------------------+
|Measr|+Student Ability|-Rater Severity    |-Rating Criteria Difficulty | S.1 | S.2 | S.3 |
|-----+----------------+-------------------+----------------------------+-----+-----+-----|
|   2 + p-P            +                   +                            +(10) +(10) +(10) |
|     |                | r-12              |                            |     |     |     |
|     |                |                   |                            |     | --- |     |
|     | p-C  p-N       |                   |                            |     |     | --- |
|     |                |                   |                            |     |     |     |
|     | p-E            |                   |                            |  8  |     |     |
|     | p-D  p-H       |                   |                            |     |     |     |
|     | p-O            |                   |                            |     |     |     |
|     |                |                   |                            |     |  8  |  8  |
|     |                |                   |                            |     |     |     |
|   1 +                + r-11              +                            +     +     +     |
|     |                | r-2               |                            |     |     |     |
|     |                |                   |                            | --- |     |     |
|     |                |                   |                            |     | --- | --- |
|     | p-J  p-B       | r-14  r-15        |                            |     |     |     |
|     | p-G            |                   |                            |     |     |     |
|     | p-S  p-I       |                   |                            |     |     |     |
|     | p-F            | r-8               |                            |     |     |     |
|     |                |                   |                            |  7  |     |  7  |
|     |                | r-6               | Value to Reader            |     |  7  |     |
*   0 *                * r-13              * Clarity of Concepts/Ideas  *     *     *     *
|     | p-L            | r-9               | Language Use and Mechanics |     |     |     |
|     |                | r-1               |                            | --- |     |     |
|     | p-K            | r-10              |                            |     | --- | --- |
|     |                |                   |                            |     |     |     |
|     |                | r-5               |                            |  6  |     |     |
|     |                |                   |                            |     |     |     |
|     |                |                   |                            | --- |  6  |  6  |
|     |                |                   |                            |     |     |     |
|     | p-M            | r-16  r-4  r-3    |                            |  5  |     |     |
|  -1 +                +                   +                            +     + --- + --- |
|     |                |                   |                            | --- |     |     |
|     |                |                   |                            |     |     |     |
|     |                |                   |                            |  4  |  5  |  5  |
|     |                |                   |                            |     |     |     |
|     |                | r-7               |                            |     |     |     |
|     |                |                   |                            | --- | --- |     |
|     |                |                   |                            |     |     | --- |
|     |                |                   |                            |     |     |     |
|     |                |                   |                            |     |     |     |
|  -2 +                +                   +                            + (1) + (3) + (3) |
|-----+----------------+-------------------+----------------------------+-----+-----+-----|
|Measr|+Student Ability|-Rater Severity    |-Rating Criteria Difficulty | S.1 | S.2 | S.3 |
+-----------------------------------------------------------------------------------------+
```

Figure 1. Vertical Rulers for All Three Criteria. Right-side column labels: S.1 = Value to Reader, S.2 = Clarity of Concept, S.3 = Language Mechanics.

Regarding fit of the data, one third of the participants fit the Rasch model (Table 2, bolded), while almost half of the participants overfit, likely giving many essays common scores of 7 and 8. A few participants underfit by giving unpredicted responses such as assigning highly scored papers low scores, or vice versa. Table 2 also shows that the infit and outfit measures by each participant varied by criterion (Table 2, fitting between MnSq 1.00+/-.50 in bold by criterion). The fit range of MnSq 1.00+/-.50 for each rater and paper is most productive to the model's overall measurement (Linacre, 2020b, p. 286). Outlying ratings were predicted to become a resource for participants to evaluate. However, data overfit was more than anticipated and this reduced the accuracy of measure. Scores of 1-5 points were rarely given. In an attempt to narrow the logit range of rater severity, the data was re-analyzed on a 7-point scale, with categories 1-4 combined. A paired sample *t*-test saw marginal improvement, but not at the level of significance. No outlying data was removed because this was not a summative assessment and because each participant's data was part of the analysis and feedback report. For more precise measures, the papers would need to be reassessed using a 6-point scale, for example, as outlined by Nemoto and Beglar (2014). For the current results, the separation of assessment for each criterion was relatively low for both raters and papers (2.29, 2.69, 1.51 and 2.19, 2.35, 1.79, respectively) due to the narrow use of the scales. This indicated that raters and papers were not sufficiently separated into groups of 3 or more. Nonetheless, the data was able to show a low-resolution separation of performance and severity. The participants were informed that finer positions between papers versus raters are not as accurate.

Table 2
Fit Statistics for Rater Severity

	Infit	Infit by Criteria			Outfit	Outfit by Criteria		
Rater	Combined	Value	Concept	Language	Combined	Value	Concept	Language
r-1	3.88	2.21	3.08	3.84	3.91	2.81	2.93	3.49
r-2	2.04	1.91	1.99	2.78	1.98	1.93	1.95	2.11
r-3	1.74	1.65	2.98	1.53	2.25	2.88	3.90	1.86
r-4	**1.46**	1.92	0.42	2.00	**1.47**	1.93	0.52	2.01
r-5	**1.22**	0.47	**1.13**	1.00	**1.21**	0.49	**1.10**	**1.00**
r-6	**1.07**	**0.81**	**0.94**	1.88	**1.19**	**1.00**	**0.97**	2.10
r-7	**0.90**	**0.71**	**1.08**	**0.53**	**0.85**	**0.74**	**1.01**	**0.52**
r-8	**0.57**	0.33	**0.50**	**0.90**	**0.57**	0.42	**0.51**	**0.88**
r-9	**0.54**	0.29	**0.71**	0.43	**0.57**	0.37	**0.72**	0.43
r-10	0.48	**0.71**	0.27	0.42	0.45	**0.67**	0.27	0.38
r-11	0.47	**1.45**	0.38	0.14	0.48	**1.45**	0.37	0.13
r-12	0.45	0.28	0.19	**1.08**	0.45	0.23	0.17	**1.02**
r-13	0.45	0.08	**0.77**	0.20	0.44	0.12	**0.77**	0.18
r-14	0.35	0.34	0.19	**0.61**	0.32	0.34	0.19	**0.52**
r-15	0.33	0.35	0.20	0.19	0.31	0.47	0.20	0.22
r-16	0.14	0.05	0.14	0.15	0.13	0.06	0.15	0.12
Mean	1.01	0.85	0.93	1.11	1.04	0.99	0.98	1.06
SD	0.94	0.73	0.95	1.06	0.98	0.93	1.07	0.98

Note. Ordered by the infit figures of the three criteria combined (1.00+/-.50 MnSq bolded). Participant group of 15 members, one instructor. Criteria = partial credit model.

The fit statistics for the papers revealed a marked result for the exemplar paper, coded p-E. As predicted, the fit statistics for p-E indicated unpredictable responses for Value to Reader (Infit MnSq = 1.99, Outfit MnSq = 2.17) and were the only underfitting figures for value among the papers (Table 3). More variation in responses could be because it was rated by all participants. It could also be due to it garnering a wider range of reactions to the paper's somewhat strongly stated arguments. The participants appeared to assess the value of p-E in different ways and this was also borne out in the essay's qualitative feedback comments as well. Both the comments and rating results for exemplar p-E were provided to all participants as an informative subsection of the analysis and feedback report.

Table 3
Fit Statistics for Participant Ability (Performance of Papers)

	Infit	Infit by Criteria			Outfit	Outfit by Criteria		
Rater	Combined	Value	Concept	Language	Combined	Value	Concept	Language
p-A	1.86	**0.58**	2.35	2.93	1.96	**0.83**	2.39	3.08
p-B	1.81	**0.83**	**1.33**	2.51	1.75	**0.70**	**1.28**	1.98
p-C	**1.55**	**0.94**	**1.10**	**1.11**	1.84	**0.72**	**1.14**	**0.99**
p-D	**1.28**	**0.97**	0.42	**0.85**	**1.45**	**1.09**	**0.55**	**0.84**
p-E	**1.25**	1.99	1.86	**1.07**	**1.27**	2.17	1.94	**1.03**
p-F	**1.06**	**1.14**	**1.24**	**0.86**	**1.10**	**1.35**	**1.21**	**0.95**
p-G	**0.81**	**0.86**	**0.79**	**1.42**	**0.85**	**1.05**	**0.82**	1.57
p-H	**0.80**	**0.82**	**0.61**	**0.80**	**0.80**	**0.81**	**0.62**	**0.80**
p-I	**0.75**	0.40	**0.77**	1.89	**0.77**	0.41	**0.78**	1.96
p-J	**0.69**	**0.67**	0.16	0.30	**0.63**	**0.78**	0.17	0.27
p-K	**0.68**	**0.71**	0.14	**0.54**	**0.67**	**0.70**	0.15	**0.54**
p-L	**0.53**	0.20	**0.69**	**0.80**	**0.51**	0.19	**0.70**	**0.73**
p-M	**0.50**	0.11	0.46	**0.54**	0.49	0.09	0.47	**0.54**
p-N	0.37	0.43	0.16	0.19	0.38	0.46	0.18	0.19
p-O	0.35	0.33	0.46	0.04	0.35	0.34	0.44	0.04
p-P	0.21	0.26	0.20	0.16	0.21	0.26	0.21	0.15
Mean	0.91	0.70	0.80	1.00	0.94	0.75	0.82	0.98
SD	0.51	0.46	0.64	0.83	0.56	0.51	0.65	0.81

Note. Ordered by the infit figures of the three criteria combined (1.00+/-.50 MnSq bolded). Fifteen papers written by the participants, one exemplar. Criteria = partial credit model.

The anonymous exit survey attempted to measure how the participants perceived different aspects of the peer-exchange process (Table 4). The results for the 6-point scale agreement items indicated that most participants could strongly endorse the five statements. However, the minimum and maximum values show that participant sentiment was not uniform. Scores of 2 and 3 showed that a few participants disagreed with at least some of the intentions of the design.

Table 4
Exit Survey Results

	Q1	Q2	Q3	Q4	Q5
Replies	15	15	15	15	15
Mean	5.00	5.07	5.13	5.07	5.13
Std. Deviation	1.00	0.88	0.92	0.88	0.92
Skewness	-1.98	-0.86	-0.94	-0.14	-0.94
Kurtosis	5.68	0.67	0.52	-1.78	0.52
Minimum	2.00	3.00	3.00	4.00	3.00
Maximum	6.00	6.00	6.00	6.00	6.00

Note. 6-point scale (1 = strongly disagree, 6 = strongly agree).

Discussion

Theoretical aims

The act of assessing and giving feedback, as in Topping's (1998) view, seemed effective for learning in the Piagetian sense where each participant could encounter *cognitive conflict* between what they think they know and new experiences. The learning environment was student-centered and it provided the participants opportunities to better resolve conflicting information independently. The data generated suggested that the participants were learning by assessing at three stages of the exchange procedure. The first stage was when they assessed and gave feedback on their peers' papers. The second was when they had to evaluate the measures and qualitative feedback of various types provided in the report. Finally, the reflection paper provided a third opportunity for the participants to consolidate their learning and re-assess their performance.

The participants successfully completed each task and submitting work on time at each stage and this helps confirm that the process was straightforward for the students to follow. To attempt this peer exchange approach in another context, the instructor would need the organizational skills to manage the exchange of data and the knowledge needed to use Rasch analysis software (see e.g., Bond & Fox, 2015; Eckes, 2015; Linacre, 2020b). Attempting to transfer this peer feedback and peer assessment design to other learning situations would require adjusting the procedure accordingly. The length and number of tasks to complete could be carefully adjusted based on learner proficiency and learning aims. Writing tasks other than essay writing are possible. One example could be to have students produce and exchange a set of pragmatically written emails. In any case, the design generally appears to be applicable to writing development in a language course given proper preparation. This includes the use of a judging plan to enable a manageable and balanced exchange of varied feedback and assessment ratings.

The participants took on the role of the distant reader, as viewed by Booth et al. (2008), but the results of the fit statistics showed that the consistency of their assessments was less than expected. For a standardized testing situation, such evaluation would not be acceptable, but there was much evidence that a mechanism for learning was achieved within this peer exchange environment. It was key for the participants to interact with their peer's writing, the rubric, and the subsequent feedback information during the procedure. The participants saw the varying interpretations of value, especially in the case of paper p-E. Generally, the ratings by the participants diverged at a finer scale, but the exit survey showed high engagement, growth in critical thinking skills, and improved self-awareness in the writing process. Pedagogical value could then be seen in formative aspects of writing development rather than summative testing.

Perceptions by participants

Through observation of the participants' work and in-class discussion, it was clear to the researcher that the approach was positively received. This impression about the participants' perceptions and engagement was supported by the group's effortful qualitative comments in many of the reflection papers. In particular, the participants seemed to value anonymity as a way to better embody the role of a reader, apart from social distraction. One participant could experience how writers sometimes structure arguments in ways that are difficult for readers to receive.

> I was able to understand how other people write their essays (good and bad) and from those problems such as unclear structure, tone used in essay as well as grammar misusing and so on, I managed to understand how it feels when seeing these problems as a reader. I was like a lost cat in the forest could not find the way out. I assume this is what happens

> to those people who were evaluating my essay, too. It must be very difficult and hard. – Participant A

Another participant reflected on the conflict within the evaluation process.

> I honestly feel happy to get those mixed feedbacks from my classmates and the scoring because it encouraged me to improve my learning ability on logical and creative thinking for understanding the articles. Exchanging feedbacks in this course has not only helped me to develop self-awareness of my writing ability, but also helped me to know how to provide value to the audience with our creative thinking and writing skill. – Participant B

While the participants generally seemed to find the process insightful, a few participants shared a preference for teacher feedback. Interestingly, the teacher's comments were in fact provided anonymously within the analysis and feedback report. Each participant also received additional and open feedback after the process from the instructor. This raises interesting questions about the perceptions of feedback, its types, and its sources. This could be an area for further investigation.

Future considerations

To remedy the issues about accuracy of measurement, the design of the rating scale would be the first consideration. The 10-point scale in this context too closely resembled a typical school grading system, leaving half of the Likert scale (points 1-5) not being utilized by most participants. The rating patterns of most participants saw many 7s and 8s (analogous to C and B grades) and almost no scores below 6 points for the three criteria. This narrow use of the ratings scales likely exaggerated the unpredictability of some other raters. Such measurement error could be avoided by choosing a point scale that more intuitively encourages broader scoring patterns, which leads to more productive ratings for Rasch analysis. This would be a highly economical solution instead of other possible strategies like time spent on rater training. Even with time and robust resources, rater training can yield limited benefits (Eckes, 2015; Farrokhi et al., 2012; Weigle, 1998). Nemoto and Beglar (2014) outlined how a 6-point scale would avoid over-stressing the working memory of raters by applying descriptors for a smaller number of categories on each scale. An even-numbered scale would also prevent participants from making neutral assessments. By simplifying the scale, learners may be able to better interact with peer work clearly through the rubric, even if interpretations vary. This interaction can become an improved mechanism for independent learning in addition to the receiving of feedback and assessment. If applied in other L2 learning contexts, the criteria of a rubric with simplified scales could be adjusted according to the task. For the procedure as a whole, considerations for L2 proficiency, writing or speaking task, and learning goals would need to be made carefully.

Conclusion

Anonymity and Rasch analysis via a judging plan were successfully employed in a peer feedback and assessment design to facilitate the aims of building awareness of reader needs and the ability to resolve differences among a variety of feedback and assessment. These features provided all participants in the group with class-wide measures on writing performance and peer rating patterns. The results of the study showed that the judging plan also helped ensure that the variation of feedback given by individuals was balanced out and allowed for a variety of peer-feedback for each member of the group. Participant reflections and exit survey responses indicated positive reception by the participants and that the goal of creating learning opportunities for participants at multiple stages was met. However, the review of the Rasch analysis figures raised concerns over lower-than-expected accuracy of the measure due to a narrow use of the rating scales. A 6-point rating scale for each rating criterion is proposed in order to better elicit discerning scoring responses by peer raters. The results of this single application cannot be easily generalized beyond the context of these learners. However, its evidence of peer-assisted learning and the feasibility of its improvements suggested that this design for peer feedback and peer assessment with L2 participant writers is worth further investigation.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2000). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Pearson.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners. Handbook 1: Cognitive domain*. Longman.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge.

Booth, W. C., Colomb, G. G., & Williams, J. M. (2008). *The craft of research* (3rd ed). University of Chicago Press.

Cheng, K.-H., & Tsai, C.-C. (2012). Students' interpersonal perspectives on, conceptions of and approaches to learning in online peer assessment. *Australasian Journal of Educational Technology*, *28*(4), 599–618. https://doi.org/10.14742/ajet.830

Eckes, T. (2015). *Introduction to many-facet Rasch measurement*. Peter Lang.

Falchikov, N. (2001). *Learning together: Peer tutoring in higher education*. Routledge.

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, *34*(1), 79. https://doi.org/10.37546/JALTJJ34.1-3

Foot, H., & Howe, C. (1998). The Psychoeducational Basis of Peer-Assisted Learning. In K. J. Topping & S. W. Ehly (Eds.), *Peer-assisted learning* (pp. 29–39). Routledge.

Garner, B. A. (2012). *HBR guide to better business writing*. Harvard Business Review Press.

Harvard Business Review. (2011). *HBR's 10 must reads on leadership*. Harvard Business Review Press.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, *39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2020a). *Facets®* (3.83.2) [Computer software]. https://winsteps.com

Linacre, J. M. (2020b). *A User's Guide to Facets Rasch-Model Computer Program*. winsteps.com.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, *3*(4), 486–512.

Liu, J., & Hansen, J. G. (2002). *Peer response in second language writing classrooms*. University of Michigan Press.

Liu, N. F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, *11*(3), 279–290. https://doi.org/10.1080/13562510600680582

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, *18*(1), 30–43. https://doi.org/10.1016/j.jslw.2008.06.002

Nemoto, T., & Beglar, D. (2014). Developing Likert-Scale Questionnaires. In N. Sonda & A. Krause (Eds.), *JALT2013 Conference Proceedings* (pp. 1–8). JALT. Available online: http://jalt-publications.org/files/pdf-article/jalt2013_001.pdf (accessed on 18 November 2020).

Piaget, J. (1971). *Biology and knowledge: An essay on the relations between organic regulations and cognitive processes*. University of Chicago Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability: Volume IV: Contributions to biology and problems of medicine* (pp. 321–333). University of California Press.

Rotsaert, T., Panadero, E., & Schellens, T. (2018). Anonymity as an instructional scaffold in peer assessment: Its effects on peer feedback quality and evolution in students' perceptions about peer assessment skills. *European Journal of Psychology of Education*, *33*(1), 75–99. https://doi.org/10.1007/s10212-017-0339-8

Topping, K., & Ehly, S. (1998). *Peer-assisted Learning*. Routledge.

Topping, K. J. (1998). Peer Assessment Between Students in Colleges and Universities. *Review of Educational Research*, *68*(3), 249–276. https://doi.org/10.3102/00346543068003249

Vygotsky, L. S. (1978). *Mind in society*. Harvard University Press.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287.

Xu, Y., & Liu, J. (2010). An investigation into anonymous peer feedback. *Foreign Language Teaching and Practice*, *3*, 44–49.

Yu, S., & Lee, I. (2016). Peer feedback in second language writing (2005–2014). *Language Teaching*, *49*(4), 461–493. https://doi.org/10.1017/S0261444816000161

Zamel, V. (1982). Writing: The process of discovering meaning. *TESOL Quarterly*, *16*(2), 195. https://doi.org/10.2307/3586792

Call for Papers

Shiken is seeking submissions for future issues on an ongoing basis. After checking the guidelines for publication below, please send your submission to the editor at tevalpublications@gmail.com.

Overview

Shiken aims to publish articles concerning language assessment issues relevant to assessment professionals, researchers, classroom practitioners and language program administrators. *Shiken* is dedicated to publishing articles on assessment in various educational contexts in and around Japan.

Article formats include research papers, replication studies, and review articles, all of which should typically not exceed 7000 words; as well as informed opinion pieces, technical advice articles, and interviews, all of which should typically not exceed 3000 words. Please contact the editor if you wish to submit an article that differs from these formats.

Novice researchers are encouraged to submit, but should aim for papers that focus on a single main issue. Please review recent issues of *Shiken* to understand the level of detail, depth of discussion and provision of empirical evidence that is required for acceptance.

Format

To facilitate double-blind peer review, the name(s) of the author(s) should not be mentioned in the manuscript. All citations and references to the author or co-author's work should read (Author, Year) e.g. "(Author, 2017)."

Articles should be formatted according to the guidelines of the *Publication Manual of the American Psychological Association (APA), 7th Edition*. Submissions should be formatted in Microsoft Word (.docx format) using 12-point Times New Roman font. The page size should be set to A4, with a 2.5 cm margin. The body text should be block justified, and single-spaced. Paragraphs should be separated by a blank line space. Each new section of the paper should have a section heading. Please review recent issues of *Shiken* for examples of section headings.

Research articles must be preceded by an abstract that succinctly summarizes the article content in less than 200 words. The reference section should begin on a new page immediately following the body text. Authors are responsible for comprehensive and accurate referencing. Tables and figures should be numbered and titled. Separate sections for tables and figures should follow the references. Any appendices should be numbered and should follow the tables and figures.

Evaluation

All papers are double-blind peer-reviewed by at least two expert reviewers. Initial evaluation is usually completed within four weeks. The whole process from submission to publication is normally completed within six months. Submissions should be sent to the editor at tevalpublications@gmail.com.