Using dictation to measure language proficiency: A Rasch analysis

Paul Leeming
Kindai University, Japan
Aeric Wong
Konan University, Japan

Groups are used widely in the language classroom and, particularly in classes where there is a wide range of English proficiency among students, teachers may want to construct balanced groups based on the language proficiency of individual students. In order to construct such groups, teachers need a reliable measure that effectively differentiates between different levels of proficiency, and yet there are contexts where information regarding student proficiency may not be available. This paper reports on the use of an in-house dictation test to measure the English proficiency of students in a Japanese university. Rasch analysis was used to determine the degree to which the dictation differentiated between the range of proficiencies in the classes, and to assess the reliability of the test. Correlation with scores from *TOEIC* and *SLEP* tests was used to confirm that the dictation tests English proficiency. Results show that dictation is a simple, cheap, and effective means of assessing the relative proficiency of students in this context, and can be used for constructing balanced groups.

Key words: testing, dictation, proficiency, Rasch analysis, group construction.

Introduction

With the spread of communicative language teaching (CLT), and more recently task-based language teaching (TBLT) (Howatt & Widdowson, 2004), small group work has become central to most teaching approaches in the foreign and second language classroom. Group work allows students greater opportunities to practice in the target language (Long & Porter, 1985), and this interaction increases opportunities for output, which has been shown to be a necessary condition for language acquisition (Swain, 2005). One has only to thumb through the many textbooks used

Email address for correspondence: paulleeming@yahoo.com

in English language teaching to see the extent to which small group work is used in almost all stages of a lesson.

The proliferation of small group work in the language classroom means that teachers are faced with several issues regarding group construction. Cooperative learning practitioners argue that teachers should construct groups for specific purposes (Cohen, 1994), and within an SLA context Jacobs (2006) argues that small groups should be heterogeneous in terms of English proficiency and personality. Many teachers support this, and seem to believe that mixed proficiency groups provide the most learning opportunities for students, as the stronger students benefit from being forced to provide metalinguistic explanations to the weaker students, who conversely are able to learn from their more proficient peers. Based on the research of Vygotsky (1986), Lantolf (2006) has shown that peer-peer learning can occur within the zone of proximal development (ZPD), and that more capable peers are able to provide scaffolded assistance leading to language learning. A body of research exists showing that the relative level of proficiencies in peer interaction will have a direct influence on the nature of the interaction, including the number of instances of negotiation of meaning, and use of the first language (Philp, Adams, & Iwashita, 2013). Generally, it seems that the benefits of heterogeneous groups will be for the weaker students who are provided with scaffolded opportunities to develop their language resources, although there is also the possibility that they will be cut off by the more competent speaker (Philp, Adams, & Iwashita, 2013). If we are going to use small group work in the language classroom then it would seem that creating mixed proficiency groups where stronger students can assist others within the ZPD will have benefits for all the students.

In order to create groups based on language proficiency we need information regarding the relative proficiency of students within our class. Although in some contexts proficiency scores may be available, there are many countries where standardized proficiency tests such as TOEIC (Test of English for International Communication) are too expensive, or the information is considered sensitive and withheld from the teacher. In developing countries in South-East Asia for example, proficiency tests can be almost two-thirds of the average monthly salary, and therefore are clearly beyond the means of many schools and universities. Free online assessments such as vocabulary tests do exist, but again these tests require a large number of computers for students and access to the internet. In a country such as Japan, institutions have become increasingly cautious with treatment of personal information, and students' test scores may be withheld from teachers, as was the case in the context for this study. In these situations teachers need a quick, simple,

cheap, and accurate means of measuring the relative proficiency of students in their class, and this paper considers dictation as a means for achieving this.

After a brief review of dictation as a test of English language proficiency, this paper introduces the dictation used in the current study, and presents a Rasch analysis of the results to determine test reliability, and to ascertain whether the test is effectively differentiating between the different levels of proficiency of the students in this context. The paper then presents the results of a correlation analysis used to determine if the test is measuring the English proficiency of the students.

Although there are many discussions in the literature concerning test validity (see Fulcher & Davidson, 2007), this paper focuses on the practical application of the test score and adopts Bachman and Palmer's (2010) Assessment Use Argument (AUA) in order to justify the use of dictation in this context. This framework was developed in order to help test developers communicate the rationale behind the decisions made during test formation. Figure 1 gives a simple overview of the framework based on Bachman and Palmer (2010). The first stage is considering the consequences of the test, which in this case is the formation of well-balanced groups to facilitate language learning. There are no perceived negative outcomes from this, although perhaps more advanced learners could be held back. Bachman and Palmer (2010, p. 87) then talk about decisions to be made based on the assessment. In the current study this is the assignment to groups and again this can be seen to have a very specific benefit for the students.

Perhaps a more difficult aspect of the argument in the case of dictation is the interpretation of scores. The interpretation used is that performance on the dictation test is representative of English proficiency, and more specifically, is likely to influence performance in class when interacting in groups. This would suggest that the dictation test is measuring speaking and listening performance. Bachman and Palmer (2010) state that at each stage there should be claims and data to support what is being claimed. The dictation clearly is aural input related to listening performance, and correlation with the TOEIC listening test will help to provide the data to support this part of the argument.

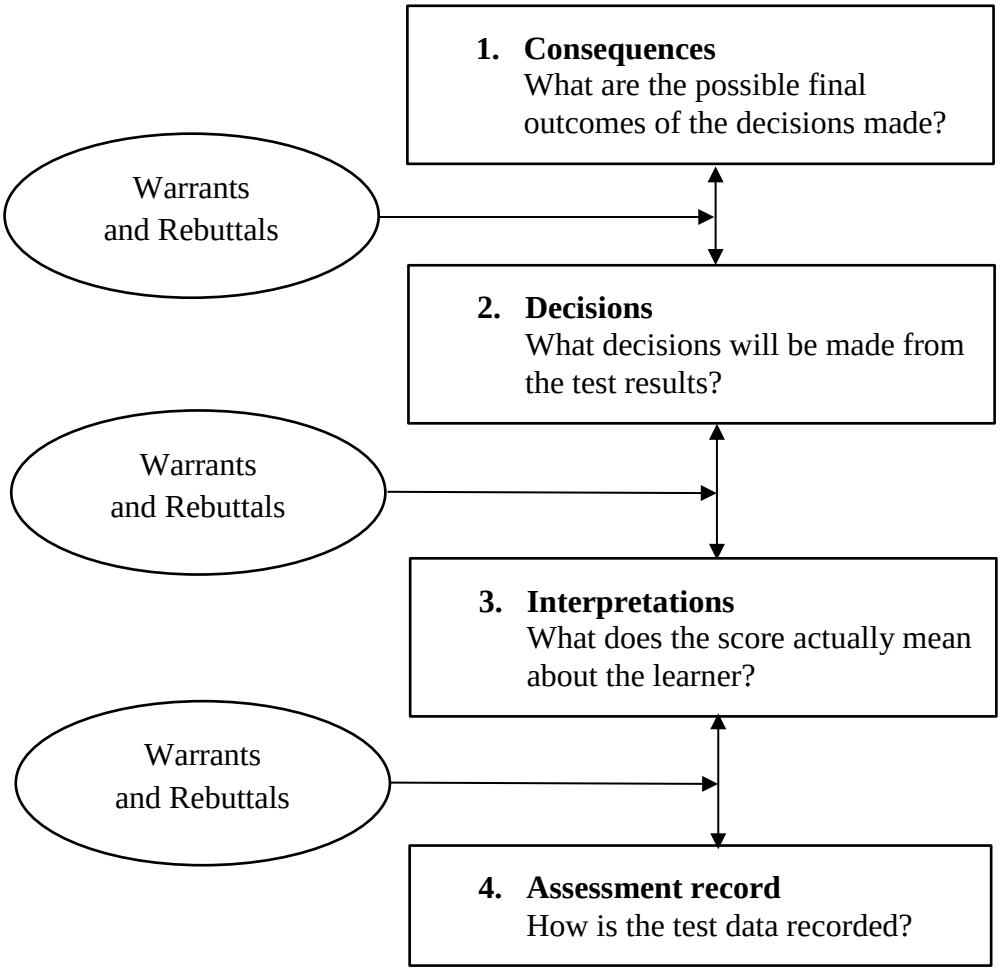

Figure 1. Assessment Use Argument (based on Bachman and Palmer, 2010)

The next part of the AUA is the assessment record, and the claim that the record (in this case the students' transcription of the dictation) represents the students' language proficiency. The dictation is aural and yet the test calls for the students to write down what they have heard. Oller and Streiff (1975) believe that students need to use their grammatical knowledge of the language in order to reconstruct what they heard, and therefore written production is reasonable approximation to production through speaking, where students must construct sentences based on their knowledge of the language system. A further issue cited by Bachman and Palmer (2010) here is that there is consistency across different test administrations, and this seems to be satisfied in this case, as all students were given the test by the same teacher. Also, the dictation was recorded rather than read, ensuring that administration was the same for all students.

In order to fulfill the AUA as described above, the current study needs to provide data to show that the scores can be used to make decisions when constructing

groups based on proficiency, and therefore that the test is effective in separating out the relative abilities of the students in this study. The other claim that needs some data is the interpretation of the scores, and correlation with established measures of English proficiency will support the argument that dictation does measure language proficiency.

Dictation and Language Testing

The use of dictation to measure language proficiency can be dated back as far as 1913, as part of the Certificate of Proficiency in English (CPE) (Weir, Vidaković, & Galaczi, 2013). Oller (1971) was the first researcher in SLA to analyze dictation as a measure of language proficiency, and used correlation analysis to compare dictation with a general English proficiency test, which along with a dictation section, comprised of vocabulary, a composition, a phonological discrimination task, and the selection of grammatically acceptable sentences. Oller (1971) showed that dictation correlated most strongly with each other part of the test, and claimed that dictation was an effective test of general English proficiency, with a .86 correlation with the overall composite score. The only test part to correlate more strongly with the overall composite score was the composition section of the test that correlated at .88.

Following some methodological criticism of Oller's 1971 paper, Oller and Streiff (1975) revisited the data and conducted a re-evaluation. After describing each of the different tests in detail, they provided their analysis which corrected some statistical errors from the previous paper, and found even stronger correlations between dictation and the combined measures of language proficiency. The dictation correlated at .96 with the overall test score, which was the strongest correlation from all parts of the test. The authors (Oller & Streiff, 1975) suggested that "dictation activates the learner's internalized grammar of expectancy, which we assume is the central component of his language competence" (p. 34). Although students are supplied with the language, they must rely on their knowledge of grammar and syntax in order to accurately reproduce what they have heard. This means that dictation tests not only decoding, but also the grammatical parsing of information (Rost, 2005).

Other researchers have continued to investigate dictation as a measure of proficiency, and Fouly and Cziko (1985) used a "graduated" dictation test, where texts became progressively longer and therefore more difficult. They checked the correlation of the dictation with the IEPT Cloze Test, IEPT Grammar Test, and the Test of English as a Foreign Language Test (TOEFL). The total TOEFL score was used along with subsets of listening comprehension, reading comprehension, and

structure and written expression. The dictation showed moderate correlations with the other measures ranging from .50 to .60, with the strongest correlation to the overall TOEFL score. This supports the prior claim of Oller (1971) that dictation is a useful measure of general English proficiency. Fouly and Cziko (1985) argued that the test can be shown to be unidimensional, and as it is cumulative, can test students with a wide range of language proficiencies. They also suggested that, although graduated dictation tests are little more difficult to construct than a standard dictation such as that used by Oller (1971), this kind of test is easier to score.

Stansfield (1985) wrote a comprehensive history of the use of dictation as a measure of language proficiency, and claimed that at the time of writing dictation was a popular measure of proficiency, and could be used as an effective placement test. He did state in his conclusion that dictation may wane in popularity due to the cyclic nature of attitudes within language teaching, and a subsequent lack of interest in dictation as a measure of proficiency ensued.

In perhaps the most recent paper to consider dictation as a test of proficiency, Cai (2012) considered partial dictation, where part of the text is provided for the students who must complete it while listening to the full text. Cai (2012) argued that generally language tests are difficult to construct and are therefore often done badly, and that the simplicity of dictation makes it appropriate for teachers with limited time. Cai (2012) claimed that partial dictation is simpler to administer and easier to score than a standard dictation, although concedes that test takers can listen for specific words without understanding the flow of information. There has been some argument as to the construct that dictation is measuring (Stansfield, 1985), and Cai (2012) used confirmatory factor analysis in an attempt to clarify what is being measured. He concluded that the data most closely represents a model where dictation is measuring both lower- and higher-order listening abilities. Cai (2012) believes that the partial dictation test is comparable to a gap fill test in terms of the abilities being measured.

Cai (2012) used partial dictation as a final measure of achievement to ascertain if students had successfully learned the course material, whereas in the current study it was used as a measure of proficiency. Although dictation can be used in standardized testing, the purpose was to construct a norm-referenced test that would allow the teacher to construct groups that would allow for a mix of proficiencies within each group to facilitate learning within the ZPD. All of the papers discussed above suggest that dictation is a somewhat useful and reliable measure of language proficiency with moderate correlations with overall scores on batteries of tests (Oller & Streiff, 1975), and with the overall TOEFL score (Fouly &

Cziko, 1985). However, in order to be useful for the purposes of group construction the test must differentiate between the differing proficiencies within a given group. After describing the context and the dictation test, we present the results of a Rasch analysis of the data. Rasch analysis will enable us to determine if the items are effectively measuring the ability of the candidates in this study, and highlight any problematic items in the test. Following this, Rasch measures for people are correlated with two established measures of English proficiency.

Research Questions

The current study sought to determine whether an in-house dictation test would be effective in differentiating between the English proficiency levels of a mixed-ability group of students. The study also investigated the correlation between the dictation and established measures of English language proficiency. The research questions are:

1. Does the dictation test successfully measure the range of proficiencies of students in this context?
2. Does the dictation test have a strong correlation with established tests of English proficiency, the TOEIC test and the SLEP (Secondary Level English Proficiency) test?

The Study

Participants

The participants in the current study were 146 first year university students from six intact classes at a private university in Western Japan. The students were majors in science in a department of science and technology and were pre-intermediate in terms of English proficiency (average TOEIC 390). The students were taking compulsory English courses in reading, communication and writing and were streamed according to their major within the department, not their English proficiency. English is included as part of the entrance exam to the university so students have achieved a basic level of proficiency, but within a given class there was a large range of English abilities. In the most extreme case in this study there was a student who had lived abroad for several years and had a TOEIC score of 895, while the lowest score in the same class was 220. The data was collected over a two-year period with two consecutive cohorts, although TOEIC scores were only made available for the first year (n=68). For reasons of privacy, the university decided to

restrict access to the students' TOEIC scores in the second year, but we did have access to scores from a short version of the SLEP test as a measure of English proficiency (n=78), administered as part of an unrelated research project.

Dictation

A dictation passage was constructed and piloted with the students to ensure that they were comfortable with the format, and that the level was appropriate. The pilot dictation also used the simple past, in line with the dictation used in the current study. The dictation for the study was constructed and administered following guidelines provided by Oller et al. (1975). The first author wrote the dictation and attempted to make it an appropriate difficulty level for the students, based on his 18 months experience teaching at the university, and more than ten years teaching experience in Japan. The aim was to make the dictation at a level appropriate for measuring the range of proficiencies of the students in this context. The dictation was 93 words in length (see Appendix A), and the vocabulary was checked to assess the level. A VocabProfile analysis using the Compleat Web VP! Lextutor routine (Cobb, n.d.) showed that only 7 words were outside the most frequent 2,000 words of English based on the New General Service List (Browne, Culligan, & Phillips, 2013). These were content words, many of which were used in Japanese as loan words (*Japan, lion,* and *panda* are all used in Japanese). The remaining words (*zoo, paddling*, *dolphin*, and *sand*) were all considered reasonably simple for these students. Past-tense narrative was chosen as this is reasonably simple and enabled the vocabulary level to be kept quite low.

Again, following directions from Oller et al. (1975), the dictation was recorded twice; once at normal speed and once with pauses to allow the students time for writing. Total time for administration was approximately 15 minutes including explanation. The dictation was read at natural speed, and the pauses were at phrasal boundaries to maintain the meaning of the text, but to allow students time to write (see Appendix A for phrasal boundaries). The dictation was read by the first author, who is from the U.K. and speaks standard British English with a northern accent. Instructions were given in the students' first language, and students were given a chance to ask questions before the test began. Instructions were also written on the board in English for confirmation. The test audio was played once at normal speed, once with pauses, and then finally at normal speed once again. Students were told not to write anything the first time, but to just listen, and were encouraged to complete the dictation on the second playing. Only one student failed to understand

the instructions and did not complete the dictation, but changed the words to a report in the third person. Her paper was not considered in the subsequent analysis.

Following on from the test administration, student papers were copied, and a norming session with the two authors was conducted in order to establish guidelines for marking. Issues such as spelling, word-order, and additional words not in the original text were considered and several papers were marked together in order to discuss any issues that arose. The decision was taken to allow for phonetic representations and not to be strict regarding spelling as we did not want the test to become a spelling test. Japanese learners typically struggle to distinguish between L/R sounds and therefore this was not penalized, as again it was felt this would become more of a spelling test and not representative of speaking proficiency. Following the norming session papers were marked independently by the two authors, and scores were inputted into a Microsoft Excel spreadsheet. A simple binary code was used to differentiate correct and incorrect responses. After all the papers had been marked, the two raters met again and all discrepancies in marking were discussed. Inter-rater reliability was very strong with a value of .98, and there were very few issues that arose with marking. The raw scores from the dictation test were then used for Rasch analysis, and subsequent logit scores derived from the Rasch analysis were used for the correlation analysis with the TOEIC and SLEP test scores. Rasch analyses were conducted using the software Winsteps (Linacre & Wright, 2007), and correlational analyses were conducted using SPSS 19.0 (IBM).

Data analysis

Previous studies have shown that dictation measures language proficiency, but none have used Rasch analysis to analyze the results of a dictation test. A test does not have inherent difficulty, and difficulty is determined by the test takers (Bond & Fox, 2007). Rasch analysis enables the researcher to ascertain the difficulty of the test in relation to the population being tested, and therefore will enable us to determine if the test constructed is of a suitable difficulty level for this population of students. Rasch analysis also acknowledges that some words in the dictation will be more difficult than others, and avoids the use of raw scores, instead giving test takers a logit measure which factors in the different difficulty of each of the items (Bond & Fox, 2007). In this analysis we use logit scores, placing students on an interval scale, and accounting for differences in item difficulty.

The dictation test scores derived from the Rasch analysis were correlated with two established measures of English proficiency. The TOEIC test is very popular in Asia and is used extensively as a measure of English proficiency in Japan (see

https://www.ets.org/toeic for more details). It was created as a test of business English, and is now used extensively at universities and even high schools in Japan, with many companies requesting TOEIC scores from potential employees. The test is two hours in length, and includes a listening section, and reading section. Although described as reading, many of the questions in this section have a strong focus on grammar and vocabulary so it is not solely a test of reading ability. All students at the university took the TOEIC test, but in the second year of our data-collection the administration at the university decided that student scores would no longer be available to teachers. In the second year of the study, as part of an unrelated research project, students took a shortened 30-minute version of the SLEP test, which has since been discontinued, and replaced with the TOEFL (Test of English as a Foreign Language) Junior test. This test was reading only, and designed to measure a wide range of proficiencies in a short time.

Results and discussion

There were two questions considered when analyzing the dictation test: Did the test measure the range of proficiencies in the student sample, and was it measuring English ability? We used Rasch analysis to answer the first question, and correlation analysis to answer the second. After reviewing the descriptive statistics, we present the results of these analyses.

Descriptive statistics

First, we consider the descriptive statistics for the three tests to ascertain normality of distribution. Table 1 shows the descriptive statistics for the three proficiency tests. The results for the dictation are in logits attained from the Rasch analysis, and show that the distribution is positively skewed with positive kurtosis. Case 65 was an outlier on this variable, with a standardized Z score of 4.65, which is greater than the 3.29 benchmark provided by Field (2009). This student was a returnee who had lived abroad for a considerable length of time and therefore, although part of the population, was highly proficient in English. Removal of this student's data results in a normal distribution for the dictation test scores. The results of the TOEIC test show that again the distribution is positively skewed and has positive kurtosis. Again case 65 was an outlier on this variable, with a TOEIC score of 895 and a standardized Z score of 4.11, exceeding the benchmark of 3.29 (Field, 2009). As with the dictation, removal of this student results in a normal distribution for the TOEIC test scores. The results of the SLEP test show that the distribution is normal.

Table 1. Descriptive Statistics for the Proficiency Tests

	Dictation Test	TOEIC Test	SLEP Test
M	1.63	387.89	19.81
SE	.10	14.82	.39
95% CI	[1.42, 1.83]	[358.27, 417.51]	[19.02, 20.59]
SD	1.30	118.58	3.48
Skewness	.71	1.51	.20
SES	.19	.30	.27
Kurtosis	1.81	3.66	.33
SEK	.38	.59	.54

Rasch analysis of the Dictation test

In order to comprehensively assess the dictation test, we must consider if it follows the principles of measurement central to the Rasch model. According to the principles of measurement in the Rasch model the test should be a unidimensional measure of a construct. Each individual item must be independent of other items and must fit the model, and the two fit statistics most readily reported by researchers are the infit and outfit mean square. The infit mean square is sensitive to the response pattern of individuals around their level of ability, while outfit mean square reflects how an individual responds to items that are either considerably above or below their actual ability. The ideal value for both measures is 1, but they can range from zero to positive infinity. Infit or outfit values of less than 1 imply that the person or item is overfitting the model, while values over 1 suggest that the person or item is underfitting the model. Wright and Linacre (1994) suggest a range of .7 to 1.3 as being appropriate for a test like the one in this study. Standardized infit and outfit statistics can have positive values indicating greater variation than suggested by the Rasch model, or negative values indicating less variation than expected. The ideal value is 0 with a standard deviation close to 1. Again Bond and Fox (2007) provided acceptable ranges for standardized infit and outfit statistics as between -2.0 and 2.0.

Rasch analysis offers two indices of reliability. The person reliability index indicates the degree to which the items separate the participants so that some perform at high levels and others at low, which means the measure must effectively distinguish between differences within the given population on the construct of interest. Values for person reliability range from 0 to 1 and are analogous to Cronbach's alpha, with values closer to 1 indicating higher reliability. Interpretation of reliability statistics follows that of Cronbach alpha with values over .90 indicating strong reliability, values over .80 indicating good reliability, and values over .70 indicating acceptable reliability (Sheridan & Puhl, 1996, p. 26).

The item reliability index indicates the replicability of item separation should the items be given to a different, but equivalent group of respondents. That is, would the items still behave in the same way with a similar sample of the population? High reliability suggests that the difference in difficulty is consistent and the researcher can be confident that the item difficulties are reasonably stable. Again, values range from 0 to 1 and follow the same principles as for person reliability.

The Rasch person separation index is a more sensitive measure of how well the items are able to differentiate between the respondents for the given variable. This index is not bound by 1 and values range from zero to infinity, with high values indicating that the measure is effective in distinguishing between the respondents. Generally, values greater than 2.0 are regarded as providing acceptable separation. Measures with low person separation are not measuring differences in the sample and therefore are of limited use. This is of relevance to the first research question, considering the ability of the test to differentiate between the proficiency of the test-takers in this context.

In a similar manner, the Rasch item separation index is an indicator of how well spread or separated the items are on the given variable. Again, values range from zero to infinity and high values indicate that the items are well spaced in terms of difficulty, with values greater than 2.0 being acceptable (Bond & Fox, 2007). As with person separation, it is important that the measures developed cover a wide range of abilities to avoid floor and ceiling effects.

The Rasch analysis followed the steps outlined by Tennant and Conaghan (2007). The model used for the current analysis was the dichotomous model. As stated previously acceptable range for infit and outfit mean square values was 0.7-1.3 (Wright & Linacre, 1994), as the test was not high stakes, but also not survey data. The item fit statistics for the first analysis are shown in Table 2. The items were assessed and deleted sequentially, beginning with the worst fitting item. After each deletion the item-fit statistics were reassessed. In this case that meant deleting, in order, the following items: 40 *went*, 75 *panda*, 25 *was*, 33 *did*, 18 *spent*, 69 *to*, 42 *the*, 41 *to*. Following deletion of these items the infit mean square values were found to be within the acceptable range. Standardized values were checked and although outfit values were within acceptable ranges, the standardized infit values revealed three items with high positive values. These items were sequentially removed from the analysis starting with the item with the highest infit value. Items were removed in the following order: (34 *take,* 60 *sand*, 23 *members*), and following this all items fit the model. Following Wright and Linacre (1994), low negative values show redundancy

but do not affect the quality of measurement, and therefore these items were therefore not removed.

Table 2. Rasch Item Statistics for the Dictation Test Items

Item	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
30 The	3.47	.22	0.98	-0.1	0.93	-0.3	0.4
35 About	3.33	.21	1.1	0.8	1.1	0.5	0.31
76 Lions	2.96	.20	1.06	0.7	0.99	0	0.36
33 Did	2.92	.19	1.29	2.9	1.57	3.1	0.14
21 Other	2.81	.19	0.98	-0.1	1.04	0.3	0.41
27 For	2.77	.19	1.09	1.1	1.1	0.7	0.33
70 A	2.73	.19	0.91	-1.1	0.87	-0.9	0.49
74 A	2.73	.19	0.88	-1.5	0.82	-1.3	0.51
18 Spent	2.70	.19	1.3	3.4	1.56	3.5	0.14
29 Of	2.66	.19	0.94	-0.8	0.95	-0.3	0.46
62 Paddling	2.52	.18	1.07	0.9	1.08	0.7	0.36
58 In	2.36	.18	1.11	1.6	1.19	1.6	0.31
3 A	2.26	.18	1.04	0.6	1.05	0.5	0.38
60 Sand	2.13	.18	1.17	2.4	1.17	1.6	0.29
81 Was	1.76	.18	0.81	-3.1	0.76	-2.7	0.58
42 The	1.69	.18	1.3	4.1	1.46	4.2	0.16
82 A	1.66	.18	0.73	-4.4	0.66	-4	0.64
28 Most	1.63	.18	0.91	-1.4	0.86	-1.4	0.5
78 Other	1.44	.18	1.03	0.5	1.04	0.4	0.39
59 The	1.41	.18	1.06	0.8	1.06	0.6	0.37
23 Members	1.38	.18	1.12	1.6	1.2	1.9	0.31
20 With	1.28	.18	0.96	-0.6	0.94	-0.5	0.45
73 Saw	1.19	.18	1.06	0.8	1.08	0.7	0.36
34 Take	1.12	.18	1.22	2.7	1.23	1.8	0.24
80 There	1.12	.18	0.8	-2.8	0.71	-2.7	0.58
63 In	1.09	.18	0.99	-0.1	0.96	-0.3	0.42
38 Off	.95	.19	0.97	-0.4	0.89	-0.8	0.44
9 Relaxing	.92	.19	0.85	-1.9	0.8	-1.6	0.52
56 Enjoyed	.92	.19	1.04	0.5	1.04	0.3	0.36
50 It	.88	.19	1	0	0.98	-0.1	0.4
64 The	.74	.19	0.82	-2.2	0.74	-1.9	0.54
19 Time	.59	.19	1.01	0.1	0.94	-0.3	0.39
10 With	.55	.20	0.93	-0.7	0.83	-1	0.45
7 This	.52	.20	0.97	-0.3	0.95	-0.3	0.4
26 Working	.36	.20	1.04	0.4	0.95	-0.2	0.35
36 One	.31	.20	0.93	-0.6	0.79	-1.1	0.44
49 And	.31	.20	1.06	0.6	1.03	0.2	0.32
6 Time	.21	.21	0.81	-1.8	0.67	-1.9	0.53
2 Had	.23	.21	1.02	0.2	0.99	0	0.34
57 Playing	.23	.21	0.96	-0.3	0.87	-0.6	0.4
37 Week	.05	.22	0.93	-0.5	0.79	-0.9	0.41

75 Panda	.05	.22	1.33	2.5	1.73	2.8	0.03
41 To	.00	.22	1.23	1.8	1.33	1.4	0.15
69 To	-.05	.22	1.11	0.9	1.43	1.7	0.22
86 Was	-.05	.22	0.86	-1.1	0.72	-1.2	0.46
14 Stayed	-.10	.22	1.04	0.4	0.98	0	0.31
84 Show	-.10	.22	1	0.1	0.82	-0.7	0.36
91 A	-.10	.22	0.96	-0.3	0.76	-1	0.39
22 Family	-.20	.23	1.04	0.3	0.88	-0.4	0.32
17 And	-.25	.23	1.1	0.7	1.06	0.3	0.25
85 Which	-.31	.24	0.98	-0.1	0.87	-0.4	0.35
51 Was	-.37	.24	0.92	-0.5	0.78	-0.7	0.39
65 Sea	-.37	.24	0.9	-0.6	0.72	-1	0.41
5 Nice	-.49	.25	0.82	-1.1	0.55	-1.7	0.47
32 But	-.49	.25	1.01	0.1	1.05	0.3	0.29
79 Animals	-.49	.25	1.03	0.2	0.86	-0.4	0.3
83 Dolphin	-.55	.25	0.92	-0.4	0.88	-0.3	0.35
45 My	-.68	.26	1.01	0.1	0.84	-0.4	0.29
47 And	-.76	.27	0.95	-0.2	0.87	-0.3	0.31
77 And	-.83	.28	1.05	0.3	0.94	-0.1	0.24
4 Really	-.91	.29	0.83	-0.8	0.56	-1.3	0.41
44 With	-.91	.29	1.06	0.4	1.08	0.3	0.22
46 Wife	-.91	.29	0.99	0	0.92	-0.1	0.28
25 Was	-.99	.29	1.19	0.9	1.58	1.4	0.07
68 Went	-1.18	.31	0.9	-0.3	0.82	-0.3	0.31
11 My	-1.28	.32	0.98	0	0.67	-0.7	0.27
53 Fun	-1.28	.32	1.02	0.2	0.64	-0.8	0.26
72 And	-1.28	.32	1.02	0.2	0.82	-0.3	0.24
88 Amazing	-1.28	.32	0.9	-0.3	0.58	-1	0.33
87 Really	-1.39	.34	0.97	0	0.77	-0.4	0.27
67 Also	-1.51	.35	0.91	-0.2	0.45	-1.2	0.33
31 Summer	-1.64	.37	1.03	0.2	0.78	-0.3	0.2
92 Great	-1.96	.43	0.98	0.1	0.55	-0.7	0.24
13 We	-2.16	.46	0.97	0.1	0.73	-0.2	0.19
52 Great	-2.16	.46	1	0.1	1.09	0.4	0.13
66 We	-2.16	.46	0.93	0	0.53	-0.6	0.24
55 daughter	-2.39	.51	0.9	-0.1	0.81	0	0.22
90 Had	-2.39	.51	0.98	0.1	0.97	0.2	0.15
61 And	-2.70	.59	0.95	0.1	0.89	0.1	0.16
1 I	-3.12	.72	0.99	0.2	0.5	-0.3	0.14
12 Family	-3.12	.72	1.03	0.3	0.94	0.3	0.07
15 In	-3.12	.72	0.93	0.1	0.39	-0.6	0.19
40 Went	-3.12	.72	1.05	0.3	2.09	-1.2	0.02
48 Daughter	-3.12	.72	0.86	0	0.13	-1.2	0.27
71 Zoo	-3.12	.72	0.97	0.2	0.28	-0.8	0.19
8 Summer	-3.83	1.01	1.01	0.3	0.53	-0.2	0.09
43 Beach	-3.83	1.01	1.02	0.4	1.14	0.5	0.03
16 Japan	-5.04	1.83	Min	Min	Min	Min	Min

24 I	-5.04	1.83	Min	Min	Min	Min	Min
39 I	-5.04	1.83	Min	Min	Min	Min	Min
54 My	-5.04	1.83	Min	Min	Min	Min	Min
89 I	-5.04	1.83	Min	Min	Min	Min	Min
93 Summer	-5.04	1.83	Min	Min	Min	Min	Min

To investigate the dimensionality of the test, a Rasch PCA of item residuals analysis was run. The results showed that 39.8% of the variance (eigenvalue = 50.9) was explained by the Rasch model, 24.4% of the variance (eigenvalue = 31.2) was explained by the items, and 2.5% of the variance (eigenvalue = 3.3) was explained by the first residual contrast. The raw variance explained by the Rasch measures fell short of 50%, but this is a weak measure of unidimensionality, and the unexplained variance in the first contrast is below the 5% criterion specified by Linacre (2007). The variance explained by items is greater than four times the unexplained variance in the first contrast (Linacre 2007), and therefore it was concluded that the measure was unidimensional. An additional means of determining that the test is unidimensional is to check the loading of items on the non-Rasch measures dimensions (the contrasts from the PCA of the standardized residuals). Loadings above ±.40 are considered to be strong, and suggestive of different factors. There were five items with positive loadings above .40. These items were 62 *paddling* (.43), 63 *in* (.51), 64 *the* (.43), 65 *sea* (.48), and 5 *nice* (.41). There were no items with a negative loading above -.40. The five items with strong positive loadings were sequential suggesting that there might have been some feature of this sequence that made it different from the rest of the dictation. Generally, the item loadings support the fact that the dictation test is a unidimensional measure of language proficiency, with only five items from the 93 item test showing strong loadings.

Rasch analysis requires that each item be completely independent of the other items, and with a dictation test there was concern that ability to write words would be linked as students remember chunks of language rather than individual items. In order to check that the assumption of independence was valid, the item correlation of standardized Rasch residuals for all possible pairs of items was checked. Item correlation analysis revealed that the item correlations were weak, and within the range of -0.3 to +0.3. Only one correlation was outside of this range, with items 1 *I,* and 66 *we,* showing a moderate correlation of 0.57. These results indicate that the assumption of local independence central to the Rasch model has been satisfied.

Although Tennant and Conaghan (2007) recommend conducting a DIF analysis, the only potential variable to consider was gender, and as this was not hypothesized to influence listening ability for individual items it was not conducted. Instead, we proceeded to analyze the targeting of the items to the target population.

Research question 1 asks if the test is able to measure the range of proficiencies of students. Therefore, the separation figures and Wright map provide the data for the claim that the test can be used to make decisions regarding assigning students to different groups (Bachman & Palmer, 2010). Figure 2 below shows the Wright map for the items and people and shows the distribution of people abilities and item difficulties.

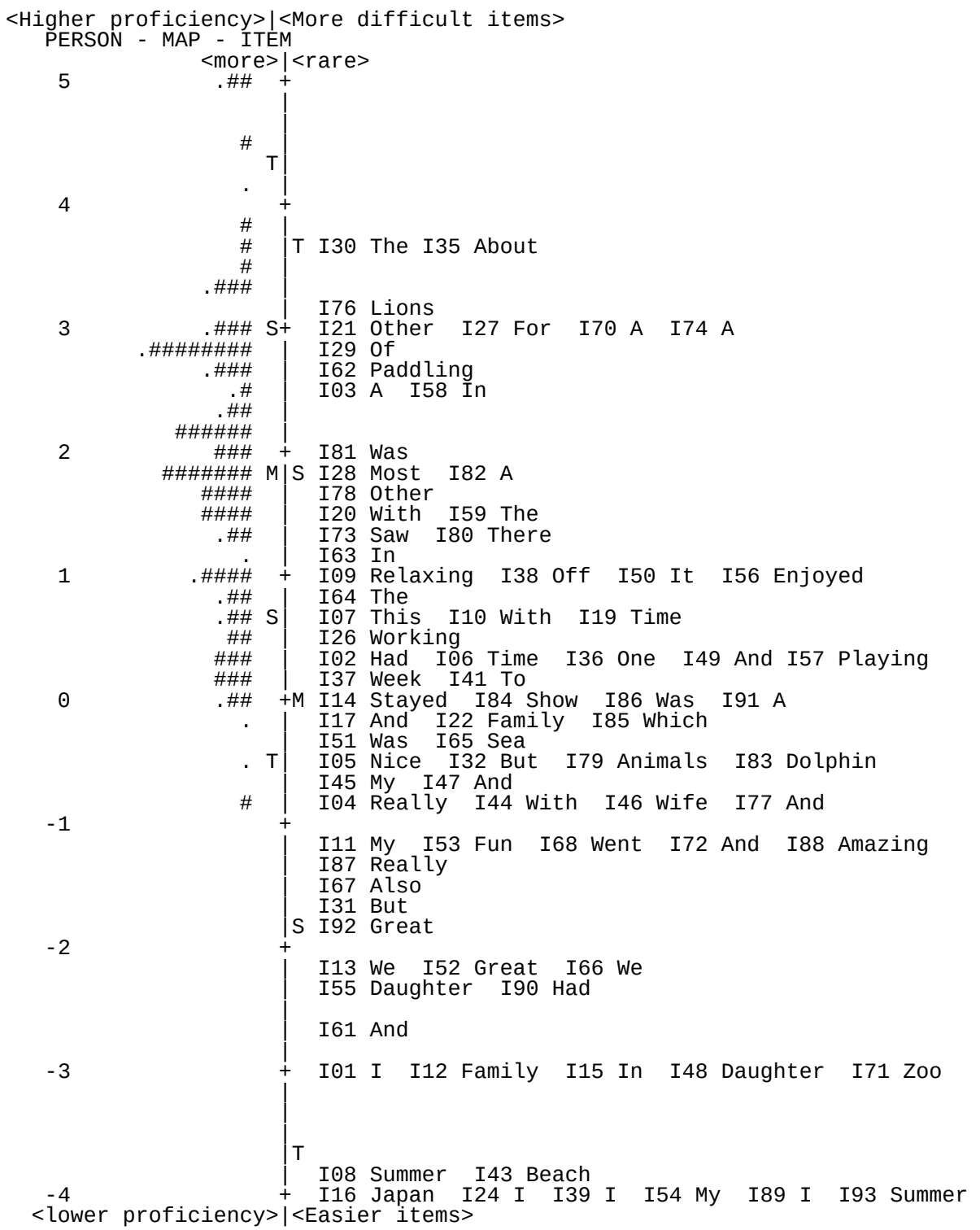

Figure 2. Wright map for the dictation test items.
M = Mean; S = 1 SD; T 2 SDs. Each X = 1 person, each # = 2 people.

As can be seen from figure 2 the mean for the people (1.96) is considerably higher than the mean for the items (.0), indicating that the dictation test was relatively easy for this group of students. There were six items at the bottom of the scale (6 *Japan*; 24 *I*; 39 *I*; 54 *My*; 89 *I*; 93 *Summer*) that were correctly transcribed by all students, and therefore did not help in the measurement of proficiency for these students. Four of these words were the first word in a sentence and therefore were highly salient to students, while *summer* was the final word in the dictation. These results support VanPatten's sentence location sub-principle from his Input Processing theory (VanPatten, 2004, p. 13). VanPatten (2004, p. 14) states that "learners tend to process items in sentence initial position before those in final position and those latter in turn before those in medial positions." *Japan* is easy in this context and therefore was highly salient.

In regards to research question 1, the test seemed to have items that measured the range of the students effectively, although there were a number of items that were easy for these students. 29 items were classified as having a difficulty below that of the lowest ability student. In order to refine the measurement model, the 6 items that were rated as minimal measure were removed. Following removal item separation was 4.98 (reliability .96) and person separation was 2.89 (reliability .89). Even after deletion of these items there were still a large number of items below the lowest measure for persons. Indeed, there was a sudden jump in item difficulty from -3.09 to -2.67, and therefore to remove a further level of redundancy, six more items with logit values below -2.67 were removed from the analysis. This meant removal of items 1 *I*, 8 *summer*, 12 *summer*, 15 *in*, 43 *beach*, 48 *daughter*, 71 *zoo*. These items are largely redundant in terms of measurement. Following their removal item separation improved considerably with only a minor loss in person separation. Item separation was 5.62 (reliability .97), and person separation 2.87 (reliability .89). This is a good level of person separation and means that the test is useful for making the decisions necessary to form groups, therefore providing evidence to support the AUA provided above (Bachman & Palmer, 2010).

Items that proved difficult for students were the definite and indefinite articles, with *the* proving to be the most difficult word, and *A* often featuring at the top end of the item difficulty scale. Articles are often not clearly enunciated in speech, and prove very difficult for most learners of English, and the knowledge of grammar will dictate whether an article is necessary. Their difficulty supports the theory posited by Oller and Streiff (1975) that in some way expectancy grammar is measured by dictation. The results again offer evidence in support of VanPatten's Input Processing Theory (2004), which claims that content words are processed before largely redundant grammatical items such as articles.

The Wright map shows that there was a certain degree of redundancy in the test, with a number of items measuring the same difficulty level, and also items that were below the ability of even the weakest student. This may be a facet of dictation, in that some words will inevitably be salient and easy for all of the candidates, particularly content words, and words at the start of a sentence (VanPatten, 2004). With a test of this nature, if all items were at a level that was optimum for measurement, then the test would be deemed to be very difficult by the students and may cause students to stop attempting to complete the test. Particularly considering that the test was zero-stakes for the students, it is important that students feel that they can understand and transcribe a reasonable amount of the text, and the current test seems to have therefore been at an appropriate level for these students, as it provides the information necessary to make the decision regarding group assignment as part of the AUA.

At the top of the person ability scale, there was one student who achieved a perfect score and therefore her proficiency was not measured by the dictation test. This student was a returnee, having spent more than six years of her childhood living in the USA and with a TOEIC score of 895. During an informal interview after the dictation test she stated that she had found the test very easy. Eight of the students had ability above the most difficult item, showing that these people are not being precisely measured by this dictation test. This shows a possible limitation of dictation, and suggests that graduated dictation as used by Fouly and Cziko (1985), may be more effective in measuring a wide range of abilities, and avoiding ceiling effects that were present in this study. It should be noted that the student described above who achieved a perfect score was an outlier according to the criteria provided by Field (2009), and aside from her no other students were able to achieve full marks on the test.

In conclusion, the Rasch analysis shows that the dictation test was able to measure the range of proficiencies in this context, and provide good separation for the person measure. There was a degree of redundancy in terms of items that were not adding to the measurement model, but the data can be used to make decisions regarding group assignment as part of the AUA (Bachman & Palmer, 2010).

Correlation with established measure of English proficiency

The Rasch analysis supports the conclusion that the dictation is a reliable measure, effectively able to differentiate between the different proficiency levels of students in this study and therefore useful to make the decision based on the assessment. In order to make the claim that students are being separated according to language

proficiency as is needed in the interpretation of scores for the AUA, we need data that supports the warrant that dictation is in fact a measure of English proficiency. To provide the data necessary to support this claim, a bivariate correlation analysis was performed with the person logit scores from the Rasch analysis and the listening and reading sections of the TOEIC test, the overall TOEIC score, and the SLEP test. As stated previously, although all the students in the study had taken the TOEIC test, the scores were only available for the first cohort due to an administrative decision to withhold scores for subsequent cohorts. The reading section of the SLEP test was given to the second cohort and used as an established measure of proficiency for these students. A correlation analysis was performed with the dictation scores and these tests. The results of the correlation for TOEIC are shown below in Table 2. The dictation test correlated with the SLEP test at .61, significant at $p < .01$ (2-tailed).

Table 3. Correlations for the Dictation and TOEIC Test

	1. DICT	2. T-LIST	3. T-READ	4. T-TOTAL
1. DICT	—			
2. T-LIST	.76	—		
3. T-READ	.72	.69	—	
4. T-TOTAL	.80	.92	.92	—

Note. DICT = Dictation test; T-LIST = TOEIC listening; T-READ = TOEIC reading; T-TOTAL is total TOEIC SCORE. All correlations significant at $p < .01$ (2-tailed).

The results in Table 3 somewhat support the argument that the dictation test is a test of English proficiency, with moderate and significant correlation with both parts of the TOEIC test. As would be expected, the listening part of the TOEIC correlated more strongly with the dictation than the reading part. Of interest is the fact that the overall TOEIC score correlates more strongly with the dictation than individual parts of the test. This supports the claim by Oller and Streiff (1975) that the dictation test also incorporates grammar, as students rely on their knowledge of grammar to complete the parts of the dictation that they were unable to hear or to remember. As mentioned previously, although described as "reading," the TOEIC actually comprises a large number of questions related specifically to grammar. The correlation helps support the argument that dictation is a measure of language proficiency, and warrants the interpretation of scores that is postulated as part of the AUA, although correlation analysis alone is insufficient to make strong claims.

Although the correlation of the dictation with the SLEP is moderate and significant, it is weaker than the correlation with the TOEIC test. The students in the current study were given a shortened 30 minute version of the SLEP test. The shortened

version of the test was only reading, with no listening component, which may also explain the weaker correlation with the dictation test.

Conclusion

There can be little doubt that small group work is central to most language teaching pedagogies. Based on most of the research investigating peer interaction, language proficiency has been found to have a large impact on the nature of the interaction (Philp et al., 2013), and therefore teachers should be interested in the relative proficiencies of students within their classes, and consider this when making groups. A norm-referenced proficiency test is needed in order to effectively construct balanced groups, and the test must effectively differentiate between the different English proficiencies of students within a given class. The current study set out to determine if a dictation test, made in-house by the first author, could effectively differentiate between the proficiency of university students in a mixed-ability class in a Japanese context.

Rasch analysis showed that the test was able to differentiate between the students in this context, despite students having a range of TOEIC scores from 220 to 895. The test did have a ceiling effect, with the most capable student in the class achieving a perfect score on the test (TOEIC 895), and therefore her proficiency was not measured. The test was reasonably easy for the students, and the average student ability was above the average difficulty of the items, although this would be expected for a test of this nature. There was also a degree of redundancy, with six items being too simple and not measuring the students, and a large number of items with the same difficulty. Clearly with a graduated dictation the test can be adapted to cover more advanced students, but redundancy will remain. Partial dictation, as described by Cai (2012), may be most effective in avoiding this redundancy in that the highly salient words can be provided, with only more difficult items used for dictation, but this test may be quite different, as students are not required to use their knowledge of grammar to recreate full sentences. A Rasch analysis of a partial dictation would be useful in examining the degree to which redundancy can be reduced. Despite the redundancy, the test provided the necessary data to make the decision that was part of the AUA, in terms of separating the students based on English proficiency, and can therefore be considered successful.

In conclusion, the in-house dictation test was simple to construct, administer and grade, and produced results that allowed for the decisions necessary for group construction based on the proficiency of students in this context. The test was also free, and although in the current context students are all required to take the TOEIC

test, there are many places where the cost of testing is prohibitive and yet teachers would like to know the relative proficiency of their students. Particularly in these situations a simple dictation test will allow teachers to differentiate between the different proficiency levels of students within their classes. The moderate correlation with the TOEIC test helps to support the interpretation that the dictation test is a measure of overall English proficiency, including grammar, reading, and listening, and is therefore useful as a general measure.

Returning finally to the concept of Assessment Use Argument introduced earlier (Bachman & Palmer, 2010), the consequences of making mixed groups of proficiency seem to be for the benefit for all students. The decision that the test is being used for seems reasonable, although the interpretation of the dictation scores as a reliable measure of how the students will perform in the oral English class may be a little difficult. This is the weakest part of the AUA model, in that dictation does not involve the spontaneous production of language that is necessary in conversation. Ultimately, however, the test is being used for a beneficial outcome for all students, and the data supports the claims made to a degree. Based on this, the AUA can support the use of dictation to assign students to mixed proficiency groups in this context.

The study highlights several limitations of using dictation as a means of testing language proficiency. The dictation test does require some knowledge of the ability of students in order to construct a test that is appropriate to the level of the students. In the current context, the first author had 18 months of experience with the students, and more than ten years teaching at various levels in Japan. The test constructed was effective in separating students, but more extensive piloting of the test would have made it more appropriate in terms of difficulty. Also inexperienced teachers or teachers with limited experience in a particular context may struggle to construct an effective test. Somewhat related to this, although a simple dictation test is easier to construct than a graded dictation, there is the risk of a floor or ceiling effect, and it is challenging to construct a short 100 word dictation that will measure a very wide range of proficiencies such as were present in the current study. As Beglar (2009) states, it should be remembered that this paper has presented a dictation test used in a specific context, and test validity and reliability are not transferable. As such, more research investigating the use of dictation in varied contexts and with different levels of student, both in terms of proficiency and age, would be beneficial. Furthermore, we recommend the use of Rasch analysis which treats each item as having different difficulty, and also converts the students' scores to an interval scale.

References

Bachman, L. F. & Palmer, A. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Beglar, D. (2009). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd Edition). Mahwah, NJ: Routledge.

Browne, C., Culligan, B. & Phillips, J. (2013). *The New General Service List.* Retrieved from http://www.newgeneralservicelist.org.

Cai, H. (2012). Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing 30*(2), 177–199.

Cobb, T. *Compleat Web VP!* [computer program]. Accessed February 16, 2016, at http://www.lextutor.ca/vp/comp/

Cohen, E. G. (1994). *Designing Groupwork: Strategies for the Heterogeneous Classroom*. New York, NY: Teachers College Press.

Field, A. P. (2009). *Discovering statistics using SPSS*. London: Sage.

Fouly, K. & Cziko, G. (1985). Determining the reliability, validity, and scalability of the graduated dictation test. *Language Learning, 35*(4), 555–566.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. Mahwah, NJ: Routledge.

Howatt, A. P. R., & Widdowson, H. G. (2004). *A History of English Language Teaching*. Oxford: Oxford University Press.

Jacobs, G. M. **(**2006). Issues in implementing cooperative learning. In S. G. McCafferty, G. M. Jacobs, and A. C. DaSilva Iddings (Eds.), *Cooperative Learning and Second Language Teaching* (pp. 30–46). New York, NY: Cambridge University Press.

Lantolf, J. P. (2006). Sociocultural Theory and L2: State of the Art. *Studies in Second Language Acquisition, 28*(1), 67–109.

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

Linacre, J. M., & Wright, B. D. (2007). WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis [Computer software]. Chicago, IL: MESA.

Long, M. H. & Porter, P. A. (1985). Groupwork, Interlanguage Talk, and Second Language Acquisition. *TESOL Quarterly, 19*(2), 207–228.

Oller, J. (1971). Dictation as a Device for Testing Foreign-Language Proficiency. *ELT Journal, 25*(3), 254–259.

Oller, J. & Streiff. V. (1975). Dictation: A test of grammar-based expectancies. *ELT Journal, 4*(1), 37–41.

Philp, J., Adams R., & Iwashita, N. (2013). *Peer Interaction and Second Language Learning.* New York, NY: Routledge.

Rost, M. **(**2005). L2 Listening. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 503-527). Mahwah, NJ: Lawrence Erlbaum.

Sheridan, B., & Puhl, L. (1996). Evaluating an indirect measure of student literacy competencies in higher education using Rasch measurement. In G.Engelhard & M. Wilson (Eds). *Objective measurement: Theory into practice, Volume 3 (pp. 19-44).* Norwood, NJ: Ablex.

Stansfield, C. (1985). A history of dictation in foreign language teaching and testing. *The Modern Language Journal*, *69*(2), 121–128.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.). (2005). *Handbook of research in second language teaching and learning.* Mahwah, NJ: Lawrence Erlbaum.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis & Rheumatism, 57*(8), 1358-1362.

VanPatten, B. (Ed.). (2004). *Processing Instruction: Theory, Research, and Commentary*. London:Taylor and Francis.

Vygotsky, L. S. (1986). *Thought and Language*. Cambridge, MA: MIT Press.

Weir, C. J., Vidaković, I., & Galaczi, E. (2013). *Measured constructs: A history of Cambridge English lanugage examinations 1913-2012*. Cambride: Cambridge University Press.

Wright, B. D., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Appendix A: Dictation test

I had a really nice time this summer / relaxing with my family. / We stayed in Japan / and spent time with other family members. / I was working for most of the summer / but did take about one week off. / I went to the beach with my wife and daughter / and it was great fun. / My daughter enjoyed playing in the sand / and paddling in the sea. / We also went to a zoo / and saw a panda, lions and other animals. / There was a dolphin show / which was really amazing. / I had a great summer.

1. **Listen to recording A-natural speed without pauses.**
2. **Listen to recording B and write down-natural speed with pauses in between clauses.**
3. **Listen to recording A and check-natural speed without pauses.**

(/ marks a pause in the recording for the second delivery)