Using Dictation to Measure Language Proficiency

Aeric Wong
Konan University, Japan

Paul Leeming
Kinki University, Japan

Abstract

Many teachers are faced with mixed-ability classes and have little information on the proficiency of their students. With group work being central to most pedagogies in second language acquisition, teachers may want to construct groups for specific purposes; knowledge of the relative proficiency of students is therefore very important. Although tools such as the Test of English for International Communication (TOEIC) are available, they may be expensive to implement, or results may be unavailable to the teacher. This paper discusses dictation tests as a possible way of measuring the relative proficiency of students and highlights the results of a study conducted in a university in Japan that used dictation as a test of proficiency and compared results with the TOEIC test. The authors explain how to design and conduct dictation in the classroom and demonstrate how dictation is a cheap, simple, and effective means of measuring language proficiency.

Group Work in the Classroom

A large number of language teachers regularly assign group work in the classroom. The various reasons for assigning small-group tasks may include adding interest, maximizing student talk-time (Long & Porter, 1985), encouraging and fostering a collaborative work ethic among peers, and demonstrating how learners can learn from each other through co-constructed scaffolding (Leeming, 2011). One possible problem with the use of small groups in the language classroom is that language teachers frequently have mixed-ability classes and so must consider the English proficiency of students when deliberately constructing groups. Often teachers do not have information regarding the relative proficiency of students within a given class, and in large educational contexts it is quite common to have students spanning a wide range of language proficiencies which therefore requires more deliberate consideration when attempting to create well-balanced groups. Factors such as linguistic ability and personality can affect the performance of the group as a whole, and often the aim is to achieve a balance within each group. Lantolf (2006) has described how peers can learn within the Zone of Proximal Development (ZPD). His research, based on the work of Vygotsky (1986), shows that the scaffolding that results from negotiating meaning can contribute to the language development of more proficient language learners. At the same time, less proficient learners can benefit by learning from their more able peers. As such, a well-balanced group consists of learners of

varying language proficiencies. Jacobs (2006) recommends deliberately constructing groups to achieve a balance of proficiency within each group, but this may prove difficult if the level of students is unknown.

One only has to glance at any of the major textbooks used within English language teaching to see that group work is used extensively in the classroom, and indeed is central to most pedagogies in second language acquisition (SLA), including communicative language teaching (CLT) and task-based learning (TBL). With groups playing such a central role, the method of group construction can be considered a central issue in the language classroom. In order for teachers to construct groups for specific purposes they need to know the relative proficiency of students in each class; however, standardized test scores such as the Test of English for International Communication (TOEIC) may be unavailable due to privacy concerns or budgetary limitations. Although there are many standardized tests such as the International English Language Testing System (IELTS), Test of English as a Foreign Language (TOEFL), and the Versant test, they all cost money. Even free online tests require a large number of computers and reliable internet access. With this in mind, how then can teachers quickly and effectively determine the relative proficiency levels of their students? It is not the authors' purpose here to compare the range of standardized tests. Rather, the authors will propose that dictation is a free alternative to standardized testing that is easy to administer and score, and assesses multiple dimensions of language proficiency. This paper discusses how to use dictation as a measure of English language proficiency and dictation's practicality in determining students' relative English levels as an aid in group construction. The aim of this paper is to provide a guide to the construction, implementation, and marking of a dictation test. The authors will also provide data supporting the fact that dictation is a test of overall English proficiency.

Why Dictation?

Upon first consideration, dictation might not appear to be a reliable, accurate, or sufficiently comprehensive method to measure a learner's language proficiency. The test itself is very simple. The words and the word order are given, so it would appear that very little is tested. Indeed, it would seem that dictation is limited to testing spelling and, perhaps, working memory. It seems reasonable to assume that listening comprehension and concentration are being assessed as well. However, Oller (1971) noted that the words and word order are only known to the person giving the dictation. This means that dictation actually requires the listener to produce the words and complete the sentences with the correct syntax and structure. Without understanding the context of the language, grammatical errors are inevitable. Oller (1971) was claiming that dictation includes knowledge of English grammar and syntax, and that by comparing the student-generated sequence with the original dictation, the differences observed can allow researchers to estimate the student's general language proficiency.

In his 1971 study, Oller conducted an evaluation of the English as a Second Language Placement Examination (ESLPE) for the University of California at Los Angeles. Oller (1971) found that dictation correlated most strongly with each part of the test, suggesting that dictation alone, more than any other part of the ESLPE, was the single strongest predictor of English-language proficiency (significant at the .001 level of confidence).

In a follow-up study by Oller and Streiff (1975) which addressed criticisms of the 1971 paper, they showed that dictation seemed to be "tapping an underlying competence" (p. 33) due to the consistently high correlations between dictation and the ESLPE. In light of these findings, they reasonably concluded that dictation is "a useful instrument for testing short-term instructional goals as well as integrated language-achievement over the long-term [sic]" (p. 34).

Tests such as TOEIC are designed to measure a wide range of English proficiency and are lengthy and comprehensive. Dictation is comparatively short, and it is difficult to construct a dictation test that will measure students from beginner to advanced level. However, dictation does appear to tap into what Oller (1979, p. 24) calls “expectancy grammar,” which is described as “the combined knowledge of the phonological, syntactic, semantic, discourse, sociolinguistic, and pragmatic rules of a language” (Fouly & Cziko, 1985, p. 556). Furthermore, dictation as a measure of language proficiency is “grounded in the similar types of knowledge and psychological processes believed to underlie both language use and dictation test performance” (Fouly & Cziko, 1985, p. 556).

As such, despite its obvious shortcomings, a well-designed dictation can yield useful measures of second language proficiency and aid in constructing well-balanced groups in the language classroom. The remainder of this paper will describe how the authors constructed a dictation test to differentiate between the relative proficiencies of students in their classes, and finally, present data supporting the fact that their dictation was an effective measure of English proficiency.

The Study

The authors were interested in the use of dictation and aimed to determine its feasibility and applicability in a Japanese context. In the following section, they briefly describe the context where the test was administered before going on to discuss the points to consider when making a dictation test, how to administer it, and the scoring method used. Following on from this, results of a correlation analysis are presented.

Context

The test was administered to 138 students majoring in science at a private university in western Japan. The students were in six intact classes that were streamed according to major within the department and not English proficiency. English classes are compulsory, with three 90-minute classes each week focusing on speaking, reading, and writing. All of the students had passed the English section of the entrance exam to the university, so English proficiency ranges could be hypothesized to be small. However within a given class there was a wider range of English proficiency, with the most extreme case being in a class where one student’s TOEIC score was 895, while in the same class the lowest score was 220.

The dictation tests were administered in the speaking classes which were all taught by the same teacher. The data were collected over a 2-year period, and in the first year, three classes took the TOEIC test (see <https://www.ets.org/toEIC> for details) and the data were made available to the teacher. Due to privacy issues, in the second year of the study, the university decided that the teachers were no longer permitted to have access to the TOEIC scores for students, and therefore the SLEP test was administered by the teacher to the three classes in that year. The SLEP test is no longer in use but was used in American middle and high schools for students whose first language was not English, and has since been replaced by the TOEFL Junior test.

Procedure

Making the Dictation

In order to avoid floor and ceiling effects (a dictation that is too easy or too difficult), the difficulty of the dictation should encompass the range of student abilities. Ideally the dictation should be generally understood by all the students, but there will be sections which may be considered slightly beyond most students’ current level of proficiency (equivalent to Krashen’s *i+1* [1981]). Most reasonably experienced teachers can typically gauge the approximate level

of a class after 1 or 2 weeks of instruction, and although specific information may not be available for a class, teachers are often aware of the general level of students within the institution where they teach. Once the approximate level of the students in the class has been determined, the teacher should compose a dictation passage approximately 100 words in length (see Appendix A for the dictation passage used in this study). It is advisable for teachers of limited experience to have co-workers check the dictation to make sure that the vocabulary and general level of English are appropriate for the students. For the students in the current study (average TOEIC = 390), a simple narrative using mainly the simple past tense was considered appropriate. The teacher who created the dictation had been teaching at the university for approximately 18 months when it was made and had over 5 years' experience teaching in Japanese schools.

After writing the dictation passage, it is then necessary to record it. Although the dictation can be read aloud, it is easier and safer for the teacher to make a prerecording of the dictation so that during administration, the teacher is free to monitor students and ensure they are all attempting the test. Recording also ensures uniformity between classes should the classes be mixed subsequently. One recording (A) should be at normal speaking speed with no pauses, and the second (B) should be at normal speed with pauses at phrasal boundaries (e.g., I had a really nice time this summer [pause] relaxing with my family) to allow time for students to write down what they hear. When making the second recording, it is advisable for the speaker to actually write down what was just said to ensure that there is sufficient time for students to complete the task; this was done for the dictation test used in this study.

Administering the Dictation

When administering the dictation, the students should be informed that they will listen to the dictation three times and that they must write down exactly what they hear. With lower-level students, it is worth explaining this in their first language, and for this study, the directions were written in Japanese to ensure that students understood the task. The following procedure was on the board for students in English or the first language, depending on the level of the students:

1. Play Recording A (natural speed). Students listen (can make notes).
2. Play Recording B (with natural pauses at phrasal boundaries). Students write down what they hear.
3. Play Recording A (natural speed). Students check their work.

It is important to emphasize that students should not attempt to copy down what is said when first listening, and that they will be given a chance to listen again when pauses are added. Writing directions on the board is advisable to ensure that all students understand.

Scoring

After successfully administering the test, teachers must then mark answer papers. In this study, two raters checked all the answer papers. Depending on how many students take the dictation, scoring can be done by one or several raters. Similarly, depending on the stakes of the test, an individual can mark the papers, although to ensure consistency, it is preferable to have at least two people mark the tests. If there is a relatively small number of students (up to 50), then two raters can mark all the papers. If there are a large number of students, then it may be necessary to have more raters and have some overlap between individual raters. However, involving several raters would necessitate the establishment of some basic guidelines for marking. Factors to consider include spelling, tenses, missing words, extra words, correctly spelled but wrong words, misspelled but phonetically accurate words, points for partial knowledge, and the

number of points that can be awarded for each word. For example, in this study, the decision was made to accept both / and *r* for spelling, as these sounds are hard to differentiate for Japanese learners of English. This meant that *relax* and *rerax* were considered acceptable.

The researchers used a dichotomous scoring scale with no partial points awarded (words can be given a score of 1 or 0). Words that were misspelled but still phonetically accurate and recognizable were given 1 point. In addition, words that were in the wrong order but written correctly (orthographically or phonetically) were also given one point. No points were awarded or deducted for extraneous words. Copies of the students' answer papers were made for both raters to mark independently.

To ensure that each rater gives the same points for the same answers (inter-rater reliability), it is important to have a norming session. For this study, the first few students' answer papers were marked, followed by discussions regarding any differences in grading. Once the norming session was complete, the raters individually graded each answer paper and input the scores on a shared scoring sheet made using Microsoft Excel (see Appendix B for a section of the scoring sheet used). Using a macro (a collection of commands to automate a procedure) for Excel, any word which was not awarded the same score by both raters was highlighted. After all of the papers were marked, the raters then met to discuss any discrepancies in their marking and decided upon a final score for each student. While marking, the raters also kept in regular contact to ensure that any issues that arose could be dealt with. It is also important to regularly compare papers marked by the same rater as a single person's grading consistency may change over time (intra-rater reliability). Overall, the raters' inter-rater reliability was 0.98, which is considered to be very high agreement, and the marking was time-consuming but not difficult. If the purpose of administering a dictation is relatively low stakes such as for the formation of groups, a single rater can mark all the answer sheets for a particular class. This would remove the need for marking guidelines or a norming session and would reduce the overall marking burden.

Correlation Analysis

To convert the scores from the dictation to a true scale as needed in correlation analysis, the researchers conducted a Rasch analysis of the dictation test using Winsteps (Linacre & Wright, 2007), and used the subsequent logit scores for the correlation (a description of Rasch analysis is beyond the scope of this paper but see Bond and Fox [2007] for a comprehensive introduction to the Rasch model). The Rasch model also allows researchers to see how well the test used is able to differentiate between the relative proficiency of the students. The results for the dictation showed a person separation of 2.91, well above the benchmark value of 2 (Linacre, 2007), indicating that the dictation is effective in separating students by proficiency level.

In order to ensure that the dictation was a test of English, the researchers conducted a correlation analysis with the student scores for the TOEIC and SLEP tests. The results are shown below in Tables 1 and 2.

Table 1***Correlations for the Dictation and TOEIC Test***

	1. DICT	2. T-LIST	3. T-READ	4. T-TOTAL
1. DICT	—			
2. T-LIST	.78	—		
3. T-READ	.72	.69	—	
4. T-TOTAL	.81	.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).

Table 2***Correlation for the Dictation and SLEP Test***

	1. DICT	2. SLEP
1. DICT	—	
2. SLEP	.60	—

Note. DICT = Dictation test; SLEP = SLEP Test. All correlations significant at $p < .05$ (2-tailed).

As can be seen from Table 1, the dictation test correlates strongly with the listening and reading parts of the TOEIC test, and the strongest correlation is with the overall score for TOEIC. This suggests that the dictation is measuring general English proficiency including grammar and vocabulary, and not just listening skills. From the strong and highly significant correlations, the researchers can conclude that the test is a good measure of English proficiency. Table 2 shows that the correlation with the SLEP test is moderate, although still significant. The SLEP test used was a short form of only 30 minutes and failed to achieve a strong distribution of scores, which may account for the weaker correlations here.

Conclusion

In conclusion, the findings of this study are in line with others that have shown a high correlation with other standardized tests (Oller, 1971), and it can be reasonably concluded that dictation is an accurate measure of English language proficiency. In this paper the researchers have provided detailed instructions on how to create, administer, and mark a dictation test. The test was easy to create and administer, and although marking was somewhat time consuming, it was a simple process. From this simple test constructed in-house after little more than a year teaching at the university, the researchers were able to gain proficiency scores for the students in their class. These scores can aid in constructing balanced groups in the language classroom, which is central to most SLA pedagogies. The results of the Rasch analysis showed that the dictation was effective in separating the students into different levels based on proficiency. Dictation itself is cheap and easy to conduct and is a viable alternative when standardized proficiency scores for students are unavailable. It should be noted that dictation does have its own limitations and may be limited in measuring all aspects of an individual's language proficiency. However, for the purposes of group construction, dictation is an appropriate alternative.

Author Note

Aeric Wong, Hirao School of Management, Konan University, Japan; Paul Leeming, Department of Economics, Kinki University, Japan.

Correspondence concerning this article should be mailed to Aeric Wong at Konan University, 8-33 Takamatsu-cho, Nishinomiya, Japan, 663-8204. E-mail: awong@center.konan-u.ac.jp

References

- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Routledge.
- Fouly, K. A., & Cziko, G. A. (1985). Determining the reliability, validity, and scalability of the graduated dictation test. *Language Learning*, 35(4), 555-566. <http://dx.doi.org/10.1111/j.1467-1770.1985.tb00361.x>
- Jacobs, G. M. (2006). Issues in implementing cooperative learning. In S. G. McCafferty, G. M. Jacobs, & A. C. DaSilva Iddings (Eds.), *Cooperative learning and second language teaching* (pp. 30-46). New York, NY: Cambridge University Press.
- Krashen, S. D. (1981). *Second language acquisition and second language learning*. New York, NY: Pergamon Press.
- Lantolf, J. P. (2006). Sociocultural theory and L2: State of the art. *Studies in Second Language Acquisition*, 28(1), 67-109. <http://dx.doi.org/10.1017/S0272263106060037>
- Leeming, P. (2011). Japanese high school students' use of L1 during pairwork. *International Journal of Applied Linguistics*, 21(3), 360-382. <http://dx.doi.org/10.1111/j.1473-4192.2011.00284.x>
- 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). Group work, interlanguage talk, and second language acquisition. *TESOL Quarterly*, 19(2), 207-228. <http://dx.doi.org/10.2307/3586827>
- Oller, J. W., Jr. (1971). Dictation as a device for testing foreign-language proficiency. *English Language Teaching*, 25(3), 254-259. <http://dx.doi.org/10.1093/elt/XXV.3.254>
- Oller, J. W., Jr. (1979). *Language tests at school: A pragmatic approach*. London, UK: Longman.
- Oller, J. W., Jr., & Streiff, V. (1975). Dictation : A test of grammar-based expectancies. *English Language Teaching*, 30(1), 25-36. <http://dx.doi.org/10.1093/elt/XXX.1.25>
- Vygotsky, L. S. (1986). *Thought and language*. Cambridge, MA: MIT Press.

Appendix A

Dictation Passage for Students with an Approximate TOEIC Score of 400

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.

Total word count: 93 words

Appendix B
Example Scoring Sheet

Student ID	I	had	a	really	nice	time	this	summer	relaxing	with	my	family
1024	1	1	0	1	1	1	1	1	1	1	1	1
1025	1	0	0	0	0	0	1	1	0	1	1	1
1026	1	1	0	1	1	1	0	1	1	0	1	1
1027	1	1	1	1	1	0	0	1	0	0	1	1
1028	1	0	0	0	0	0	0	1	0	1	1	1
1029	1	1	0	1	1	1	1	1	1	1	1	1
1030	1	0	1	1	1	1	1	1	1	1	1	1
1031	1	1	0	1	1	1	1	1	0	1	1	1
1032	1	1	0	1	1	1	1	1	1	1	1	1
1033	1	1	0	0	0	1	1	1	1	1	1	1
1034	1	0	0	1	1	0	0	1	0	1	1	1
1035	1	1	0	1	1	0	1	1	1	0	1	1
1036	1	1	1	1	1	1	1	1	1	1	1	1
1037	1	1	0	1	1	1	1	1	0	1	1	1
1038	1	1	1	1	1	1	1	1	1	1	1	1
1039	1	1	0	0	0	0	1	1	0	1	1	1
1040	1	0	0	1	1	1	1	1	1	0	1	1
1041	1	1	1	1	1	1	1	1	1	1	1	1
1042	1	1	1	1	1	0	1	1	0	1	1	1
1043	1	1	0	1	1	1	0	1	1	1	1	1
1044	1	1	0	0	0	0	0	1	0	1	0	1
1045	1	0	1	1	0	0	0	1	0	1	1	1
1046	1	1	1	1	1	1	0	1	1	1	1	1