＜論文＞

Developing a Dictation Test to Stream Learners: Quick and Dirty or Smart and Efficient?

Harumi KIMURA

Introduction

In language teaching/learning communities, dictation has been used both as an exercise for learning and a tool for assessment. In dictation, typically learners receive aural input, keep the information in the mind for a little while, and reproduce it in writing (Nation, 1991). Traditionally, dictation is of three different types: full dictation, partial dictation, and dicto-comp/dictogloss. In full dictation, learners produce in writing what they hear, sentence or passage, word-by-word. In partial dictation, learners fill in the blanks in a written passage as they listen. In both full and partial dictation, learners are supposed to replicate the original language. In dicto-comp/dictogloss, however, learners do not reproduce the exact text they hear but reconstruct the content in grammatically accurate forms with correct spelling, paying attention to text cohesion. If students do this activity individually, it is called dicto-comp. If learners work in small groups to reconstruct the meaning of text among members or compare different versions of text reconstruction among different groups in discussion and then refine their own individual versions accordingly (Wajnryb, 1990), the activity is called dictogloss. Whereas all these types of dictation have been used as classroom learning activities, partial dictation has been most widely used for assessment purposes. In this paper, I focus on partial dictation as an assessment technique primarily for streaming students. First, I quickly review the controversy over the use of dictation for assessment. Second, I explain how I have developed a dictation test battery for this study. Third, I analyze the test and the test items statistically and argue that partial dictation is a smart and efficient way to assess students' basic listening proficiency, at least for classroom use. I also share some ideas for using the dictation test results for diagnostic assessment purposes.

Controversy over Dictation Use

Teachers and researchers have expressed concern over what dictation measures (Stansfield, 1985). Teachers who use dictation as part of a test battery usually take a bottom-up view of listening skills, which states that listening comprehension starts with speech perception and word recognition and continues on to higher levels of processing—syntax, discourse, and schematic levels. A dictation test is designed to measure how well each participant can identify and isolate separate linguistic elements in a stream of speech. At first glance, dictation seems like a discrete-point test (Buck, 2001) in that it measures the ability to recognize elements of the language without much contextual information. Teachers who oppose the use of dictation for language-testing purposes argue that it falls short of measuring functional language abilities, namely, how well listeners can use pragmatic, discourse, and world knowledge to comprehend the expressed meaning, and making inferences on a speaker's intention in context—i.e., top-down, or higher-order processing skills. They believe that listeners construct meaning from their personal experiences, with references to schematic connections that accompany this understanding, and that dictation cannot tap into this process.

However, these two processes are not independent or separate but intertwined and coordinated. They occur simultaneously and influence each other (Rost, 2005). Listeners structurally process the relationships among the linguistic elements in hearing sound sequences, and in this process listeners' structural knowledge is at work (Irvine, Atai, & Oller, 1974; Oller, 1971, 1979). It is arguable that dictation helps assess this particular ability and that it is neither a simple word recognition test or a spelling test in context.

Oller and his colleagues are eminent proponents of dictation. Dictation is considered to elicit data on appropriate sound perception and the correct spelling of words (bottom-up, lower-order processing). In addition, Oller (1971) and Oller and Streiff (1975) have contended that dictation makes test takers to segment and analyze words and word sequences and synthesize the information to make sense of the sound sequence (top-down, higher-order processing). Listeners formulate hypotheses about meaning based on recognized language units and structural patterns while analyzing and synthesizing information sequentially. Oller and Streiff (1975) named this process as "grammar-based expectancy" (p. 77), or expectancy grammar, and argued that dictation activates this internalized grammatical knowledge and, thus, dictation is a kind of integrative test (Buck, 2001). In fact, Irvine, Atai, and Oller (1974) demonstrated that a dictation test correlated with both listening comprehension sub-section TOEFL score (r = .69) as well as the total TOEFL score (r = .69) and concluded that dictation taps comprehensive language knowledge.

In a more recent study, Wong and Leeming (2014) demonstrated their findings that full dictation test scores correlated highly with TOEIC test scores. A dictation test was correlated with the TOEIC listening comprehension sub-section score (r = .78) and with the total TOEIC score (r = .81) and the researchers concluded that dictation could be an appropriate alternative to standardized tests. They also found that classroom teachers could use the scores to create well-balanced, heterogeneous groups according to listening proficiency.

In another recent study, Cai (2012) demonstrated that partial dictation and gap filling on a summary tapped into the same underlying construct—general linguistic ability. Cai (2012) gave both dictation and gap filling on summary to his participants. In dictation, participants were asked to fill in the blanks with the exact words they listened to. In gap filling on a summary, on the other hand, participants were instructed to fill in the blanks on summary of a passage they had just heard. This forced the participants to process meaning. A confirmatory factor analysis indicated that the general factor, in this case general linguistic ability, explained the majority of the intercorrelations among the test items. The two different method factors—in this case, dictation and gap filling on summary—that were irrelevant to the general ability accounted for part of the residual correlations. In general terms, both tests measured the same ability. Furthermore, all of the missing words both in the dictation and gap filling on summary were content words. Thus, it should be safe to say, Cai argues, that these results indicate that participants used both bottom-up and top-down skills to process the aural input for meaning.

Cai's (2012) study provided empirical evidence that dictation can assess more than phonemic discrimination, word recognition, and spelling in context, and the expectancy grammar provides a theoretical underpinning for the use of partial dictation for assessment purposes. Other researchers such as Weir (2005) and Cohen (1994) have also expressed support for the use of dictation by reporting that it provides a good supplement to other listening test forms. A dictation test is easy to make and partial dictation is not as time-consuming to score as full dictation (Cai, 2012).

Furthermore, partial dictation would be useful in assessing listeners' ability to identify not just content words but also function words in connected speech. In natural speech, a speaker naturally connects words into a smooth flow of speech production and includes features like word linking, and vowel and consonant reduction (Brown & Kondo-Brown, 2006). These features make aural word segmentation difficult, and identifying such features constitutes an important element of L2 listening proficiency.

Moreover, the abovementioned grammar involves effective processing of function words because such grammar knowledge is syntax-based (Oller, 1979) and function words are processed differently from content words (Field, 2008). As a result, it should be more appropriate to measure both content and function words in dictation. Field (2008) theorized that two distinct routes of processing aural language exist. Content words are processed for meaning while function words are processed for pattern matching. Although it is beyond the scope of this paper to discuss the theoretical basis and empirical evidence of expectancy grammar and the two different processes separately posited for content words and function words, I would argue that it makes more sense to test both groups of words in dictation to measure listening proficiency. In the next section, I describe in detail how I designed and developed a partial dictation test so that classroom teachers can follow the steps, or only the key steps, to create their own.

Creating a Dictation Test

In this section, I explain how I created a dictation test for the purpose of measuring English learners' listening proficiency in a short period of time.

Test Takers

The test takers, or participants, in this study were 1,177 predominantly Japanese university students of English who are currently taking required English classes (557 male students, 603 female students, 17 students unknown, M_{age} = 19.4 years, age range: 18–28). The students, from 15 different Japanese schools, had completed at least six years of formal English education prior to entering the university. Their English proficiency, English learning history, and majors varied widely.

Dictation Test

The test consists of 20 sentences and each sentence has three successive blanks to fill in for a total of 60 blanks (Appendix). Each blank is counted as one item. The three linguistic forms do not necessarily constitute a linguistically well-formed unit but include both content and function words; thus, the dictation test is an assessment of the participants' phonemic discrimination and word recognition abilities as well as structural knowledge.

Using readability indices, sentence length, number of syllables per word, lexical frequency analysis, and sentence complexity estimates, the sentences were grouped into three levels of difficulty in order to measure all of the participants' basic listening proficiency reasonably precisely (See Table 1). Three sentences, one from each level, as well as a summary of the sentence analyses are displayed below.

Easy sentence 1: Come (in) (and) (sit) down.
Intermediate sentence 1: A lot of (people) (around) (the) world speak English fluently.
Difficult sentence 1: Freedom of speech is the most (important) (thing) (in) a democracy.

Table 1 *Descriptive Statistics for the Dictation Test Sentence Groups*

	Easy Group	Intermediate Group	Difficult Group
Average number of words	6.4	10.0	12.3
Flesch reading ease score	102.3	88.5	75.1
Flesch-Kincaid grade level	0.5	3.4	5.8
Average syllables per sentence	6.9	12.7	16.2
Average number of syllables per word	1.2	1.3	1.4
1st 1,000 words	93%	90%	86%
2nd 1,000 words	2%	9%	3%
Other words	5%	1%	11%
Reading speed	159.8 wpm	167.6 wpm	141.8 wpm

Moreover, because the sentence structure changes from simple to more complex, the blanks become increasingly more challenging for the test takers to fill in correctly. Deleted items include those with inflectional and derivational morphemes. For example, Sentence 3 in the intermediate group has a plural noun, *dangers*, that sounds similar to one of the words in the same word family, *dangerous*. The same sentence also has an adjective phrase and the test takers must fill in the adjectival marker. (Intermediate sentence 3: Do you know about the (dangers) (associated) (with) smoking?) Sentence 4 in the difficult group has an embedded sentence with a relative pronoun, and an unstressed relative pronoun is one of the missing words. It is also followed by a verb with a third-person ending. (Difficult sentence 4: The (addiction) (that) (affects) most people is said to be chocolate.) In sum, the test assesses participants' syntactic knowledge and parsing skills as well as word recognition skills.

One may want to argue that it is better to take the three-word sequence as one test item, but I do not take that approach. Each of the 20 three-word sequences is a combination of content word(s) and function word(s), and as content word processing and function word processing are considered to be distinct as discussed in the previous section, each word is considered to consist of one test item in the sentence-level context. Thus, for example, D28, (addiction) in difficult Sentence 4 in the above example, should be taken as the word (addiction) embedded in the following sentence: *The addiction that affects most people is said to be chocolate*.

To address the issue of incongruence between word segmentation skills and spelling knowledge, each item is awarded a score of zero for no response or a wrong answer, one point for a partially correct answer, or two points for a fully correct answer. Mistakes on verbal inflections, plural markers, and spelling are counted as partially correct and scored as one.

The dictation test takes six-and-a-half minutes to administer. First, the test starts with instructions in Japanese, the participants' native language, followed by an example. Each sentence is repeated twice with a pause in between: 3 seconds for the easy sentences, 4 seconds for the intermediate sentences, and 5 seconds for the difficult sentences. After the second reading, a longer pause is inserted before the next sentence is read: 5 seconds for the easy sentences, 6 seconds for the intermediate sentences, and 7 seconds for the difficult sentences. The pauses are meant to provide time for writing the missing items stored in the test-takers' short-term memory in the process of understanding the sentence, but not for conjuring up ideas about the meaning of the sentence or retrieving information from long-term memory.

The reading speed of the easy sentence group was 159.8 wpm, that of the intermediate sentence group was 167.6 wpm, and that of the difficult group was 141.8 wpm. The difficult sentences were slowest because their sentential structures were more complicated and therefore contained longer pauses. For example, there was a long break between the subject, *Freedom of speech*, and the predicate, *is the most important thing in democracy*, in Sentence 15, and there was also a long break between the main clause, *The company employees have recently held a strike*, and the following adverbial clause, *because they didn't get a pay raise*, in Sentence 19. These internal breaks made the difficult sentences slower. Audiobooks are read at 150–160 words per minute, which is the range that people comfortably hear and vocalize words (Williams, 1998). The listening test recoding was considered to be similar to such speeds.

In the next section, I describe how I analyzed the test quantitatively and qualitatively. The statistical procedures are rather highly technical, but they are presented (a) to demonstrate to readers with statistical knowledge that a short dictation test can be of good quality, and (b) to invite general readers to consider the aspects of knowledge each item is tapping into. The second purpose leads to the claim that dictation can be a diagnostic tool for classroom teachers.

Analysis

In this section, I examine the dictation test using the partial credit Rasch model. The Rasch model is based on a probabilistic procedure in which people's ability and an items' difficulty are estimated against each other—in other words, a trade-off. A test-taker with a greater ability (in this case, listening proficiency) will have a higher probability of answering an item correctly than a less able one whereas it is less probable that a difficult test item will be answered correctly by the same test-taker than an easier item (Bond & Fox, 2007). The partial credit model is chosen to award a score of zero for no response or a wrong answer, one point for a partially correct answer, or two points for a fully correct answer.

Dimensionality

The dimensionality of the 60-item dictation test was inspected using the Rasch PCA of item residuals. This is an important step in examining whether a test measures a single psychological construct, in this case, listening proficiency. The results indicated that the Rasch model explained 56.2% of item variance (eigenvalue = 76.8) and the first residual contrast accounted for 1.8% of the variance (eigenvalue = 2.5). The eigenvalue of the first contrast is well below the 3.0 criterion proposed by Linacre (2009) and thus suggests that the dictation test measures a single dominant construct.

Gap-filling dictation tests, most importantly, tap into listeners' word recognition skills, which are "the basis of spoken-language comprehension" (Rost, 2002, p. 20). In fact, one of the essential subskills of listening is segmenting speech while recognizing words online without access to blank spaces between them as in written language (Cutler, 1998). Computational models of language perception and language learning are based, first and foremost, on word recognition (Brent, 1999). As such, word recognition is the most prominent area of difficulty for L2 listeners, particularly those whose L1 phonology differs significantly from the L2 phonological system. Rost and Ross (1991) reported that beginning and intermediate L2 listeners stated that word recognition is often the most problematic process in listening. A participant in the present study made the following mistake for Sentence 10.

Correct: Do you know about the (danger) (associated) (with) smoking?
Incorrect: Do you know about the (*dangerous*) (*your*) (*health*) smoking?

The listener was unable to identify the word boundary between *danger* and *associated* and possibly resorted to a compensatory strategy such as activating familiar schemas in search of target words. As a result, erroneous word recognition occurred. However, dictation requires listeners not only to analyze speech into chunks but also to synthesize the information from the recognized chunks to arrive at the meaning of the speech (Neisser, 1967). Processing the language successfully requires syntactic processing, as the incoming speech must be mapped onto appropriate grammatical structures. Listeners draw upon a set of grammatical and semantic cues when engaged in form-function mapping (Rost, 2002). Two examples of trial-and-error syntactic processing, which Oller and Streiff (1975) called a creative error, can be found in the dictation test data. The first example concerns Sentence 9.

Correct: You should go and see a doctor (when) (you) (feel) sick.
Incorrect: You should go and see a doctor (*while*) (you) (*are*) sick.

In the example above, the participant failed to correctly recognize the words but succeeded in syntactic processing on the first and third target words when she guessed that an adverbial clause should come after the sequence, "You should go and see a doctor," and that the adverbial clause would denote time. She also correctly guessed that there must be a connecting verb between you and *sick*.

Another example concerns Sentence 17.

Correct: They must be hungry (and) (exhausted) (after) a long day of work.
Incorrect: They must be hungry (and) (*thirsty*) (after) a long day of work.

In this example, the participant used both syntactic and semantic cues. She failed to identify the word, *exhausted*, but knew that some adjectival phrase should occur in the gap. She also inferred an appropriate gap in meaning from the context; that is, people can get both hungry and thirsty when they have worked a long time.

In comprehending language, people link utterances to structurally appropriate patterns through inferencing (Oller, 1971). Skillful language users are able to infer the meaning of a message because they have developed a grammar of expectancy. The data indicate that the gap-filling dictation actively involved test takers in using both semantic and syntactic cues in a search for meaning and that teachers can gain information about the cues the test taker has missed at least in some cases.

In the next subsection, all of the items are checked for item fit statistics. This step is needed to investigate whether the items are functioning well to assess test-takers' ability. When an item is judged otherwise, it will be removed because it does not fit the good measurement model. The "new" test with the rest of the items will be inspected for dimensionality and for fit statistics of each item until the data fit the model. I will demonstrate this rather cumbersome, repetitive procedure to invite readers to speculate on why each item did not contribute to the effective measurement of test-takers' abilities. Readers can skip the technical jargon and numbers if they wish. Furthermore, in my humble opinion, classroom teachers do not have to go through these statistical processes. By hand-scoring tests, experienced teachers can distinguish between good items and bad items based on students' test performance.

Rasch Descriptive Statistics

The fit statistics of the 60-item dictation test were inspected. All items except two, D32 and D51, satisfactorily met the infit MNSQ criterion of .70-1.30 (within the range of two standard deviations, McNamara, 1996). Item D32, *at*, which appeared in the sentence, *I met some friendly students on the* *first (day) (at) (school)*, had an infit MNSQ statistic of 1.51. The word is not stressed and it is almost inaudible. The most common incorrect answer was *of*, which makes the sentence meaningful and is grammatically correct. It was likely that the item had unexpected responses because more able students missed the item while less able students guessed it correctly. Because of its poor fit to the model, the item was deleted.

The second Rasch analysis was conducted with the remaining 59 items. The PCA of item residuals showed that the Rasch model explained 57.1% of item variance (eigenvalue = 78.4) and the first residual contrast accounted for 1.9% of the variance (eigenvalue = 2.5). Item D51, *after*, which was outside the appropriate range in the first analysis as well, had an infit MNSQ statistic of 1.46 and was still identified as misfitting because it did not meet the infit MNSQ criterion of .73–1.29. D51 was tested in the sentence, *They must be hungry (and) (exhausted) (after) a long day of work*. Although the word was pronounced clearly, it is possible that some participants who were able but who had struggled with the previous word, *exhausted*, could not understand or retain the word while some less able students, who failed to process one or both of the first two missing words, wrote the word correctly. As the item displayed excessive randomness, it was therefore deleted.

The third Rasch analysis was conducted with the remaining 58 items. The PCA of item residuals showed that the Rasch model explained 58.0% of item variance (eigenvalue = 80.0) and the first residual contrast accounted for 1.8% of the variance (eigenvalue = 2.6). At this point, another item, D49, *and*, was found to be misfitting with an infit MNSQ statistic of 1.32; it did not meet the infit MNSQ criterion of .75-1.27. The item appeared in the same sentence as item D51, *They must be hungry (and) (exhausted) (after) a long day of work*. The word was unstressed as it occurred between the content words, *hungry* and *exhausted*. The two other items, D02 and D20, also tested the participants' ability to accurately perceive *and*, but they fit the model well with infit MNSQ statistics of .89 and .99, respectively. Although the reason for the randomness was not clear and D49 was deleted, this incongruence unwittingly demonstrated that each tested word was embedded in a context and that the same word can be of different difficulty depending on the specific context.

The fourth Rasch analysis was conducted with the remaining 57 items. The PCA of item residuals showed that the Rasch model explained 58.7% of item variance (eigenvalue = 81.0) and the first residual contrast accounted for 1.9% of the variance (eigenvalue = 2.6). As Table 2 shows, all of the items met the infit MNSQ criterion of 77–1.25. The item difficulties ranged from 37.7 to 63.9 CHIPS, with a mean of 50.0 (*SD* = 6.6). The participants' responses ranged from 36.6 to 78.0 CHIPS, with a mean of 52.6 (SD = 4.8). The point-measure correlation coefficients ranged from .14–.67. The items were of the same polarity and appeared to measure the same latent variable effectively. The results supported the content validity of the test.

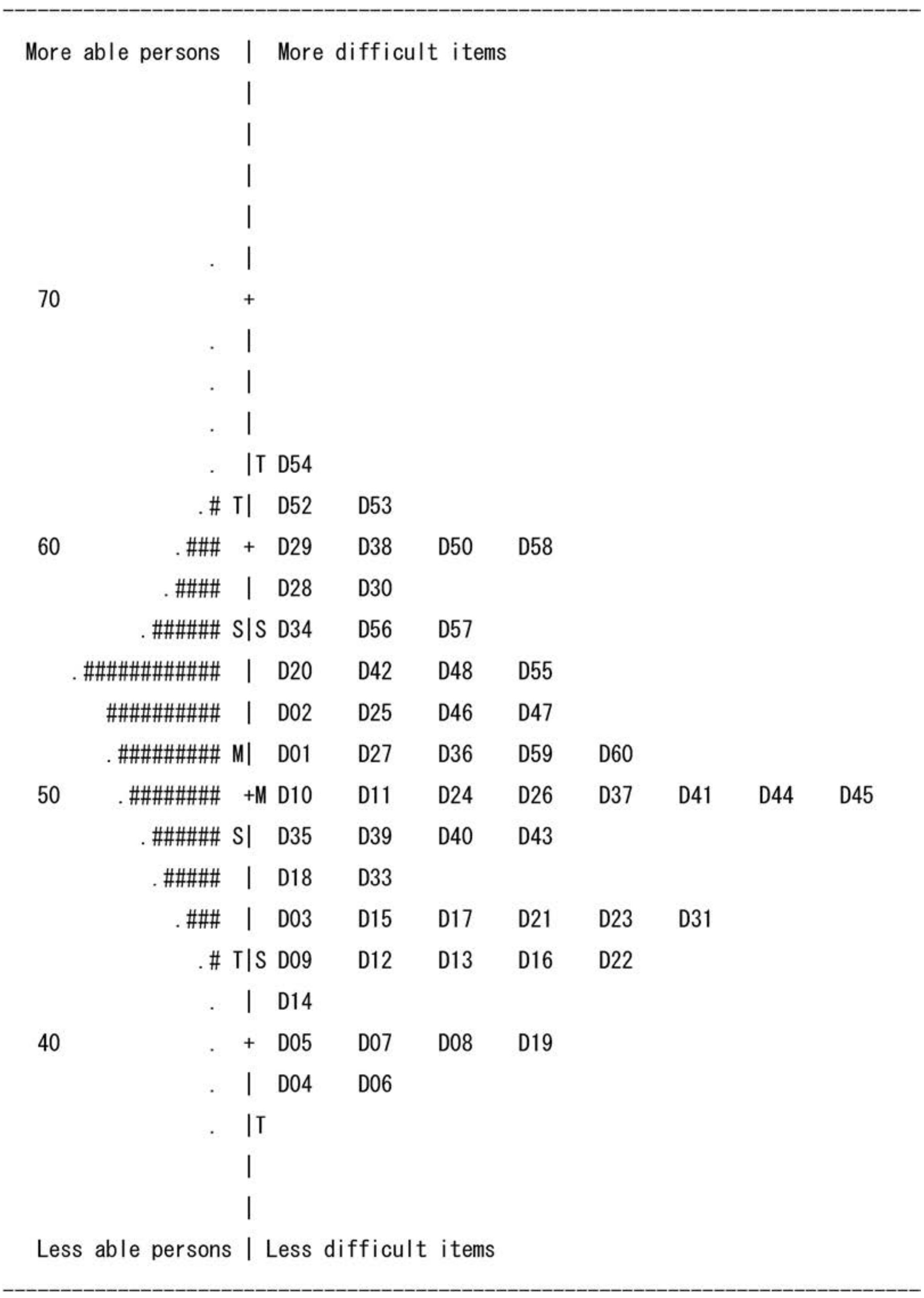
Table 2 *Rasch Descriptive Statistics for the 57-Item Dictation Test*

Item	Difficulty estimate	*SE*	Infit MNSQ	Outfit MNSQ	Infit *t*	Outfit *t*	Pt-measure correlation
D54	63.9	.4	.78	.49	-2.2	-3.2	.48
D53	61.7	.3	.93	2.06	-.7	2.2	.40
D52	61.1	.3	.88	.51	-1.4	-2.3	.47
D38	60.3	.3	1.23	1.79	2.8	1.9	.35
D29	60.3	.3	.85	.73	-2.9	-3.5	.56
D50	59.8	.3	.95	.81	-1.0	-2.4	.52
D58	59.3	.2	.96	.64	-.6	-1.7	.49
D28	59.1	.3	.85	.82	-3.9	-3.8	.57
D30	57.5	.2	1.02	1.05	.4	.3	.51
D57	57.1	.2	.96	.78	-.8	-1.8	.55
D56	55.9	.2	1.17	1.50	4.0	3.9	.47
D34	55.9	.2	1.01	.97	.2	-.2	.55
D20	55.8	.2	1.01	.90	.1	-.5	.56
D55	55.2	.2	1.10	1.10	2.6	2.3	.47
D48	54.8	.2	.92	.88	-2.0	-.9	.60
D42	54.2	.2	1.00	1.00	.0	.1	.59
D25	53.4	.2	1.07	1.12	1.7	1.0	.56
D02	53.2	.2	.91	.84	-2.4	-1.6	.63
D46	53.0	.2	.89	.84	-3.0	-2.7	.63
D47	52.8	.2	1.07	1.08	1.8	1.0	.56
D36	52.2	.2	1.12	1.28	3.1	2.6	.55
D59	52.0	.2	.88	.84	-3.5	-2.7	.63
D60	51.9	.2	.99	.94	-.4	-.7	.60
D01	51.0	.2	1.22	1.80	5.0	6.3	.51
D27	51.0	.2	1.10	1.07	2.3	.7	.56
D10	50.6	.2	.77	.74	-6.3	-3.0	.67
D41	50.1	.2	1.19	1.38	4.6	5.9	.46
D26	50.1	.2	.93	.79	-1.6	-2.5	.61
D45	50.0	.2	.89	.78	-2.5	-1.8	.62
D11	49.9	.2	.89	.75	-2.6	-2.2	.63
D44	49.5	.2	.79	.81	-5.3	-2.0	.65
D24	49.5	.2	1.08	1.14	1.7	1.0	.55
D37	49.4	.2	1.03	1.06	.9	1.4	.49
D35	49.1	.2	.97	1.02	-.6	.4	.55
D43	49.0	.2	.79	.58	-4.8	-3.7	.64

D40	48.4	.2	1.17	1.34	3.5	3.5	.47
D39	47.8	.2	1.16	1.71	3.3	6.6	.44
D33	46.5	.2	1.13	1.35	2.3	2.3	.45
D18	46.0	.2	.92	1.82	-1.3	2.7	.51
D23	45.3	.3	.85	.67	-2.7	-3.5	.55
D17	45.3	.3	.88	.58	-1.8	-2.2	.51
D21	45.2	.3	1.06	1.01	1.0	.1	.44
D03	44.7	.3	1.13	1.16	1.8	.8	.39
D15	44.4	.3	.96	.83	-.5	-.4	.43
D31	44.2	.3	1.26	1.83	3.5	4.4	.30
D16	44.1	.3	.96	.64	-.4	-1.1	.43
D09	42.9	.3	1.07	1.57	.8	2.0	.34
D12	42.8	.3	1.14	1.16	1.5	1.3	.31
D22	42.6	.3	1.12	1.18	1.2	.7	.32
D13	42.5	.3	.91	.83	-1.8	-2.2	.49
D14	41.1	.4	1.08	1.77	.8	3.0	.26
D08	40.7	.5	1.03	.92	.2	.0	.26
D05	40.4	.5	1.11	3.08	.7	3.2	.22
D19	39.9	.4	1.16	1.31	1.6	1.8	.21
D07	39.6	.5	1.10	2.43	.6	2.1	.18
D06	38.1	.6	.94	1.33	-.2	1.0	.21
D04	37.7	.7	1.08	3.93	.4	4.1	.14

Note. Statistics are based on Rasch CHIPS.

Figure 1 shows the Wright map with the participants and items placed on the same linear scale. The items are lined, for easy inspection, on the right side of the map from more difficult (top) to easier (bottom) according to the item-difficulty estimates. The sign # represents test takers, and the position demonstrates that the test taker has a 50-50 chance to answer that level of item correctly. The letter, M, on the vertical line shows the mean. Although the mean of the person-ability estimate was slightly higher than the mean of the item-difficulty estimates, there is a good match between the participants' ability estimates and the item-difficulty estimates.



```
------------------------------------------------------------------------------
  More able persons  |  More difficult items
                     |
                     |
                     |
                     |
                  .  |
  70                 +
                  .  |
                  .  |
                  .  |
                  .  |T D54
                 .# T|  D52    D53
  60            .###  +  D29    D38    D50    D58
              .####  |  D28    D30
            .###### S|S D34    D56    D57
      .############  |  D20    D42    D48    D55
         ##########  |  D02    D25    D46    D47
         .######### M|  D01    D27    D36    D59    D60
  50      .########  +M D10    D11    D24    D26    D37    D41    D44    D45
            .###### S|  D35    D39    D40    D43
             .#####  |  D18    D33
               .###  |  D03    D15    D17    D21    D23    D31
                 .# T|S D09    D12    D13    D16    D22
                  .  |  D14
  40              .  +  D05    D07    D08    D19
                  .  |  D04    D06
                  .  |T
                     |
                     |
   Less able persons | Less difficult items
------------------------------------------------------------------------------
```

Figure 1　*Wright map for the Dictation test.*

A close inspection of the item-difficulty estimates and the item ordering also support the claim that the listening skills measured by the dictation test are more than mere word recognition. The most difficult item was *addiction* (D54), which is not among the 2,000 most frequent English words. Inflected words such as *exhausted* (D50), *associated* (D29), and *spent* (D37) were among the most difficult items, as were plural nouns such as *employees* (D55), *dangers* (D28), and *animals* (D47). It is reasonable that less frequent, structurally more complex items posed difficulties for the test takers. Phonologically more salient, high-frequency forms, on the other hand, were relatively easy to answer. The first person pronoun *I* (D05 & D08), the modal verb *may* (D07), and high-frequency words such as *sorry* (D04) and *go* (D19) were among the easiest items. The content words tested in the dictation test seemed, in general, to exhibit clear contrasts of difficulty in terms of frequency, complexity, and saliency.

However, unstressed function words were dispersed in terms of difficulty estimates. The relative pronoun *that* (D53) and the preposition *on* (D38) were among the five most difficult items. Although these words were not stressed or phonologically salient, they showed structural relationships in the sentence. Test takers whose expectancy grammar was still underdeveloped likely encountered difficulty identifying such structural patterns. On the other hand, *to* (D16) was easier to recognize probably because it was part of a common phrase, *be going to*, and was identified as such. Even less proficient listeners were able to comprehend the word presumably because they successfully perceived and parsed the aural input.

Next, the functioning of the partial credit categories was examined. This procedure is needed to examine the partial credit model with one point award for imperfect answer options. Table 3 provides a summary of the category structure. Each category had more than 10 observations, the average measures advanced monotonically with higher categories indicating more of the latent variable, and the Outfit MNSQ statistics were acceptable as they were less than 2.00. The threshold distance was 11.86, which is greater than the required 6.37 CHIPS for a 3-point scale.

Table 3 *Summary of Category Structure of the Dictation Test*

Category Label	Observed Count (%)	Observed Average	Infit MNSQ	Outfit MNSQ	Structure Calibration	Category Measure
0	405 (39)	49.6	1.22	1.58	NONE	(46.38)
1	27 (3)	50.8	.81	.49	11.82	51.00
2	616 (59)	54.5	1.23	2.25	-11.82	(55.62)

Note. Statistics are based on Rasch CHIPS. 0 = incorrect; 1 = partially correct; 2 = correct.

Lastly, the Rasch person reliability estimate was .94 and the person separation statistic was 4.01. The Rasch item reliability estimate was 1.00 and the item separation statistic was 21.85. The person reliability is equivalent to traditional test reliability (online Winsteps manual). These Rasch reliability estimates are usually rather conservative and therefore more generalizable. The dictation test spread out the participants according to their listening proficiency reasonably well.

Listening comprehension entails more than information processing and syntactic parsing. In fact, "listening is a process involving a continuum of active processes" (Rost, 2002, p. 1) on the part of the listener and different types of linguistic and non-linguistic knowledge are involved in listening comprehension (Buck, 2001). A six-minute, one-way, gap-filling partial dictation test cannot adequately tap into, for example, the interactive aspect of listening. However, considering the fact that speech-processing in meaning-making constitutes the fundamental component of listening skills, the present results indicate that the dictation test was a reasonably reliable and valid instrument for measuring participants' basic listening abilities. Experienced teachers may want to provide feedback to their students not just about their proficiency levels, but also in regards to the skill area the particular student needs to develop.

Conclusion and Suggestions

I demonstrated that a short partial dictation test is easy to create and can be an efficient assessment tool for language teachers. Teachers, with some solid knowledge of word frequency and structural difficulty of the target language, can make a quality dictation test for classroom purposes without needing a sophisticated knowledge of statistics or statistical procedures. I showed the statistical procedure to demonstrate that partial dictation can be of reasonably high quality and is useful for streaming students.

Here are some suggestions for readers who are interested in making a dictation test of their own.

1. Create one dictation test that fits a wide range of proficiency levels.

The dictation test (Appendix) can measure a wide population range of Japanese university students. Among the participants in the current study, only two students had all the correct answers. According to the demographic data, these students had spent more than five years in an English-speaking country and the onset age was less than ten years old. No student scored zero. If teachers keep their students' scores and continue using the same test, they can accumulate the appropriate data and learn to make a good guess about students' listening proficiency and even their past English learning experiences.

2. Check the word frequency levels of the test items.

Teachers' intuition about word frequency is not always accurate (McCrostie, 2014). Free vocabulary profilers such as Lextutor (http://www.lextutor.ca) are user-friendly and helpful. Some discrepancy may exist between the words students should have learned in their previous school education and the words chosen from the frequency-based perspective. It is each teacher's decision as to which group of words to prioritize for the test.

3. Include inflected and derived words for test items.

These items will be informative in making a rough estimation of students' structural knowledge. Teachers not only need to know students' levels but also their strengths and weaknesses in different skill domains. For example, some students may have a good knowledge of vocabulary but lack structural knowledge.

4. Use different structural patterns from simple to complex.

The items are embedded in context. The same item is pronounced differently in a different context and can be of different levels according to the specific context. Usually the same item is more difficult in a longer (and structurally more complex) sentence than in a shorter sentence because of the cognitive load.

5. Test both content words and function words.

The two groups of words are processed differently (Field, 2008) and relate to different subskills of listening (Buck, 2001). Cai (2012) used only content words for test items to demonstrate that listeners are processing for meaning, but meaning is constructed with both groups of words, thus, dictation should assess both types of knowledge.

6. Make successive words blank to better tap into connected speech.

Connected speech is the norm of everyday speech and therefore the ability to comprehend it should be assessed in listening tests, especially in dictation in which word perception and recognition is primarily targeted.

7. Make each blank one item.

This suggestion might be controversial. However, there are two reasons I recommend this particular approach. First, a 60-item test is more reliable than a 20-item test. Second, each item should be considered as a word in that particular context. For example, D01 should be interpreted as (in) in the sentence, *Come in and sit down*. Likewise, D02 is (and) in the sentence, *Come in and sit down*.

8. Sequence the sentences from simple to more complex so that the blanks become increasingly more challenging.

Some less skillful students may give up at their threshold level in the middle of the test. If the sentences and the items are mixed in terms of difficulty, these students may miss easy ones that come later in the test. This situation can be avoided if the sentences are ordered according to difficulty levels.

9. The first items should be quite easy.

Even when an example test sentence is provided before the actual test, students need an easy item when they start. Likewise, the last item should ideally not be the hardest.

10. Readability and other features will usually take care of themselves if vocabulary and structures are controlled.

Listening is not the only skill a student may need in class or need to develop in class, but a certain amount of listening skills are often a prerequisite for joining a group of learners. A short and efficient dictation test is useful to quickly assess students' listening proficiency for classroom purposes such as streaming. Although partial dictation is not designed to assess global dimensions of listening comprehension, affective or strategic aspects of listening comprehension, or interpersonal and cultural aspects, I recommend teachers keep one in their toolbox.

Acknowledgements

This work was supported by Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (C) Number 25370695.

References

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

Brent, M. R. (1999). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences, 3*(8), 294–301. doi:10.1016/S1364-6613(99)01350-9

Brown, J. D., & Kondo-Brown, K. (2006). *Perspectives on teaching connected speech to second language speakers*. Honolulu, HI: University of Hawaii, National Foreign Language Resource Center.

Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.

Cai, H. (2012). Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing*, *30*(2), 177–199. doi: 10.1177/0265532212456833

Cohen, A. D. (1994). *Testing language ability in the classroom*. Rowley, MA: Newbury House.

Cutler, A. (1998). Prosodic structure and word recognition. In A. Friederici (Ed.), *Language comprehension: A biological perspective* (pp. 41–70). Heidelberg, DE: Springer.

Field, J. (2008). *Listening in the language classroom*. Cambridge, UK: Cambridge University Press.

Linacre, J. M. (2009). A user's guide to WINSTEPS: Rasch-model computer program (Version 3.69.0). Chicago: MESA Press.

McCrostie, J. (2014). Investigating the accuracy of teachers' word frequency intuitions. *RELC Journal, 38*(1), 53–66. doi: 10.1177/0033688206076158

McNamara, T. F. (1996). *Measuring second language performance*. Harlow, UK: Pearson Education.

Nation, I. S. P. (1991). Dictation, dict-comp, and related techniques. *English Teaching Forum, 29*(4), 12–14. http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/1991-Dictation.pdf

Neisser, U. (1967). *Cognitive Psychology*. New York, NY: Appleton-Century-Crofts.

Irvine, P., Atai, P., & Oller, J. W. J. (1974). Close, dictation, and the test of English as a foreign language. *Language Learning*, 24(2), 245–252. doi: 10.1111/j.1467-1770.1974.tb00506.x

Oller, J. W., & Streiff, V. (1975). Dictation: A test of grammar-based expectancies. *ELT Journal, 30*(1), 25–36. doi: 10.1093/elt/XXX.1.25

Oller, J. W. (1979). *Language tests at school: A pragmatic approach*. London, UK: Longman.

Oller, J. W. J. (1971). Dictation as a device for testing foreign-language proficiency. *ELT Journal, 25*(3), 254–259. doi: 10.1093/elt/XXV.3.254

Rost, M. (2002). *Teaching and researching listening*. London, UK: Longman.

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

Rost, M., & Ross, S. (1991). Learner use of strategies in interaction: Typology and teachability. *Language Learning, 41*(2), 235–273. doi: 10.1111/j.1467-1770.1991.tb00685.x

Stansfield, C. W. (1985). A History of Dictation in Foreign Language Teaching and Testing. *The Modern Language Journal, 69*(2), 121–128. doi: 10.1111/j.1540-4781.1985.tb01926.x

Wajnryb, R. (1990). *Grammar dictation*. Oxford, UK: Oxford University Press.

Weir, C. J. (2005). *Language testing and valuation: An evidence-based approach*. New York, NY: Prentice Hall.

Williams, J. R. (1998). Guidelines for the use of multimedia in instruction. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, 1447–1451.

Wong, A., & Leeming, P. (2014). Using Dictation to Measure Language Proficiency. *Language Education in Asia, 5*(1), 160–169. doi: 10.5746/LEiA/14/V5/I1/A13/Wong_Leeming

（2016 年 9 月 22 日受領、2016 年 10 月 26 日受理）

（Received Septenber 22, 2016; Accepted October 26, 2016）

APPENDIX
THE DICTATION TEST

Example:

You hear: Are you ready to go?

Question: Are you ( ) ( ) ( )?

Answer: Are you (ready) (to) (go)?

1. Come (1 in) (2 and) (3 sit) down.
2. I'm (4 sorry) (5 I) (6 don't) know.
3. (7 May) (8 I) (9 speak) to you?
4. Do you live (10 with) (11 your) (12 family)?
5. I enjoy (13 playing) (14 tennis) (15 with) you.
6. What are you going (16 to) (17 do) (18 this) weekend?
7. Let's (19 go) (20 and) (21 see) a movie?
8. A lot of (22 people) (23 around) (24 the) world speak English fluently.
9. You should go and see a doctor (25 when) (26 you) (27 feel) sick.
10. Do you know about the (28 dangers) (29 associated) (30 with) smoking?
11. I met some friendly students on the first (31 day) (32 at) (33 school).
12. Have you talked to the (34 manager) (35 about) (36 the) trouble?
13. A large amount of money was (37 spent) (38 on) (39 space) exploration?
14. The boy helped his (40 grandmother) (41 escape) (42 from) a forest fire.
15. Freedom of speech is the most (43 important) (44 thing) (45 in) a democracy.
16. Shortly before an (46 earthquake), (47 animals) (48 are) known to go crazy.
17. They must be hungry (49 and) (50 exhausted) (51 after) a long day of work.
18. The (52 addiction) (53 that) (54 affects) most people is said to be chocolate.
19. The company (55 employees) (56 have) (57 recently) held a strike because they didn't get a pay raise.
20. Computers will be solving a wide range of our (58 current) (59 problems), (60 won't) they?

Developing a Dictation Test to Stream Learners: Quick and Dirty or Smart and Efficient?

Harumi KIMURA

Abstract

This article reports on creating a short and effective dictation test to stream students for classroom purposes. The use of dictation for assessment has long been controversial in SLA. Some researchers have asserted that dictation can only assess word recognition and vocabulary knowledge in context while others insist that it can also be helpful in evaluating structural knowledge or syntactic parsing skills and thus used as an integrative test. Recent research has demonstrated that a dictation test can measure both bottom-up and top-down listening skills—i.e., basic listening proficiency—and can be a valid, reliable, assessment tool (Cai, 2012). For this study, a six-minute partial dictation test was developed, administered, and scored for statistical analyses to demonstrate that such a test, which classroom teachers can create with reasonable ease, can be of good quality: The Rasch person reliability estimate was .94. I provide ten suggestions for making a short and efficient partial dictation test that can be used as a component of placement tests. I recommend teachers make their own partial dictation tests for both assessment and diagnostic purposes.