『中国地区英語教育学会研究紀要』No.18 (1988.9.1)

An Evaluation of a Graduated Dictation Test as a Criterion-Referenced Test

Hiroshima University, Graduate School Toshiaki Takahashi

1. Introduction

According to C. W. Stansfield (1985) dictation is one of the oldest means of testing. Dictation was originally used as a means of transmitting course content from a teacher to pupils. Then, it came into the second language classroom in the sixteenth century. Attitudes toward dictation have been cyclical. For example, in the nineteenth century, the influence of the natural method advocated by Gouin resulted in less popularity of dictation because the natural method assigned minimum importance to reading and writing. But dictation regained its popularity at the end of the nineteenth century due to the infulence of the direct method, and again became popular during the 1930's and the 1940's due to the influence of the reading method. During the 1960's dictation became unpopular again because of the influence of the audio-lingual method. At the same time, dictation tests were criticized for not being discrete-point tests. Lado (1961, p.34) criticized dictation as follows:

> Dictation is favored by many teachers and students both as a teaching and testing device. However, on critical inspection it appears to measure very little of language. Since the order of words is given by the examiner as he reads the material, it does not test words. Since the words are given by the examiner, it does not test vocabulary. It hardly tests aural perception of the examiner's pronunciation, because the words can in many cases be identified by context if the student does not hear the sounds correctly. The student is less likely to hear sounds incorrectly in the slow reading of the words which is necessary for dictation.
>
> (Lado, 1961, p.34)

Since a dictation test is not a discrete-point test, it is "impossible to tell what the results of the test really shows" (River, 1968, pp.290). Similarly, Harris (1969, p.5) regarded dictation as "generally both uneconomical and imprecise" because a dictation test doesn't provide the teacher with much systematic diagnoses about the phonological, grammatical, and lexical weaknesses of the students. Furthermore, Heaton (1975, p.185) complains about the difficulty in interpreting the responses that the students make in taking dictation:

> It is difficult to judge, for example, whether a mistake in a dictation has been made because of the student's inability to
> (i) spell a word
> (ii) "catch" what has been said

(iii) remember a word by the time he writes it
(iv) understand the general context

(Heaton, 1975, p.185)

Thus, attitudes toward dictation remained negative throughout the 1960's. Yet, dictation gained popularity again mainly because of numerous empirical studies conducted by Oller in the 1970's. Since then, dictation has been a popular device in foreign language teaching and testing (again, see C. W. Stansfield (1985), pp.121-7 for a more detailed discussion on the history of dictation). The basic assumption of Oller seems to have been accepted without being greatly questioned. For example, dictation is still being used as a pedagogical device in the Japanese classroom as well as a part of the university entrance examination. However, Cziko (1982, pp.367-9) pointed out three problems of a dictation test. The first problem is that a dictation test has been used as a norm-referenced test with scores interpreted only with reference to group norms. That is, it is difficult to identify in any meaningful way the level of the student's proficiency based on a particular test score, since the test scores are not meaningful in and of themselves. Thus a dictation test is not a criterion-referenced test. The second problem is that dictation "requires considerably more time and care to score than most other tests requiring written responses" such as cloze tests. The third problem is the choice of an appropriate level of passage for a particular group of learners. The first problem is related to the current tendency to construct criterion-referenced tests, as opposed to norm-referenced tests. The scores on norm-referenced tests can only be interpreted in comparison to the scores of the others who have taken the same tests. For example, 80 points out of 100 in a test is regarded as high score when the mean score of the test is well below 80. However, the same score is regarded as a low score when the mean score is well above 80, so it is difficult to determine the level of language proficiency of a given subject without reference to the scores of other subjects. A criterion-referenced test can, on the other hand, tell us how proficient the subject is at given point of time, as well as how proficient the subject has become in the course of his language learning. The following graded criteria are an example of what a candidate might be asked to do in a criterion-referenced test. The number of the questions up to which a candidate can respond correctly may indicate his / her level of language proficiency.

1. The candidate should be able to reproduce *ba* sound.
2. The candidate should be able to reproduce any single word presented in isolation.
3. The candidate should be able to reproduce four words given in a sentence.
4. The candidate should be able to reproduce more than ninety percent of the words in a sentence.
5. The candidate should be able to summarize what is said on a TV programme in English.

Suppose that a candidate could respond yes up to question no.2 two years ago, and now can respond yes up to question no.4. You can easily know the progress of the candidate's language proficiency over the two years without any reference to performance of other

candidates.

In order to overcome the three problems described above (especially, the problem of norm-referenced tests), a series of studies by Cziko (1982, 1984, 1986) were conducted. From the result of these studies, Cziko proposes a "graduated dictation test" as a measure of testing language proficiency. A graduated dictation test is formed of segments of increasing length. Therefore the task of writing down each segment becomes harder and harder. It is then assumed that the longer segments the subject can cope with, the more proficient he is. Thus, Cziko proposed a "graduated dictation" as an excellent testing device of language proficiency based on the results of his studies. This testing technique is new and worth discussing because it may give us a device to measure the proficiency level of the student in a criterion-referenced manner. It may also allow us to know the extent to which the student has become proficient in the course of language learning. However, this testing technique assumes that the length of segments corresponds to the difficulty level of the task or to the proficiency level of the subjects, which may not be the case. The main purpose of this paper is to take a critical look at the assumptions behind graduated dictation tests, based on findings of an empirical study.

2. Purpose and Method

This paper first tries to deal with the procedure of investigating students' response patterns in a graduated dictation test. This will be conducted to see the extent to which the students' response patterns can reveal reliably their "language proficiency" (the problem of whether a graduated dictation test can measure "language proficiency" is beyond the scope of this paper). The graduated dictation test is formed of approximately 12—14 segments of increasing length (see Table 1). It is expected that the difficulty increases in proportion to the length of a segment. For example, Segment 1 is the least difficult because it is formed of the fewest number of words. It is also expected that the more proficient the subject is, the more difficult or the longer a segment he can write down. However, such a complete correlation between

Table 1
Length of Segments and Theoretical Difficulty

Segment No.	No.of Words	Difficulty Level
1	2	least
2	3	difficult
3	3	↑
:	:	:
11	15	↓
12	17	most
13	19	difficult

the length of segment and its difficulty may not be expected. Therefore, in this paper, the

difficulty of the segment is defined as being the number of the subjects who have written it down correctly. Thus it is assumed that S1 is the most proficient because he wrote down accurately every segment in the test passage. Therefore, he is placed first on the chart. It is also assumed that T1 is the easiest task because every subject can do it accurately. Therefore, it is placed on the left of the chart. If Sx has the ability to write down Ty correctly, he should be able to write down T(y-1), T(y-2) ···, T1, since Ty is more difficult than T(y-1), T(y-2) ···, T1. For example, S2 has the ability to write down T4, he is expected to write down correctly T3, T2 and T1 (see Table 2). However, a complete match as shown on Table 2 cannot always be guaranteed because unexpected errors are inevitable as with any other statistical devices. If Sx could write down Ty correctly, but couldn't write down

Table 2
Schematic Presentation of Expected Responses

		easy←				→difficult
		Task				
	Subject	T1	T2	T3	T4	T5
most	S1	1	1	1	1	1
proficient	S2	1	1	1	1	0
↑	S3	1	1	1	0	0
↓	S4	1	1	0	0	0
least	S5	1	0	0	0	0
proficient	S6	0	0	0	0	0

any one or two of T(y-1), T(y-2) ···, and T1, these unexpected responses are said to be *deviations* because the subject either wrote down something he shouldn't have been able to do, or couldn't write down something he should have been able to do. If the chart contains a lot of such deviations, it is difficult for us to be confident in determining the proficiency level of a subject. The subject's proficiency may be higher because he wrote down more difficult segments than the segments he couldn't write down. Or the subject's proficiency may be lower because he couldn't write down less difficult segments than the segments he could write down. Therefore, the number of these unexpected responses or deviations can be used as an index of the extent to which wrong predictions are produced concerning the proficiency level of the subject. The measure of this kind of errors is referred to as the Guttman coefficient of reproducibility (Rep) and is defined by the following formula (Kaiho, 1986, pp.64-5, Hatch and Farhady, 1982, pp.178-9, also see Sato, 1975):

$$\text{Rep} = 1 - \frac{\text{total number of deviations}}{\text{total number of responses}}$$

$$= 1 - \frac{\text{total number of deviations}}{(\text{number of subjects})\ (\text{number of segments})}$$

However, there are two additional steps that must be taken before we can conclude that our scale is real. First, the minimal marginal reproducibility (MMrep) must be computed by adding all the responses which are more frequent in each segment and dividing by the total number of responses. The minimal marginal reproducibility indicates real reproducibility excluding errors and is defined by the following formula (cf. Kaiho 1986, p.66):

$$\text{MMrep} = \frac{\text{number of more frequent responses}}{(\text{number of subjects})\ (\text{number of segments})}$$

The difference between Rep and MMrep is referred to as the percent improvement in reproducibility and is computed by subtracting the latter from the former. The last step is to find the coefficient of scalability, which indicates whether a set of data is scalable, and is computed by the following formula (ibid. p.67, Hatch and Farhady, p.183):

$$\text{coefficient of scalability} = \frac{\%\ \text{improvement in reproducibility}}{1 - \text{MMrep}}$$

In order to conclude that a set of data is scalable, the Rep must be well above 0.90 and the coefficient of scalability must be well above 0.60. The question is whether there is a scale at all.

If the difficulty order on the chart differs greatly from the theoretical order, then it indicates that the length of segment is not a reliable index of the difficulty level of the task.

If no scale is found, then it indicates that the length of segment can not reliably predict the level of language proficiecy.

3. Procedure

A graduated dictation test was administered to 60 EFL (English as a Foreign Language) students at Kagawa University in Japan. The subjects were asked first to listen to the whole passage which was presented without interruptions, at a speed considered normal for a careful oral reading of a text. The second time, the passage was read with pauses for the subjects to write down exactly what they heard. The third time the passage was read with pauses after each sentence to allow the subjects to check their work. After the third reading, the subjects were given 1 minute to correct their work. The test was scored by giving 1 point for each segment written without an error. The passage contained 118 words, and was divided into 14 segments of increasing length, ranging from 3 words to 18 words (see Table 3).

Table 3

Segment No.	Number of Words	AUTOMATION
1	(3)	In today's world
2	(3)	human beings depend
3	(4)	very much upon machines.
4	(5)	The best example of this
5	(5)	is the increase in automation
6	(7)	New, machines do much of the work
7	(6)	that people did fifty years ago.
8	(8)	Machines make soup, assemble cars, and carry messages.
9	(10)	There are even machines which are designed to make hotdogs
10	(10)	and to test the meat for color, flavor, and quality.
11	(11)	The main advantage of automation is that it makes products cheaper.
12	(12)	Automation is efficient, because the machine does the same thing every time.
13	(16)	However, automation also has disadvantages such as the high cost of buying and maintaining the equipment.
14	(18)	Also, the use of machines may replace people, so in the future many workers may lose their jobs.
TOTAL	(118)	

Each segment was formed by dividing the passage at natural points provided by phrase, clause, or sentence boundaries.

The performance of the 60 EFL students on a graduated dictation test is compared with that of 20 ESL (English as a Second Language) students at Georgetown University on the same test which was conducted in April of 1985 by Sasaki Miyuki (unpublished).

Then, the Guttman Implicational Scaling was applied to the results of scores of both the EFL and ESL students on the graduated dictation test in order to first investigate whether there is a scale at all : whether the difficulty level of the task corresponded to the level of his / her language proficiency, and secondly to compare the theoretical difficulty order with the actual difficulty order exemplified on the graduated dictation test.

4 . Results and Analysis

Using the Guttman implicational scaling, the coefficient of reproducibility, the minimum reproducibility, and the coefficient of scalability for the scores of EFL students on the graduated dictation test were respectively 0.868, 0.7, 0.560 (see Table 4 and Table 5). The Rep, the MMrep, and the coefficient of scalability for the scores of ESL students on graduated dictation test are respectively 0.971, 0.955, 0.368 (see Table 4 and Table 6). As to the ESL students, the Rep is well below 0.9. So there is no scale. As to the EFL students, the Rep

is well above 0.9 but the coefficient of scalability is well below 0.6. Again, there is no scale. Therefore, we couldn't reliably predict the level of a subject's language proficiency by

Table 4
Rep, MMrep, Coefficient of Scalability for ESL and EFL Students on the Graduated Dictation Test

	Rep	MMrep	Scalability
ESL	0.868	0.7	0.560
EFL	0.971	0.955	0.368

knowing his position on the chart. Therefore it follows that the difficulty level of the task does not correspond to the level of language proficiency.

Table 5 ESL students

Segment		3	4	6	7	1	2	5	10	11	12	14	9	8	13
Student	S13	1	1	1	1	1	1	1	1	1	1	0	1	①	0
	S12	1	1	1	1	1	1	1	1	1	1	1	0	0	0
	S11	1	⓪	1	1	1	1	1	1	1	1	①	0	0	0
	S20	1	1	1	1	1	1	1	1	1	0	0	0	0	0
	S14	1	1	1	1	1	1	1	1	⓪	0	①	0	0	0
	S08	1	⓪	1	1	1	1	1	1	1	0	①	0	0	0
	S18	1	1	1	1	1	1	⓪	⓪	1	0	①	①	0	0
	S19	1	1	1	1	1	1	⓪	1	①	0	0	0	0	0
	S15	1	1	1	1	⓪	1	1	⓪	0	0	①	0	0	①
	S17	⓪	1	1	1	⓪	⓪	1	1	0	①	0	①	①	0
	S10	1	1	⓪	1	⓪	1	①	0	0	①	0	0	0	0
	S16	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	S06	1	1	1	1	0	0	0	0	0	0	0	0	0	0
	S09	1	1	⓪	①	0	0	0	0	0	0	0	0	0	0
	S03	⓪	1	1	0	①	0	0	0	0	0	0	0	0	0
	S02	1	1	⓪	0	0	0	0	0	0	①	0	0	0	0
	S04	1	⓪	①	0	0	0	0	0	0	0	0	0	0	0
	S01	⓪	①	0	0	0	0	0	0	0	0	0	0	0	0
	S05	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	S07	0	0	0	0	0	0	0	0	0	0	0	0	0	0

○= deviant score

As is evident from Table 6, the graduated dictation test was very difficult for EFL students. They could write down at least some part of each segment correctly, yet not many of the students could reproduce what they heard without an error. Requiring the EFL students to write down each segment without a single error is considered to be one possible explanation for such a result. A different cut-off point in scoring may have produced different test

results for the same test since many students could write down at least a part of each

Table 6 EFL students

Segment		1	4	2	6	3	5	7	8	9	10	11	12	13	14
Student	S12	1	1	1	1	0	0	0	0	0	0	0	0	0	0
	S10	1	1	1	⓪	0	1	0	0	0	0	0	0	0	0
	S40	1	⓪	1	1	①	0	0	0	0	0	0	0	0	0
	S01	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	S33	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	S17	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	S07	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	S42	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	S60	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	S31	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	S20	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	S22	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	S44	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	S09	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	S41	⓪	①	0	0	0	0	0	0	0	0	0	0	0	0
	S49	⓪	①	0	0	0	0	0	0	0	0	0	0	0	0
	S21	⓪	①	0	0	0	0	0	0	0	0	0	0	0	0
	S15	⓪	①	0	0	0	0	0	0	0	0	0	0	0	0
	S39	⓪	①	0	0	0	0	0	0	0	0	0	0	0	0
	S04	⓪	①	0	0	0	0	0	0	0	0	0	0	0	0
	S30	⓪	0	①	0	0	0	0	0	0	0	0	0	0	0
	S45	⓪	0	①	0	0	0	0	0	0	0	0	0	0	0
	S25	⓪	0	①	0	0	0	0	0	0	0	0	0	0	0
	S52	⓪	0	①	0	0	0	0	0	0	0	0	0	0	0
	S02	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	S03	0	0	0	0			0	0	0	0	0	0		
	S05	0	0							0	0				

○=deviant score

segment. Nevertheless, there are many deviations suggesting that the theoretical order is not identical with the difficulty order indicated on the chart (Table 6). As for ESL students, the graduated dictation test is considered appropriate in difficulty. Yet, the theoretical order of the graduated dictation test differs greatly from the actual order shown on the chart (Table 5 and 7). The comparison between the theoretical difficulty order (based on the number of words in a segment) and the actual order (based on the number of segments correctly reproduced) was made using rank order correlation.

Table 7

The Comparison between the Theoretical Difficulty Order (T) of Segments and the Actual Difficulty Order (A) on Scores of ESL Students

Segment	T	A	Segment	T	A
1	1.5	5.5	8	8	13
2	1.5	5.5	9	9.5	12
3	3	1.5	10	9.5	8
4	4.5	1.5	11	11	9
5	4.5	7	12	12	10.5
6	7	3.5	13	13	14
7	6	3.5	14	14	10.5

The theoretical difficulty order correlated with the actual order at the level of 0.732 ($p < 0.05$, $df = 12$). This result indicates that the number of words in a segment has something to do with the difficulty level of each segment. However, the correlation is not strong enough to say that the number of words in a segment alone enables us to predict the actual difficulty level of the segment.

5. Conclusion

As George Miller (1956) wrote in his famous article "The Magical Number Seven, Plus or Minus Two : Some Limits on Our Capacity for Processing Information", we can hold abut 7 chunks in our short-term memory at one time. Therefore, if one word is counted as an independent unit, the listener would possibly have difficulty holding more than 7 words in his / her short-term memory unless he uses some scheme of grouping, such as a single unit into a smaller number of chunks possibly with support from higher-order knowledge such as syntax and semantics. Thus it seems logical to assume that the longer the segment length becomes, the more difficult writing it down would become, and that the longer segment the subject can reproduce correctly, the more proficient he should be. Therefore the graduated dictation test seems to be a criterion-referenced test of "language proficiency" because the amount of information that the subject can hold in the short term memory can be considered as an index of his language proficiency. An analysis of a graduated dictation test, however, shows that longer segments do not necessarily mean more difficult task. It also reveals that the difficulty level of the task cannot be a reliable index of language proficiency since the longest segment the subject can answer correctly does not reliably predict how many of segments he could answer correctly. Therefore this result leads us to doubt the assumption that the number of words that the subject can process at one time can be an index of the level of language proficiency.

In short, a graduated dictation test cannot be considered appropriate as a criterion-referened test. Therefore, it is necessary to devise another kind of criterion-referenced test that tells us how proficient the subject is at a particular point of time as well as how proficient

the subject has become in the course of his language learning. Since manipulation of segment length may not necessarily provide us with such a criterion-referenced test, it is necessary to further investigate fundamental elements that would make such tests possible.

ACKNOWLEDGEMENT

The author is very grateful to Prof. Tatsunori Takenaka at Kagawa University, who was helpful in conducting the experiment, and to Ms. Miyuki Sasaki for providing me with the test material and invaluable data obtained at Georgetown University.

REFERENCES

Cziko, Gary A. (1982)"Improving the Psychometric, Criterion Referenced, and Practical Qualities of Integrative Tests"; *TESOL Q*, 16, pp.367-379.

Cziko, Gary A. (1984)"An Improvement over Guttman Scaling : A Computer Program for Evaluating Cumulative, Non-parametric Scales of Dichotomous Items" *Journal of Education and Psychological Measurement*, 44, pp.157-161.

Cziko, Gary A. (1986)"Determining the Reliability, Validity, and Scalability of the Graduated Dictation Test", *LL*, 35, 4, pp.556-566.

Harris, David, P. (1969) *Testing English as a Second Language*, McGraw-Hill Book Company

Harris, David, P. (1970)"Report on an Experimental Group-Administered Memory Span Test," *TESOL Q*, 4, September, 3, pp.203-213.

Hatch, Evelyn & Hossein Farhady (1982) *Research Design and Statistics for Applied Lingustics*, Newbury House Publishers, Inc.

Heaton, J. B. (1975) *Writing English Language Tests*, Longman Group Limited.

Lado, Robert (1961) *Language Testing*, Longman.

Miller, G. A. (1956)"The Magical Number Seven, Plus or Minus Two : Limits on Our Capacity for Processing Information", *Psychological Review*, 63, pp.81-97.

Oller, J. W. Jr (1979) *Language Tests at School*, Longman.

Stansfield, Charles W. (1985)"A History of Dictation in Foreign Language Teaching and Testing", *The Modern Language Journal*, 69, 2, pp.121-128.

海保博之(1986) 『心理・教育データの解析法10講』 (応用編) 福村出版.

佐藤隆博(1975) 『S P表の作成と解釈－授業分析・学習診断のために』 明治図書.