『中国地区英語教育学会研究紀要』No.18 (1988.9.1)
〈第13回全国英語教育学会自由研究発表〉

An Evaluation of Dictation as a Means of Testing

Hiroshima University, Graduate School Toshiaki Takahashi

1. Introduction

Dictation is one of the oldest means of testing. It came into the second language classroom in the sixteenth century. Attitudes toward dictation have been cyclical : positive, then negative, then positive again (see C. W. Stansfield(1985)). Dictation gained popularity again in the 1970's because of numerous empirical studies conducted by Oller. Since then, dictation has been a popular device of foreign language testing, especially in the testing of auditory related skills such as listening comprehension. Yet, questions of whether this testing technique is valid, and whether the theory that supports this technique is appropriate have been often overlooked in the second language classroom. This paper deals with these two questions concerning dictation tests.

2. Validity of Dictation Tests.

Usually common kinds of test validity includes face validity, content validity, concurrent validity, construct validity. Face validity refers to the extent to which a test looks like it measures what it purports to measure. Yet, face validity is often determined impressionistically without an empirical bases, since there is no statistical measure of face validity. (cf. Henning 1987, pp. 94–96, Palmer and Bachman, 1981, pp. 135–6). Therefore, an examination of face validity is not included in the present paper, which focuses on the other three types of test validity.

2.1. Content Validity

Content validity is related to the question of whether a test requires the examinee to perform tasks that are really the same or fundamentally similar to the sorts of the tasks one normally performs in exhibiting the skills or ability that the test purports to measure (Oller, 1979, pp. 50–51). Then, specification of the ability that the test purports to measure should be required as a logical necessity in order to investigate content validity. However, in the case of dictation it is not at all clear exactly what ability is being tested.

One way to tackle the problem of content validity is to examine whether dictation is a task which people normally perform when they use a language. Oller claims (1971, pp. 257–8) that dictation tests basic language processing mechanism (analysis by sinthesis), and provides a comprehensive sampling of structural and lexical items in a meaningful context. However, a dictation test doesn't give "any convincing proof of the candidate's ability to actually use a language, to translate the competence (or lack of it) which he is demostrating into actual performance in ordinary situations" (Morrow, 1979, pp. 148–9). Since the task required in dictation tests is not a part of everyday language use, it cannot be considered content-valid (see Clark, 1983, p. 433).

Similarly, Canale and Swain (1980) and Canale (1981) argue that dictation testing requires

verbatim recall and writing down an auditory representation of what has been said, hardly reflecting a task which we normally perform in everyday language use. Furthermore, the subject is required to pay too much attention to the surface features rather than to the meaning of the dictation passage (see Heaton, 1975, pp. 185–6). There are findings that show that listeners remember the meaning rather than the surface features of the dictation passage (see J. Anderson (1974), Jarvella (1970, 1971), Sachs(1967)). Jarvella(1970, 1971), for instance, demonstrated that people retained the general meaning of the passage rather accurately, but their verbatim recall was poor soon after a sentence boundary. The result of these experiments show that people remember words better if they are from the part of the sentence being processed. At the same time, the findings also indicate that listeners can indeed retain the surface features of what they heard unless it exceedes a sentence boundary. Therefore, it is possible for listeners to write down what they heard verbatim (possibly with the help of rehearsal) if required, and if the ability of verbatim recall is what the test exactly purports to measure. However, listeners are not normally required to hold the surface features of the sound input exactly. They forget them soon after they finish processing the input. Therefore, dictation is considered to be a more demanding activity than the sort of task that people normally need to do. It is necessary to prove that a dictation which requires verbatim recall may still be a good measure of, say, overall second language proficiency.

There is also a problem concerning the procedure of dictation tests from the viewpoint of language use. In a dictation test, students are asked first to listen to a test passage. The second time the same test passage is read with pauses for them to write down exactly what they have heard. The third time, they are asked to listen to the passage without pauses while they check what they have written down. The procedure doesn't seem to resemble any of what people do in normal language use. However, Cohen (1980, p. 111) argues for the procedure of dictation tests, citing an example of note taking :

> One such task is note taking. Students hear information, try to process it accurately, and write down what they have heard. It is true that in real communicative situations, the learner may write down a translation of salient points, whereas the dictation task calls for complete and exact reproduction of what is said. But unlike the typical lecture-hall situation, telephone call, or whatever, the learner is given three opportunities to hear the dictated material, the second time with pauses between chunks of information. Such repetition is a "luxury" in the real world, particulary when student listeners are unable to stop a lecture, say every time they have a question as to what the lecturer has said.
>
> (Cohen, 1980, p. 111)

His argument can be considered invalid because of his inappropriate interpretation of the word "accurately" and of what Cohen calls a "luxury". The student may be required to pay careful attention to the content of what is said, but is not required to exactly reproduce all the surface features by which the content is conveyed as Cohen himself pointed

out. What Cohen calls a "luxury" actually means that the students are not doing what they do in normal language use. Furthermore, unlike cloze tests, dictation tests usually ask the listener to reproduce exactly what has been said (only one correct response), as opposed to what might have been said (one of many possible responses derived from the context). This indicates that this testing method focuses on the surface features rather than the meaning of the test passage. If the primary focus of dictation is to be on the meaning, then the counting-backward task, a task designed to prevent rehearsal in short-term memory, should be included after each pause. Alternatively passage recall, a task that requires the listener to recall freely and as much about the passage as possible could be used because this task may force the listener to remember the content of the passage rather than the means by which the content is conveyed. Yet, there is one problem to the second alternative : the students may still try to remember the surface features of the passage rather than focus on the meaning though he cannot remember all of the passage verbatim. In sum, dictation tests don't seem to be content-valid.

2.2. Concurrent Validity and Construct Validity

2.2.1. Concurrent Validity

Concurrent validity is usually determined by computing the extent to which two tests that purport to measure the same skill correlate statistically with each other. For example, in order to determine the concurrent validity of a listening comprehension test one could determine the correlation between scores of testees on the test with their scores of a test which has already been validated and known to be reliable as a measure of listening comprehension. Oller (1972, pp. 346–54) claims that dictation tests can measure overall language proficiency based on high correlations (0.88) with traditional discrete-point tests (the University of California at Los Angeles (UCLA) ESLPE (English as a Second Language Placement Examination). The UCLA examination consisted of five parts : 1) a dictation, 2) a composition, 3) a vocabulary test, 4) a phonology test based on the discrimnation of minimal pairs, 5) a grammar test requiring the identification of correct or incorrect sentences :

> The surprising result was that the dictation correlated more highly with each section of the test than did any other section. In other words, when the correlation between parts were rank ordered, the dictation came out first in every possible category. On the basis of these data, the dictation clearly seems to be the best single measure of the totality of the language skills tested by ESLPE Form 1.
>
> (Oller, 1972, pp. 347–8)

Irvine, Atai, and Oller (1974) intercorrelated the scores on dictation of 159 students of English as a foreign language in Iran with their scores of the various sections of the Tests of English as a Foreign Language (TOEFL) and reported that dictation as well as cloze correlated more than any other sub-sections of the test with the Listening Comprehension which is said to be a highly integrative task, and the total score of the test. Similarly, Oller and Streiff (1975) reported that the dictation test correlated

highly with a traditional discrete-point test (UCLA English Language Institute Placement Battery), thus proposing dictation tests as a measure of overall language proficiency.

Thus, dictation tests were claimed to have concurrent validity. However, there are problems concerning correlational studies of this kind. These points will be further discussed in the following sections.

2.2.2. Construct Validity

Construct validity is related to the following process of validating a test :

> Construct validation begins with a psychological construct that is a part of formal theory. The theory enables certain predictions about how the construct variable will behave or be influenced under specified conditions. The construct is then tested under the conditions specified. If the hypothesized results occur, the hypotheses are supported and the construct is said to be valid.
>
> (Henning, 1987, p. 98)

Since the theory of dictation tests is proposed in an elusive way, and is not specified enough to make predictions under which the theory can be tested, it is difficult to see to what degree the items in a dictation test reflect the essential aspects of the theory on which the test is based. However, there is one exception by which Oller claims that a dictation test is construct-valid : the high correlation of a dictation test with traditional discrete-point tests which have been cited above in 2.2.1.. Oller (1979) summarised the evidence from many of the correlational studies and the principal component construct validation studies, and examined the following three hypotheses based on it (pp. 425—458) :

1) The Divisibility Hypothesis (H_1)
2) The Indivisibility Hypothesis (H_2) (or The Unitary Competence Hypothesis)
3) The Partial Divisibility Hypothesis (H_3)

The first hypothesis means that language proficiency is divisible into a number of distinct components, such as knowledge of phonology, knowledge of vocabulary, knowledge of grammar, listening and reading. The second hypothesis means that language proficiency cannot be broken down into a number of subcomponents which can be independently measured. The third hypothesis means that a major portion of test variance is shared by all tests, but the small amount of variance can be attributed to another specific factor that various tests don't measure in common. After investigating these hypotheses based on the data, Oller concluded that the second hypothesis can better explain the data than the other two hypotheses, thus the unitary competence hypothesis is supported. Oller explains why a single test (or factor) can explain the whole test variance : the central component of language competence is what he calls "expectancy grammar". Therefore tests that measure language competence should highly correlate with each other. Oller also argued that since this hypothesis is proved to be correct, then the theory on which the dictation is based has an independent claim to validity, and that

the test has construct validity. (cf. Oller, 1981, p. 127).

However, both the correlational studies and the principal component construct validation studies have been criticized (see, Farhady, 1983, Vollmer and Sang, 1983 and G. Hatano, 1987). Otomo (1981, p. 10) pointed out two problems of the correlational studies. The first problem is related to the interpretation of the coefficient of correlation. Many books on psychology state that if two tests correlate with each other at the point of more than 0.70, both tests are regarded as being "highly correlated" with each other, and are said to be measuring the same thing. Yet, what the coefficient of 0.70 really means is that about fifty (i.e. the square root of 0.70) percent of the total variance can be explained by one of the tests. The remaining fifty percent either cannot be accounted for or can be explained by some other factors.

The second problem refers to the interpretation of correlation. What correlation really means is that one is related to the other, not that one is the cause of the other. If this is so, what is the meaning of the high correlation found between dictation and other tests?

Recall that Oller argued for the unitary competence hypothesis based on the evidence from the principal component construct validation studies. These principal component construct validation studies were also criticized. Bachman and Palmer criticized this technique and, based on their research, made clear that second language proficiency cannot be accounted for by a single qeneral factor:

> One general problem is that principal component analysis cannot be used to examine any kind of structural model in which the elements in the model are correlated (as appears to be the case in models of language proficiency). The reason for this is that principal component analysis looks only at variance structure, not covariance structure..... Another general problem is that of commonalities—this is, the amount of variance the analysis attributes to something the various measures have in common. The reason this is a problem is that the common variance in a principle component analysis contains measurement error and method variance, which inflate the magnitude of the common variance.
>
> (Bachman & Palmer, 1981, p. 138)

Therefore, Bachman and Palmer (1981) investigated the construct of second language proficiency by using multitrait-multimethod convergent-discriminant design which overcomes the problems cited above. Test scores are expected to reflect not only what it is that one is attempting to measure (the trait), but also the effects of the methods of measurement. In order to assess the relative contribution of trait and method to test scores, at least two or more traits must be measured by a minimum of two distinct methods. The results of their study rejected the hypothesis that a single language variable underlies language proficiency. This indicates that dictation tests lack construct validity.

A special issue of *The English Journal* (1982, pp. 19—21) reported how third-grade senior high school students (1,013) with no experience of studying abroad perform differently

in the Michigan Test from those with more than half year experience of studying abroad (59). The results (see, Table 1) indicate that the second language proficiency of Japanese students with no experience of studying abroad cannot be explained by any one of the variables, since any single variable does not correlate highly with any other variables. A high score in listening doesn't necessarily mean a high score in structure or reading, thus suggesting each variable is measuring an independent factor, not a single factor of the second language proficiency. Hatano (1987, p. 15), based on his experiment, suggests that Japanese learners seem to acquire language skills independently rather than integratively.

Table 1

Correlation Matrix for Michigan, structure, vocabulary, reading, listening*

	Structure	Vocabulary	Reading
Vocaburary	0.489 (0.773)	———	———
Reading Comprehension	0.453 (0.759)	0.430 (0.700)	———
Listening	0.411 (0.875)	0.388 (0.674)	0.361 (0.705)

* The figures in parenthesis refer to the coefficient of correlation for students with experience of studying abroad

2.3. A Small Experiment

A small experiment was conducted in order to test the hypothesis that a single variable can account for second language proficiency, thus resulting in higher correlations between dictation tests and other tests of language proficiency. The experiment examined whether the scores of 60 EFL students at Kagawa University on a dictation test correlate with their scores on subsections of CELT using product moment correlation. In the dictation test, the subjects were asked to first listen to the test passage which was read at conversational speed. The second time the passage was read with pauses for them to write down exactly what was said. The third time they heard the same passage without pauses while they checked what they had written down. The topic of the test passage was of general interest, and the test passage is taken from *Intermediate Stories for Reproduction : American Series* in which grammatical structures as well as vocabulary are carefully controlled to be appropriate to students of lower-level proficiency. The segments were formed by dividing the passage at natural points provided by phrase, clause, or sentence boundaries. The test was scored by giving 1 point for each word written without an error.

Six people were traveling in a compartment on a train./Five of them were quiet and well behaved/but the sixth was a rude young man/who was causing a lot of trouble to the other passengers./

At last this young man got out at a station/with his two heavy bags./None of the other passengers helped him,/but one of them waited/until the rude young man was very far away,/and then opened the window and shouted to him./"You left something behind in the compartment!"/Then he closed the window again./

The young man turned around/and hurried back with his two bags./He was very tired when he arrived/but he shouted through the window,/"What did I leave behind ?"/

As the train began to move again,/the passenger who had called him back/opened the window and said,"A very bad impression!"/

(The slash indicates each pause.)

(Hill, 1980, p. 56)

The results obtained show that the dictation test correlated with CELT Listening at the point of 0.481 ($p < 0.001$, $df=59$) and with CELT Structure at the point of 0.360 ($p < 0.01$, $df=59$). Although the dictation test has some relationship both with CELT Listening and CELT Structure, the correlation is not strong enough to make reliable predictions of how a subject would perform in CELT Listening and CELT Structure, based on his performance in the dictation test. It seems that dictation tests are not measuring the same ability as measured in other tests, such as CELT Listening, and CELT Structure. One may argue that these tests cited above do not tap underlying linguistic competence or that they do so to an insufficient extent (Oller and Streiff, 1975, p. 33), thus resulting in low intercorrelation. Nevertheless, when these tests are used to determine the second language proficiency of Japanese students, any one of the tests alone is not sufficient to explain the total variance of the second language proficiency because each of their language skills seems to develop independently unlike ESL students or those who have been exposed to English for a longer time in a natural setting (see Table 1.).

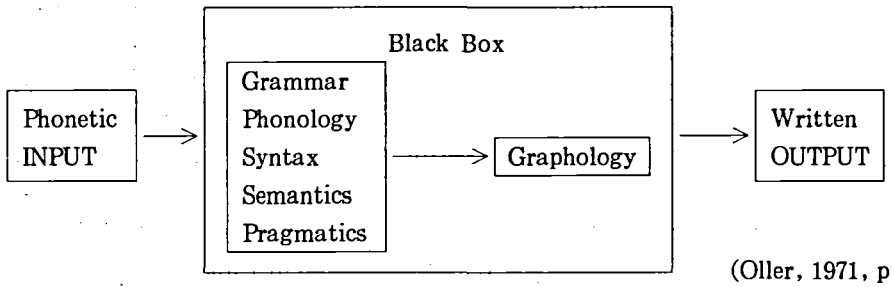
An examination of the concurrent validity and the construct validity of dictation tests shows that second language proficiency may not be composed of a single variable (see Bachman and Palmer, 1981, 1982) and therefore cannot be measured by a single test, such as dictation tests. Thus it seems that dictation tests don't have construct validity. Oller reported high correlations between dictation tests and other tests (see. 2.2.1.). However, the correlations reported cannot be considered high enough to conclude that these tests are measuring the same thing (see Otomo, 1981). The correlations obtained between a dictation test and CELT above were even lower. Thus it seems that dictation tests don't have concurrent validity.

3. The Theory of a Dictation Test : What Makes Dictation Work ?

This chapter deals with the theory of dictation as a test measurement and its problems with predominant focus on dictation as an auditory related task since the notion of dictation as a test of overall language proficiency is doubtful. Oller (1979. p. 266) explains the reason that dictation works well is that a dictation test faithfully reflects *crucial aspects* of the very activities that one must normally perform in processing the language auditorily. If a dictation test is basically the sort of task that theories of language processing defines as characteristic of human discourse, then the dictation test may still be a good device for measuring language proficiency, however unfavourable the conclusion of the previous discussions may appear. Therefore it is also important to examine the theory of dictation as means of testing from the pychological perspective. This chapter first deals with the theory of dictation tests proposed by Oller, then compares the theory with some of the findings of research studies on speech perception.

3.1. Oller's View

According to Oller (1971), the process involved in dictation tests is as follows :



(Oller, 1971, p. 258)

It is suggested that the complex interactions between phonology, lexicon, grammar and graphonology as indicated in the schematic representation, are required for writing a dictation. The dictation test measures the student's ability to (a) discriminate phonological units, (b) make decisions concerning word boundaries in order to discover sequences of words and phrases that make sense, i.e. that are grammatical and meaningful, and (c) translate this analysis into a graphemic representation. Oller meant by (b) that the listener actively participates in listening to speech : Oller and Streiff (1975, p. 34) suggest that the listener first formulates his synthesis (hypothesis) based on 'grammar' generated expectancies, and compares the synthesis with the incoming sound sequence (however, note that what Oller meant by grammar includes semantic and pragmatic facts as well). If the synthesis is not radically at variance with the acoustic materials, then the synthesis will be accepted. This process is called 'analysis by synthesis'.

Anguing against Lado's (1961, p. 34) view of dictation testing, Oller (1972) suggests that a dynamic process of analysis by synthesis is involved in dictation testing based on (1) his personal observation of students errors in taking a dictation (2) familiar problems of speech perception, such as coarticulation, acoustic-phonetic non-invariance and, consequently, the difficuly of segmentation :

> Even in briefly glancing at the errors students make in taking dictation it becomes quite clear that the student does not merely hear words in a particular order and write them down. Rather, he hears sound sequences bounded occasionally by silence or pauses, which are otherwise strung together without obvious boundaries between them ; he actively sequences into word, phrases, and sentences that make sense to him. Clearly, common errors suggest a dynamic process of analysis-synthesis. The student not only receives auditory information, but he processes this information in order to generate a sentence (or a sequence of them) that has meaning.
>
> (Oller, 1972, pp. 351−2)

Although he argues that a dynamic processes of analysis-synthesis is involved in dictation, no direct evidence supports his argument. What the errors in taking dictation really mean may be that "analysis-synthesis", even if this dynamic process is actually involved, does not guarantee accurate perception. The question of when and what the role process of analysis-synthesis can play in the task is not clear.

The problems of speech perception referred to above ((2)) include 1) lack of invariance condition, 2) the problem of coarticulation, 3) lack of linearity conditions. Lack of invariance condition refers to the condition that phonetic segments do not have invariant properties (Clark and Clark, 1977, p. 176) which is called acoustic-phonetic noninvariance. There is no simple one-to-one mapping of sound units onto phonetic units (Sawusch, 1986, p. 52). The problem of coarticulation refers to parallel transmission and context conditioned variation (Pisoni and Luce, 1986, p. 4). The term parallel transmission refers to the tendency for the phonemes to be sent in parallel (Matlin, 1983, pp. 133−4). Thus, each phoneme is not pronounced in isolation, because its sound is modified by the surrounding phonemes. The term context conditioned variation refers to the problem of variability resulting from coarticulation that presents enormous problems for segmentation of the speech signal into phonemes or even words based only on analysis of the physical signal. Adjacent phones are typically coarticulated so that there is no single point that can be identified as dividing the two (Samuel, 1986, p. 92.). Lack of linearity condition refers to the tendency that phonetic segments are not identified sequentially (cf. Clark and Clark, 1977, p. 176. see also Pisoni and Luce, 1986, p. 19).

3.2. Active Models of Speech Perception

Resulting from studies of the problems described above, active models of speech perception such as motor theory, and analysis by synthesis were proposed. However, today early models of speech perception such as motor theory and analysis by synthesis are no longer considered to be appropriate (Pisoni and Luce, 1986, p. 30). Motor theory of speech perception was once proposed as a powerful model of speech perception to overcome the problems such as lack of invariance. However, categorical perception which was often cited as a support for the theory is now put in doubt, and as the link between empirical data and theory is not strong, the theory is not considered to be appropriate. (ibid. pp. 14−15 & pp. 30−1). Similarly, analysis by synthesis is criticized because little direct empirical evidence has been found to support the model (ibid p. 30). Neisser (1967) suggested

that humans perceive speech by the active process of analysis by synthesis based on the study of Hall and Stevens (1964). Then, Oller based on Neisser (1967) claimed that this active process of analysis by synthesis is involved in dictation. However, Neisser (1976) admitted that his argument may not be correct. As he pointed out himself, it seems incorrect to suppose that the listener first formulates hypothesis about what comes next and modifies his original hypothesis only when it is greatly at variance with the acoustic materials, and formulates another hypothesis. If the process of analysis by synthesis is involved in speech perception, incrediblely large number of incorrect hypotheses would be generated as a logical necessity. Since the theoretical model of analysis synthesis is put in doubt, the validity of dictation based on such a pychological model is also questionable.

3.3. Interaction of Knowledge Sources

Later, new active models of speech perception began to suggest the interaction of various knowlege sources involved in speech perception : overriding support from higher-order knowledge is essential because the acoustic signal is so impoverished and noisy as the problems such as coarticulation seem to indicate. Thus, these models assign an important role to a higher order knowledge such as context (see Cole & Jakimik (1980), Garnes & Bond (1975), Warren & Warren (1970), Pollack & Pickett (1964) for more information on the role of context on speech perception). If higher order knowledge has an important role to play in speech perception, dictation can still be a good test of language proficiency, in spite of the fact that the theory of analysis by synthesis, on which dictation claimed validity, is no longer considered appropriate. However, the role of higher order knowledge, such as context is not clear. For example, Garnes & Bond (1975, pp. 214—225) collected the following examples of misperception found in casual speech.:

Original	*Misperception*
wrapping service	wrecking service
meet Mr. Anderson	meet Mr. Edison
I'm covered with chalk dust	I'm covered with chocolate
get some sealing tape	get some ceiling tape

(quoted from Clark & Clark 1977 : 214)

Clark & Clark (1977) pointed out that although most misperceptions bear some phonetic relation to the original, many of the changes went beyond explicable changes, seeming to be as much determined by sense as by sound. However, note that one may interpret the result in a different way : a misperception at the phonetic level finally leads to a mismatch at the semantic or syntactic level. Takahashi (1987, pp. 17—24) asked 17 university students of English as a Foreign Language to write down a dictation from tapes in their own time. An analysis of students' errors in the dictation revealed that while 91 percent of errors had some phonetic relation to the original, only 15.4 percent of their errors match the original both at the semantic and syntactic levels. Thus the subjects errors didn't seem to be determined as much by sense as by sound. Furthermore, there are several studies that suggest that overiding support from higher order knowledge is not necessarily

essential. On the problem of acoustic-phonetic invariance, for example, one-to-one mapping is not necessary as long as phonetic perception is viewed as the probalistic process of matching phonetic features to prototype representations in the memory (Massaro and Oden (1980), p. 131). The problem of corarticulation or segmentation too can be solved without resorting to higher-order knowledge (Samuel (1986), p. 94). On non-linearity conditions there is an experiment that demonstrated left to right processing effect within words (Cole (1973)). The role of higher-order knowledge is further questioned by a spectrogram-reading experiment. Cole, Budnickey, Zue, and Reddy (1980) reported that without prior knowledge of the specific words present, or the sentence context, Zue, an expert spectrogram reader could produce a phonetic transcription of about 90% of all phonetic segments (Pisoni and Luce, 1986, p. 26). Based on this, Pisoni and Luce suggest that highly accurate bottom-up phonetic analysis of an utterance is actually possible without resorting to higher-order knowledge, such as prosodic, syntactic, semantic, and pragmatic knowledge. Thus it is necessary to reevaluate the longstanding assuption that the speech signal is so "noisy" that speech perception is only possible with overiding support from higher order knowledge, which is actively used to generate lexical hypothesis (ibid. p. 26). If this is so, what does dictation measure? Dictation has been considered as an excellent testing device, with the assumption that writing down dictation unavoidably requires much support from higher order knowledge. However, if higher order knowledge is not required in writing down dictation, what is the real value of dictation as a testing device of English language proficiency?

4. Conclusion

In section 1, the validity of dictation tests was examined. The content validity is questioned mainly because what people do in taking dictation is not the kind of activity that people normally perform in day-to-day language use. The concurrent validity together with the construct validity of dictation are also questionable. It may be possible to argue that when a low correlation is found between language tests, one of the tests being correlated, such as CELT Listening does not tap underlying linguistic competence. Nonetheless, since high correlation between two tests does not necessarily mean they are measuring the same competence, and since the second or foreign language proficiency cannot be accounted for by a single factor, it seems that dictation is not measuring overall language proficiency. The study of Bachman and Palmer (1981) support the partially divisibility hypothesis. Thus, the four skills can and should be independently measured.

Some may argue that a dictation test measures listening comprehension in addition to the knowledge of the lexicon, and the adility to discriminate sounds. This problem is discussed in section 3. If writing down dictation unavoidably requires the listener to comprehend the spoken utterances, dictation can be regarded as a test of listening comprehension. Oller suggested that the listener expects what comes next in the spoken discourse based on the grammar of expectancy. Since the words are run together in nomal speech, and are consequently difficult to segment, such grammar generated expectations are said to be essential. The grammar of expectancy enables the listener to formulate a hypothesis about

what comes next in spoken utterances. The hypothesis is then be compared with the sound input. Thus the process of speech perception was explained by the active process of analysis by synthesis, which is no longer considered appropriate.

Let us forget the inappropriate argument of Oller for a while, and suppose that writing dictation requires understanding the meaning of the spoken utterances. Many books on pychology propose the interactive model of speech perception such as Cole and Jackimik suggest. A model of this kind assumes that higher order knowledge such as syntax and semantics is essential for speech perception. However, as Samuel (1986) and others pointed out, basically the bottom-up class of theory of speech perception may be possible. This implies that dictation is a test of lower-order skills such as decoding acoustic input, possibly with support of the knowledge of the lexicon, rather than a test of comprehending the meaning of the spoken input. One may argue that one has to rely on higher-order knowledge such as syntax and semantics in order to "chunk" perceived words and keep them in the short term memory until he has finished writing them down. As George Miller (1956) wrote in a famous article "The Magical Number Seven, Plus or Minus Two : Some Limits on Our Capacity for Processing Information", we can hold aboud 7 chunks in the 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 the short-term memory unless he uses some scheme of grouping such a single unit into a smaller number of chunks. At the first glance it seems that a dictation test can be a test of comprehending the spoken utterances since higher order knowledge is actually necessary for holding in the short term memory more than 7 words of sound stimulus. However, this is not correct. The process of comprehending the spoken utterances is normally carried out in a different way : even if the listener can hold in the short term memory the surface features of spoken utterances of more than seven words, this does not mean that the listener is processing the spoken utterances in the same way. As soon as the listener takes propositions out of the perceived words, the surface features of these words are soon erased from the short term memory, while the meaning may be stored in long term memory. Thus dictation requires the listener to do a more demanding activity of holding the surface features in short term memory. It is possible to test directly how well the listener can comprehend the spoken message by simply asking questions on some of its propositions. There is no need to indirectly infer the level of listening comprehension by asking the listener to do what isn't necessary for understanding what he heard. If the purpose of the test is to know how much information one can store in the short term memory, possibly with the help of higher order knowledge, we should rather ask the student first to read the sentence or part sentence, then to reproduce what he has read. Yet it remains to be shown that the the performance on such a test is in any meaningful way related to a pychological ability such as listening comprehension.

In short, dictation test are better left unused since there are many problems as discussed so far. Yet, it is necessary to investigate the relationship between the ability to hold speech verbatim and the role of higher-order knowledge so that we may know more clearly what kind of ability can be measured by dictation.

ACKNOWLEDGEMENT

The author is very grateful to Prof. Tatsunori Takenaka at Kagawa University, who was helpful in conducting the experiment.

BIBLIOGRAPHY

Anderson, J. R. (1974) "Verbatim and Propositional Representation of Sentences in Imediate and Long-term Memory," *Journal of Verbal Learning and Verbal Behavior*, 13, pp. 149–162.

Bachman, L. & A. Palmer (1981) "The Construct Validation of the FSI Oral Interview", *Language Learning*, 31, pp 67–86.

Canale, M. (1981) "From Communicative Competence to Communicative Language Pedagogy", in J. C. Richards and R. Schmidt (eds.) *Language and Communication*. London : Longman.

Canale, M. & M. Swain (1980) "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing", *AL*, 1, pp. 1–47.

Clark, John L. D. (1983) "Language Testing : Past and Current Status–Directions for the Future", *Modern Language Journal*, 67, 4, pp. 430–443.

Clark, Herbert H. & Eve V. Clark (1977) *Psychology and Language : An Introduction to Psycholinguistics*. Harcourt Brace Jovanovich.

Cohen, A. D. (1980) *Testing Language Ability in the Classroom*, Newbury House Publishers, Inc.

Cole, R. A. (1973) "Listening for Mispronunciations : A Measure of What We Hear juring Speech", *Perception and Psychophysics*, 14, pp. 153–156.

Cole, R. A. and J. Jakimik (1980) "A Model of Speech Perception", in R. A. Cole (ed.) *Perception and Production of Fluent Speech*, Hillsdale, NJ : Erlbaum, pp. 133–163.

Cole, R. A., A. I. Rudnicky, V. W. Zue, and D. R. Reddy (1980) "Speech as Patterns on Paper", in R. A. Cole (ed.) *Perception and Production of Fluent Speech*, Hillsdale, NJ : Erlbaum, pp. 3–50.

Farhady, Hossein (1983) "On the Plausibility of the Unitary Language Proficiency Factor", in John W. Oller, Jr. (ed.) *Issues in Language Testing Research*, Newbury House Publishers, Inc., pp. 11–28.

Garnes, S. and Z. S. Bond (1975) "Slip of the Ear : Errors in Perception of Casual Speech", in *Papers from Eleventh Regional Meeting, Chicago Linguistic Society*, pp. 214–225.

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

Henning, G. (1987) *A Guide to Language Testing*, Newbury House Publisher, Inc.

Hill, L. A. (1980) *Intermediate Stories for Reproduction : American Series*, Oxford University Press.

Jarvella, R. J. (1970) "Effects of Syntax on Running Memory Span for Connected Discourse", *Psychonomic Science*, 19, pp. 235–236.

Jarvella, R. J. (1971) "Syntactic Processing of Connected Speech", *Journal of Verbal Learning and Verbal Behavior*, 10, pp. 409–416.

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

Massaro, D. W. and G. C. Oden (1980) "Speech Perception : A Framework for Research and Theory," in N. J. Lass (ed.) *Speech and Language : Advances in Basic Research and Practice* Vol. 3, New York : Academic Press, pp. 129–165.

Matlin M. (1983) *Cognition*, GBS College Publishing.

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

Morrow, K. E. (1979) "Communicative Language Testing : Revolution or Evolution ?" in C. Brumfit and K. Johnson (eds.) *The Communicative Approach to Language Teaching*, London : Oxford University Press, pp. 143−57.

Neisser, U. (1967) *Cognitive Psychology*, New York : Appleton.

Neisser, U. (1976) *Cognition and Reality*, Freeman.

Oller, J. (1971) "Dictation as a Device for Testing Foreign Language Proficiency", *English Language Teaching Journal*, 25, 3, pp. 257−8.

Oller, J. W. Jr (1972) "Dictation as a Test of ESL Proficiency", in Allen, H. B. and R. N. Campbell (eds.) *Teaching English as a Second Language*, Second ed., McGraw-Hill, pp. 346−354.

Oller, John. W. Jr. and P. Atai (1974) "Cloze, Dictation, and the Test of English as a Foreign Language," *LL*, 24, 2, Dec., pp. 245−252.

Oller, John. W. Jr. and V. Streiff (1975) "Dictation : A Test of Grammar-Based Expectancies", *ELTJ*, 30, 1, pp. 25−36.

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

Oller, J. W. Jr. (1981) "Language Testing Research (1979−80)", in R. B. Kaplan, R. L. Jones, G. R. Tucker (eds.) *Annual Review of Applied Linguistics*, Newbury House Publisher Inc., pp. 124−150.

Pisoni, D. B. and P. A. Luce (1986) "Principal Issues in Speech Perception", in Schuwab, E. C. and H. C. Nusbaum (eds.) *Pattern Recognition by Humans and Machines : Volume 1, Speech Perception*, Academic Press, Inc., pp. 1−50.

Pollack, I. and J. M. Pickett (1964) "Intelligibility of Excerpts from Fluent Speech : Auditory vs. Structural Context", *Journal of Verbal Learning and Verbal Behavior*, 3, pp. 79−84.

Samuel, A. G. (1986) "The Role of the Lexicon in Speech Perception", in Schuwab, E. C. and H. C. Nusbaum (eds.) *Pattern Recognition by Humans and Machines : Volume 1, Speech Perception*, Academic Press, Inc., pp. 89−111.

Sachs, J. R. (1967) "Recognition Memory for Syntactic and Semantic Aspects of a Connected Discourse", *Perception and Psychophysics*, 2, pp. 437−442.

Sawusch, J. R. (1986) "Auditory and Phonetic Coding of Speech", in Schwab, E. C. and H. C. Nusbaum (eds.) *Pattern Recognition by Humans and Machines : Volume 1, Speech Perception*, Academic Press, pp. 51−88.

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

Vollmer, H. J. and F. Sang (1983) "The Competeing Hypotheses about Second Language Ability : A Plea for Caution", in John W. Oller Jr. (ed.) *Issues in Language Testing Research*, Newbury House Publishers, Inc., pp. 29−79.

Warren, R. M. and R. P. Warren (1970) "Auditory Illusions and Confusions", *Scientific American*, 223, pp. 30−6.

アルク(1982)別冊 The English Journal :『英語教師読本』, pp. 19−21.

大友賢次(1981)「英語能力の構造」『英語教育』大修館書店, 30, 9, 11月, pp. 9−11.
高橋俊章(1987)「The Function of Stress −ストレスは語彙認識に重要な役割を果たしているか」『中国地区英語教育学会研究紀要』, No. 17, pp. 17−24.
波多野五三(1987)「Unitary Competence Hypothesis 再考−日本人英語学習者の聴解力と読解における未知語推測力の関係から−」『英語教育研究』, No. 30, pp. 6−26.