NOTES ON THE MEASUREMENT OF ACHIEVEMENT IN FOREIGN LANGUAGES

John B. Carroll

(August, 1954)

NOTES ON THE MEASUREMENT OF ACHIEVEMENT IN FOREIGN LANGUAGES

John B. Carroll

(August, 1954)

I. Introduction

This introduction will sketch the history of foreign language achievement measurement, survey the various approaches which have been made, and outline recommended procedures for constructing achievement tests suitable in different situations.

II. History of achievement testing in foreign languages

Almost since time immemorial, examinations in foreign languages have taken the form exercises in translation and composition. The examinations set by the College Entrance Examination Board, for example in 1928 (College Entrance Examination Board,1928) in various ancient and modern languages, will provide good examples of the type of examination which had been in vogue up to that time. These examinations required straight translation of connected discourse and easy answers to questions about grammar. Nevertheless, with the development of objective psychological testing numerous instances of attempts to make more objective and reliable foreign language examinations are reported even as early as 1920 (Handschin, 1920). Books on the construction of standardized objective examinations began to have things to say about the construction of such tests in foreign languages (Munroe, DeVoss, and Kelley, 1917; Pressey and Pressey, 1923; Ruch and Stoddard, 1927; Symonds, 1927).* There were even attempts to construct semi-objective tests of oral and aural work. A "Committee on Resolutions and Investigations" appointed by the Association of Modern Language Teachers suggested in 1917 a revised plan for an oral and aural test for admission to college in

---

*More recent textbooks in educational which have included extensive sections on testing in foreign languages are the following: Hawkes, Lindquist and Mann, 1936, pp. 264-336; Odell, 1940, Chap. IV; Jordan, 1953, pp. 207-244; Greene, Jorgensen, and Geberich, 1954, pp. 465-482. Buros has included numerous reviews of foreign language tests in his series of yearbooks: Buros 1941, Items 1340-1375; Buros, 1949, Items 176-213; Buros 1953, Items 232-266.;

French, German, and Spanish. (Committee on … 1917). These are only a few of the early developments listed by Buchanan and McPhee on pages 374 to 394 of their Annotated Bibliography of Modern Language Methodology (Buchanan and McPhee, 1928. They represented a revolt from the subjectivity, unreliability, lack of comprehensiveness, and general cumbersomeness of the old-style examinations.

One of the first large-scale experiments with objective foreign language tests was conducted by Wood (1927) for the Board of Regents of the State of New York. Wood and his collaborators constructed a number of paper-and pencil tests measuring vocabulary, grammar, and reading comprehension in French and Spanish. These tests were administered to thousands of high-school students in New York State; the data were analyzed and reported with a thoroughness and detail for which current publication costs would be nearly prohibitive. Of particular interest and usefulness are the data on the individual questions of the test; the difficulty and validity of each question is reported in extensive tables of Chapter IV. The tests developed in this investigation are still available as the Columbia Research Bureau Tests in Modern Languages, published by the World Book Company. Thee are still highly regarded by modern language teachers, with the limitation, of course, that they measure only skills in the written language.

Another major effort was represented by a series of studies made under the leadership V.A,C. Henmon (1929) for the Modern Language Study. Henmon´s report consists mainly of extremely detailed analyses of a certain group of tests (the "Alpha" tests in French, German, and Spanish) designed for U.S. and Canadian high-schools and colleges and which were published by the World Book Company (these tests are still available). Like the tests developed by Wood, the Alpha tests are also tests of vocabulary, reading, and grammar, but the Henmon report also presents useful data on the individual items of the tests. The volume also reports developments in other kinds of achievement testing. The work on quality scales for written composition, reported in Chapter III, is notable and unquestionably still useful. Experiments in auditory comprehension tests in French and Spanish are reported in Chapter IX.

Other tests of vocabulary, reading comprehension, and grammar in the commonly taught modern languages, patterned after those made by Wood, Henmon, and their collaborators, were widely used, imitated, and even improved upon during the 30´s and early 40´s. They continue to be popular even up to the present time. For example, the Cooperative Test Service, first as an independent agency of the American Council of Education, and latterly as a division of the Educational Testing Service, has had a long history of developing such achievement tests, for high school and college levels. Similar tests have been developed by the Bureau of Educational Research and Service, State University of Iowa, and by the United States Armed Forces Institute. The development of various types of frequency counts (of vocabulary, idioms, and syntax) have made it possible to control the sampling of test content more rigorously than might otherwise be the case. However, a frequent criticism of these frequency counts and their use in the construction of tests is that they have been based almost exclusively upon printed materials; thus, it is often claimed that these tests constitute a handicap to students who have been trained in courses emphasizing oral-aural skills. Shaeffer (1948) for example, blames this feature of the Cooperative Tests for the relatively low standing of students taught by oral-aural methods in the Agard-Dunkel Investigation of the Teaching of a Second Language (Agard and Dunkel, 1948). But he points out that they do poorly on vocabulary but relatively better on grammar because their oral-aural training, he claims, is adequate to give them structure points of the language.

By 1942 even the College Entrance Examination Board had changed over to the new style of testing in its entrance examination program (Fuses, 1950, p. 156). In 1954 the College Board published a useful little pamphlet describing its tests in French, German. German, Latin, and Spanish; the pamphlet (CEEB, 1954) contain a variety of sample item types, all for reading, vocabulary, grammar, and syntax. It is worth noting, incidentally, that the College Board has objectified, to a considerable extent, even its test in English

composition. As we shall see, the so-called "inter-linear exercises" utilized in the CEEB English composition test might possible be adapted for use in connection with foreign language examinations.

We may conclude that the techniques of constructing paper-and-pencil foreign language tests of vocabulary, reading, and grammar are highly perfected at the present time.* It is true, of course, that all the customary problems of item writing apply with equal force to the writing of items in foreign language examinations; for example, in multiple-choice items, the distractors should be equally attractive, and should represent mutually exclusive ideas. In addition to the general problems of item writing, there are certain problems which are special and unique to tests of foreign languages. One particularly vexing problem is that of cognates. Cognates in various languages related to English, as well as borrowed terms in almost any language, will often "give away" the answer more easily than the test constructor may be aware. For example, the spoken sentence "Est schneiet im Winter" in a true-false test is almost certain to be answered correctly even by a person who knows no German. On the other hand, so-called "false cognates" (words in the foreign language which have similar form to a word in English, but a different meaning) may be used to form good distractors, if the French "se dérober" does not mean what it seems to mean; it actually means "to steal away, escape".

Suppose, further, that one is constructing an achievement test in Rumanian, a Romance language which has very strong overtones of Latin. It is difficult to make this test sufficiently free of cognates with other Romance languages to prevent high scoring on the part of persons who know no Rumanian (as such) but who know something about other Romance languages.

In the meantime, progress in the development of tests of aural comprehension and oral production has been considerably slow. It has

---

*A useful and interesting discussion of problems of item writing is Paula Thibault´s article which appears in he monograph edited by Hill (1953)

sometimes been claimed that the lack of progress in these phases of testing has been due to the inherent difficulty in them. It seems more likely that this lack of progress stems from the following considerations: a) Standardisation of auditory tests would require the use of recorded stimuli: only in the last few years has recording equipment of satisfactory flexibility and fidelity become available in the form of the tape recorder, but in any case it may take some time before such equipment is used widely. b) In view of the fact that auditory tests require special equipment, they have not been considered feasible in large-scale testing programs; consequently, test construction agencies have not been willing to invest research effort in this form of testing; c) Oral-aural testing has become of interest to foreign language teachers chiefly since the advent of World War II, when courses stressing oral-aural skills began to occur more widely. In short, the lack of progress in oral-aural testing is simply due to the lack of effort. There is no reason why good tests of oral and aural skills can not be made more readily and in the near future.

We have already mentioned several auditory comprehension tests in French and Spanish, developed under the sponsorship of Henmon´s Committee (Henmon, 1929). Cole and Tharp (1937, pages 345-44) lists a number of other aural comprehension tests in French, Spanish and German; none of these tests have phonographic recordings available, and few of them seem to be commercially published. Nevertheless, some of them incorporate features which appear to be as useful now as they ever were. For example, the Rogers-Clark American Council French Aural Comprehension Test and the Lundeberg-Tharp Audition Tests in French, Spanish and German deserve examination. Some of their features will be described below. The chief drawback of these tests is that they did not have wide use and consequently did not get the benefit of adequate research. One reason for their lack of use seems to have been the apathy of many modern language teachers about tests in general, particularly tests of aural comprehension.

It seems not to be widely known that the College Entrance Examination Board established, in 1930, an English competence examination for foreign language nationals who hoped to come to the United States for study. This examination included measurements of aural comprehension and oral ability, and was designed for administration in foreign countries. The examination was discontinued after 1933 because the number of candidates to be tested annually (averaging about 30 per year) did not justify its continuation (Fiske, 1934).

World War II brought an increased emphasis on the spoken language and a corresponding interest in oral-aural testing. Doubtless there were numerous institutions teaching foreign languages which also sponsored the development of appropriate tests of achievement, but few of these efforts were reported in the literature. The text-books in spoken foreign languages produced by the Intensive Language Program of the ACLS (now published as the Holt Spoken Language Series) incorporated a series of small-scale, non-standardized testing devices. At Harvard, P.J. Rulon constructed a series of highly interesting tests in German and Russian under a contract with the War Department; these tests exist on professionally-produced phonograph records, but they were never used, owing to the fact that the ASTP program for which they were designed was closed down before they were fully completed.

The work of Sandri and Kaulfers (1945, 1946) with auditory comprehension and oral production tests in Spanish deserves special mention, as well as Kaulfers's (1944) oral fluency scale in Spanish. These tests seem particularly well designed; norms and statistical data are not as complete as might be desired, but this is simply because the tests have apparently not been widely used. They are not commercially available, but since they have been presented in Sandri and Kaulfers's articles, they presumably could be recorded by any teacher who might wish to use them. Furthermore, they provide models which could readily be adapted for use in other languages. In many respects, these tests seem to be better than several tests constructed at a later date.

We will conclude this brief history by citing a number of still more recent developments in the oral-aural phases of testing. The auditory comprehension tests constructed by Agard and Dunkel(1948) for their Investigation of the Study of a Second Language are fairly widely known, owing to their wide-spread use in connection with that investigation. It does not seem to be widely known, however, that they are still available from the Veterans Testing Service, 5741 Drexel Ave., Chicago 37, Illinois. These tests, in French, German, Spanish and Russian, will be described and commented upon below. Agard and Dunkel also started work on tests of oral production, but practically no usable materials remain from their efforts in this direction

About the same time, in 1948-1949, the War Department developed a series of so-called proficiency examinations in some 20 or 25 modern languages. These were designed not so much as end-of-course examinations but rather as aids to locating Army personnel with foreign language qualifications. (The present writer happens to be connected with the development of these tests.) It was extremely difficult to get some normative or validation data for these examinations, but from all reports they have served their purpose adequately, despite their somewhat hasty construction. Each of the tests consisted of three parts, of which the first two parts were, respectively, true-false statements recorded on a phonograph record, and questions with multiple-choice options in English, the questions being recorded phonographically. The writer does not recall the nature of the third part of the examination. The Army has recently become interested in tests of oral production (Kaplan and Berkhouse, 1954).

In connection with an extensive study of foreign language aptitude for the Army, performed by Dorcus at al. (1952), a series of special proficiency tests were constructed by the Army Language School in Russian, Japanese, Hungarian, Serbo-Croatian, Arabic, and Mandarin Chinese. Thee tests appear to incorporate a rather wide variety of testing devices. It is not known whether the examinations are available outside the Army.

At the University of Michigan, in the English Language Institute, Robert Lado (1950, 1951, 1953)has developed a series of examinations in English as a foreign language. These examinations appear to have high reliability and validity, and they have particularly well solved the problem of distinguishing between knowledge of lexicon and knowledge of language structure.

Bovée (1947, 1948) has constructed several ingenious tests in reading and inn aural comprehension. While these tests are not in any sense standardized, they may provide some useful ideas for the construction of future test.

Finally, Nelson Brooks, as chairman of a committee on tests (of the Northeast Conference on the Teaching of Foreign Languages; see Kellenberger, 1954) has sparked the development of an aural compre hension test in French which has recently been accepted by the CEEB as part of its placement series. Parallel examinations in German and Spanish are promised in the near future. These tests incorporate no particularly novel features, being quite similar, for example, to the Lundegerg-Tharp Audition tests. They have been subjected to the statistical analyses, and it is probable that the wide-spread enthusiasm about them will assure them a more permanent place than some of the tests which were proposed for a similar purpose as long ago as 1919 (se Doyle, 1927).

III. Dimensions of Foreign Language Achievement

It is assumed that we are concerned solely with the acquisition of a foreign language, not with the acquisition of the culture of foreign people, nor the appreciation of its literature. In another memorandum (Carroll, 1954, the writer has pointed out that the type and level of mastery achieved in a foreign language must be considered in at least three dimensions:

1. Mastery in terms of auditory comprehension, oral production, reading, and writing.
2. Mastery of the linguistic structures (phonology, grammar, syntax) vs. mastery of the *lexical* aspect of the language.

3.The actual level of ability achieved for any aspect, i.e., this is the third dimension, or "independent variable" which has to be specified for each cell of the following chart:

Type of Behavior	LANGUAGE ASPECT	
	Linguistic Structure	Lexicon
Auditory Comprehension	(1)	(5)
Oral Production	(2)	(6)
Reading	(3)	(7)
Writing	(4)	(8)

In theory, it should be possible to obtain measures of the level of mastery of an individual in each cell of the above table. In practice, it is probably easier to distinguish between the types of behavior represented by the vertical dimension of the table than the various kinds of language mastery represented by the horizontal dimension. This is because the vertical dimension represents differences in the active (productive) and passive (receptive) behavior with reference to the two aspects of language - spoken language and written language, while the horizontal dimension actually represents a single highly complex continuum along which all the facts about a language can be arrayed, including its phonology, its morphology, its syntax, and all the ramifications of its lexicon. Furthermore, even though it is possible to maintain a fairly sharp distinction between linguistic structure and lexicon, in the practical situation of testing it is obvious that one must use lexical items in testing linguistic structure. If one takes the point of view that "knowing a language" is chiefly knowing its structure, rather than its vocabulary, testing the linguistic structure becomes more important. Yet, this can only be done by using lexical items. The difficulty can be resolved by agreeing in advance that the lexicon to be utilized in testing linguistic structure will be kept as restricted as possible, or restricted to an agreed vocabulary list. Preferably, it should be restricted to the vocabulary items learned in the particular course in which achievement is to be tested. If one is interested in measuring proficiency out of the

context of a particular language course, the lexicon must be kept to items of high frequency in the language, as determined by appropriate frequency counts (when available) or by expert judgment.

It is very probable that most tests of language proficiency, regardless of which aspect they measure, are difficult for the examinee in proportion as the lexicon is difficult. For example, both auditory comprehension and reading comprehension tests can readily be made difficult by including relatively infrequent or unfamiliar vocabulary items. Likewise, oral production and writing tests can be made difficult by requiring the subjects to produce language about foreign language. It is probably for this reason, chiefly, that different kinds of foreign language achievement tests are usually found to be highly correlated. Furthermore, it is a general rule that where a number of different abilities are taught in a foreign language course, students progress in those abilities more or less equally, with the result that correlations among different types of achievement tests are high. On the other hand, one would expect relatively lower correlations between two abilities, one of which is taught in a course and the other of which is hardly emphasized at all. These considerations must be taken account of in interpreting certain kinds of data which have been reported in the literature. Fichen (1937), for example, found correlations averaging about .8 between tests of vocabulary and reading, as well as between grammar and vocabulary was only .7. These findings can be simply explained by pointing to the large lexical component in all these tests. Likewise, Bovée´s (1948) finding of a correlation of .792 between "audio" and "visual" thought comprehension in French is probably to be ascribed largely to the common lexical element and the fact that the class was taught by an aural-oral method. A similar interpretation may be made of the findings of Kamman (1953), who found that in a group of American-speaking students of Spanish, tests of various abilities written in Spanish tended to have a stronger general factor than tests of the same abilities written in English. By controlling the type of training and the types of text per-

formances, one could probably produce any factorial structure that one might desire in a battery of foreign language achievement tests. This, of course, is only a hypothesis, but it seems to be consistent with the results obtained to date. (Compare also Wittenborn and Larsen's finding (1944) of a singly factor of German ability in the subtests of the Cooperative German Test and grades in college German.)

On the subject of factorial composition of language achievement tests, it is probably worth while to insert a word of warning. A number of different factors of "verbal ability" have been demonstrated for native speakers of English; these include "verbal knowledge", "word fluency", "ideational fluency", "fluency of expression", and the "naming factor". (See Carroll, 1941; French, 1951). Of these, the only factor truly represents knowledge of English is the verbal knowledge factor; the others represent specific kinds of behavior which probably reflect variations in cognitive processes or in personality rather than in mastery of the language. In measuring achievement in a foreign language, one wishes to measure the analogue of the verbal knowledge factor. The measurements should not reflect variations in word fluency, ideational fluency, etc. For example, an oral production test in a foreign language is probably influenced by the factor we call "ideational fluency" if we require the examinee to "think up" a series of ideas; it is better for us to contrive to put the ideas in the subject's head, asking him only to express them in the foreign language.

The fact that under many conditions different kinds of language achievement are highly correlated will often make it possible to rely heavily on the more easily constructed and reliable tests, with less stress on tests of such abilities as oral production, which seem more difficult to construct or to administer. For example, Evans (1937) found a correlation of .80 between scores on the phonetic accuracy subtest of the Lundberg-Tharp test (a group paper-and-pencil test) and ratings of recorded samples of pronunciation. Likewise, Lado (1953-54) has cited the fact that his paper-and-pencil

pronunciation test, included in his English Language Test for Foreign Students, correlates highly (r= .89) "with three combined tests, two of which are auditory ones, and it can therefore be used alone whenever it is not practicable to use in addition a test of aural comprehension". This would imply when oral instruction has been given, a group test in which the examinee merely selects among differently pronounced words is almost as good as an individual test in which he pronounces words himself.

Nevertheless, in developing a series of achievement tests for a given situation, it is probably wise to construct tests for the various kinds of mastery in which one is interested; as experience accumulated with regard to the correlations between different kinds of tests, the tests which are less predictable or less reliable can be dropped if it is seen that the dimensions they measure are adequately measured also by the more reliable and feasible tests.

I should like to insist again, however, on the necessity for careful control of vocabulary. First I should like to offer the hypothesis (without guaranteeing in any way that it might be con--firmed)that can individual with a very limited vocabulary and structure might be able to do extremely well in an aural comprehension test which would be limited to that vocabulary and structure. For example, it is conceivable that an individual might be able to react very quickly and efficiently to a series of simple directions phrased in a simple terminology, even when the rate of speech might be quite fast. If this hypothesis could be confirmed, it would show that beyond a certain point of acquisition of a foreign language is almost solely a matter of obtaining a larger vocabulary and all that implies. It would also imply that in any battery of foreign language achievement tests, the measurement of vocabulary should be kept as fa as possible independent of other aspects of language achievement. The assumption here is that if an individual knows the general phonological and orthographic system of a language acquisition of a vocabulary item by one mode (e.g. visual) will immediately transfer to another mode (e.g. auditory),

separate testing of these response systems unnecessary.* (There is, of course, the special case of languages which are written ideographically; Chinese and Japanese are the only well-known examples. Here it is necessary to test separately for auditory and visual vocabulary.

We will now proceed to a survey of the various kinds of language achievement test performances, classified in terms of the type of behavior sampled, and the kind of linguistic knowledge measured.

IV. Types of Foreign Language Achievement Tests.

As has been indicated before, foreign language tests may be usefully categorized into tests of reading, of writing, of aural comprehension, and of oral production. This classification recognizes a division into tests of competence with the spoken language and tests of competence with the written language, and a further division into tests measuring passive control and tests measuring active control. In practice, it is no always easy to maintain these distinctions, nevertheless, this survey will attempt to follow this classification.

Tests of Reading Skills

The common element in the tests to be considered here is the fact they involve responses to foreign language materials in written or printed form and that they can be administered as group paper-and-pencil tests. They involve what may be called "passive" control of foreign language lexicon and structure, in the sense that decoding of the foreign language is emphasized, the only encoding being into the native language.** The foreign language stimuli involved range from single words and phrases to long passages of

---

* It is true that, as Anderson and Fairbanks (1937) have shown for native speakers of English, there are discrepancies between English "reading" and "hearing" vocabularies, depending upon ability in reading. These are so small, relatively. as not to overthrow the assumption made here

** For convenience, we shall usually speak of English as the native language, since this paper is oriented chiefly around problems of measuring foreign language skills of native speakers of English. The reader may mutare mutandis in case he is concerned with tests to be applied to native speakers of other languages besides English.

connected discourse. The tests in which single words are the stimuli are, in the main, vocabulary tests, while tests with more ex-tended stimuli are usually called reading comprehension tests, but even these often turn chiefly on knowledge of vocabulary rather than of grammar and syntax.

a. Tests of vocabulary -Vocabulary tests exist in a variety of forms. They may be set up either as recognition tests or as recall tests. It has been established a number of times that these two types are very highly correlated (Henmon, 1929, 15, 346). It may be harder to make direct inferences about the size of an individual's vocabulary from a recognition test than from a recall test, but recognition tests order the examinees vary reliably. (It should be noted that in speaking of recall tests at this point, we are thinking only of those tests in which the stimulus is a foreign language word, and the subject must recall its English meaning. Tests of recall where the subject supplies the foreign language word are treated under the category of *written tests*.)Recognition tests of vocabulary are found in the following varieties: the foreign language words may be matched either with English words in the foreign language, or with pictures or other non-linguistic representations. There seems to be six possible types, as follows:

1.A printed foreign language word is to be matched with one of a number of words in English.

Example (German) fast

1. probably
2. extremely (Item # 1, CEEB, 1954)
3. usually
4. nearly
5. often

2. A printed English word is to be matched with one of a number of options in the foreign language:

Example: (German) explain

1.ausgeben
2. empfinden (Item# 9, CEEB, 1954)
3. entlassen
4. erklären
5. erschöpfen

3.The item is composed solely of foreign language words, among which the subject must identify synonyms, or eliminate words not belonging with the remaining words.

Example (a) (German) (Item # 13, CEEB, 1954)

Pick the pair of synonyms:

(1) gedämpft (2) nass (3) feucht (4) angstvoll

Example (b) (French) (from Bonnardel, 1951)

Eliminate the word which doesn't belong with the rest:

1.gai 2. rieur 3. sourient 4. trist 5. amusant 6.joyeux

4. A foreign language word or phrase is to be matched with one of a number of options, also in the foreign language. This type of item is favoured by some because it need not involve any translation into English, at least in theory. Actually, there is no guarantee that the student will not resort to a quick mental translation into English; recent evidence seems to show that some highly practiced bilinguals continually practice translation from one language to another.

Examples: (French)

chagrin

1. expression moqueuse
2. minéral combustible
3. animal à deux cornes
4. vieux morceau d'étoffe
5. état désprit douloureux

(Item # 18, CEEB, 1954)

agir sans réflexion, c'est être

1. aveugle
2. acharne
3. épris
4. étourdi
5. brouillard

(Item # 21, CEEB, 1954)

Example: (German) (Item #36, CEEB, 1954)

Sie haben Ihre Uhr verloren:

1.Wie viele Stunden dauert das?
2. Ich hätte vorsichtiger sein sollen.
3. Ich muss sie zum Uhrmacher bringen.
4. Wie befinden Sie sich?

(Item # 36, CEEB, 1954)

The last item might be classified as a reading comprehension item, but it is just as likely to turn on knowledge of vocabulary.

5.A foreign language word is to be matched with a number of options in the form of pictures, with which the foreign language word is associated.

Example: (German) Teppich"

1. (picture of table)
2. (picture of rug)
3. (picture of stairway)
4. (picture of a painting)

6. One might also present a picture with four possible foreign language words as options. No example will be given here; this type of item is the opposite of type 5. It would, at least, require few pictures to be drawn than type 5.

These six types present endless possibilities for item construction; each type may be made easy or difficult depending upon the commonness or rarity of the key words and upon the extent to which the options require fine discriminations. Type 1 is undoubtedly the easiest to construct, and it is the commonest; Type 2 tends to be slightly more difficult for the examinee. Both types 1 and 2 presume that English is the native language of the examinee (or at least that the examinee is highly fluent in English); thus, these types are not appropriate where one has individuals of differing language backgrounds. For example, a native speaker of French might fail miserably on a French vocabulary test composed of items of type 1. Types 3 and 4 do not suffer from this disadvantage, nor do the items of types 5 and 6, if the pictorial material is sufficiently neutral with respect to the cultural content. Type 3 is likely to measure more than simply knowledge of a foreign language; items composed in this way may actually turn out to be intelligence tests, in the sense that they might differentiate individuals who have equal knowledge of the foreign language vocabulary involved. Type 4 is probably a good format, but since it involves so much foreign language material, it would not be useful in a diagnostic sense; that is, for example, one might not know whether an individual fails an item because he did not know the stimulus word or because he misunderstood some word in one of the options. Thus, such items might tend to be more unreliable.

Vocabulary items involving the use of pictures are more cumbersome to construct, to be sure, but they are free of certain disadvantages shared by the other types. For example, they are independent

of the examinee's control of the English language, and they minimize dependence on his control of the foreign language orthography.

Completion tests of foreign language vocabulary practically always require the supplying of an English translation (in written form) for a given foreign language word. Sometimes these words are embodied in a sentence or larger context, in which case the subject can use the context for making inferences about the meaning of the word. This often has the effect of testing the subject's ability to infer the meanings of new words. Such a test, however, is not diagnostic of actual word knowledge, since supplying a correct translation may depend either on actual knowledge of the word as such or upon an ability to infer the meanings of unfamiliar words from context; one has no way of ascertaining which it is. If one really wants to test ability to infer meanings of new words from context, of the work of Werner and Kaplan (1950) where one presents the subjects with items such as the following:

What is a corplum? Notice its use in he following sentences:

1. A corplum may be used for support.
2. Corplums may be used to close off an open place.
3. A corplum may be long or short, thick or thin, strong or weak. (etc.)

It happens that Gibbons (1940) has already developed such a test; he finds that the ability to consruct the meaning of a strange word from context is very specific.

If there is any real reason to construct completion test of vocabulary, it would be probably be a more valid test if the foreign language word were not presented in context(except to the extent necessary to specify the particular meaning intended). But even such a test has little to offer beyond what can be measured by an ordinary multiple-choice vocabulary test in one of the varieties described above. Stalnaker and Kurath (1935) found these types correlated highly and were approximately equal in reliability.

b. Tests of reading comprehension. Items designed to measure reading comprehension are tests in which the foreign language stimulus is longer than a phrase. The assumption is that comprehension of the foreign language material depends on knowledge of the structure of the language. Actually, it most often depends upon knowledge of lexical items. For example, in the following German sentence, which the student is supposed to indicate as True or False, the student who knows all the structural characteristics of German, such as the morpheme for the superlative of adjectives, might still mark the statement incorrectly if he does not know certain words. such as Kerzen:

Das kleinste Streichholz brennt als die grösste Kerze.

Therefore, in order for reading comprehension items to be truly diagnostic the structure points to the language, the vocabulary must be carefully controlled. These items exist in a number of varieties. Perhaps the commonest is the one where the statement in the foreign language is to be indicated as true or false. An example of such an item was given above. One must be careful, in constructing such items, to limit them to statements whose truth or falsity will be well within the experience of the persons who are likely to be tested. In other variants, there may be statements in the foreign language (one or more sentences) after which occur multiple-choice, true-false, or completion questions on the statement. The lead and the options may be either in English or in the foreign language, Finally, there may is the Van Wagenen technique (Henmon, 1929, p.301) in which the subject is supposed to check whether statements (either in English or the foreign language, but usually the former) contain ideas expressed in the paragraph or derivable from them or not. Ruch and Vander Beke (see Henmon, 1929, p. 30) performed an experiment on the relative reliability and validity of these various types of reading comprehension items. The conclusion was that the type where the items are in the form of T-F statements, based on the original paragraph, was the most reliable and valid type, when allowance was made for time required, administrative feasibility, etc. Since the Van Wagenen technique was not included in these comparisons but partake of the same characteristics as the T-F item.

it may have something to recommend it (except for the danger that it may involve too much "reasoning", independent of actual language knowledge).

One great difficulty which is encountered with all tests of reading comprehension (whether a foreign language is involved or not) is the possibility that the questions can be answered extremely well even when the subject answers by "common sense" or by noticing certain internal characteristics of the questions. Reading comprehension questions should always be tested on individuals who have not read the paragraphs on which they are based.

Finally, we may classify translation tests of a recall type as measures of reading skill. About the only way in which this type of test can be made objective is to set up a sentence in a foreign language, the subject being required to choose which of a series of English sentences is a correct (or the best)translation of the foreign language sentence; an even better refinement is to construct a paragraph with a number of words and phrases supplied with "translations", the subject being required to state in each case whether the translation is accurate or not.

Several other types of reading tests present themselves as interesting possibilities. One would be an adaptation of the technique used in the Minnesota Speed of Reading test (Eurich, 1936, which consists of a series of paragraphs which contain, at random intervals, words which do not fit with the sense of the paragraph. The subject is supposed to read the material as rapidly as possible, indicating his progress by underlining the nonsensical words. Although to the best of my knowledge this technique has not been used in measuring foreign language achievement, it ought to be useful, particularly for discriminating among the more advanced students. The technique would also lend itself to scaling with reference to the performance of native speakers of the foreign language involved.

Tests of Writing Skills

What might be sent by "writing skill"? In the context of this paper, we certainly do not mean the skill displayed by a Shakespeare or a Faulkner, or even a writer on the New Yorker staff. What the foreign language teacher usually means hereby is simply the ability to "put one's thoughts on paper" in the foreign language. The ability to write a reasonably intelligent letter, without betraying the limitations imposed by imperfect knowledge of the foreign tongue, exemplifies one of the aims of the foreign language instruction. A direct attack on the measurement of this ability, then, might be simply to ask the student to write such a letter, or similar composition. But of course, such a direct approach has its disadvantages, not the least of which is the consequent subjectivity of scoring and the great labor in securing reliable ratings. Besides this, the task imposed on the student has a minimum of restriction; if he is smart, he will say whatever his knowledge of the language permits him to say, and one will never be the wiser if he can't say some things he might otherwise say. The testing situation is entirely too uncontrolled to obtain any reliable pointer-reading - it is gross, unstructured, and the results will be unconvincing or misleading. Many current tests and examinations of foreign language "composition" ability suffer from these effects.

What, in essence, are the behaviors and knowledge required for intelligent setting down of thoughts in a foreign language? First, there must be "thoughts". But this will be true whether the student is writing in his own language or in another language. We are not interested in testing for the presence of "thoughts". (Even if we ask students to write themes in English, their native language, one finds tremendous variations in performance.) Some students do not seem to have much to say, at least not while they are in process of being examined - and perhaps this deficiency can be excused; we shall not go into the possible psychological explanations for this. Let us, therefore, put thoughts into their heads. We shall have to do this by using the student's native language (if one

objects to the use of the native language, one will have to be prepared to construct pictorial materials so admirably contrived that they will constitute stimuli which can be expected to lead unequivocally to certain verbalizations, and this is a difficult if not impossible feat.)

Secondly, there must be facility in active recall of the various structural and lexical features of the language to be used. We need, therefore, to construct our tests out of English phrases or sentences which in translation will demand the knowledge of specified structure points and lexical items. Normally, these will be the structure points and lexical items which have been taught in the language course in which we are testing achievement. When we are testing students of unknown or heterogeneous foreign language experience, we shall have to resort to sampling from the more common elements of the language to which they may be expected to have been exposed, unless, of course, we are interested in discriminating among the upper levels of ability.

We may now consider methods of testing knowledge of language structure. Many of these methods were considered earlier in connection with the measurement of vocabulary and reading comprehension. What remains to be considered here are those types of tests which emphasize encoding into the foreign language, either by active recall of the proper foreign language forms and expressions (in written form) or by a kind of active recognition of such forms and expressions. The evidence leads us to expect that recall and recognition will be highly correlated here as in other cases.

There are a number of varieties of objective or semi-objective test items measuring active knowledge of foreign language grammar:

We give examples:

1a. (Suppy missing element)

Give me the pen knife. (_____) le canif. (Wood, 1927, p. 62)

1b. Speak to the men. Parlez (_____) hommes. (Henmon, 1929, p. 304)

2a. She opened all the windows.
Elle a ..... toutes ls fenêtres.
(1) ouvrit
(2) ouvri#
(3) ouvert (CEEB, 1954, #13)
(4) ouverte
(5) ouverts

2b. He arrived without anybody's knowing it. (CEEB, 1954, # 31)
(1) Er kam an, ohne dass es jemand wusste.
(2) Er kam an, ohne dass es jeder wusste.
(3) Er kam an, ohne jemand es zu wissen.
(4) Er kam an, ohne von jemandem erkannt zu werden.

3a. (Complete missing elements.)

___?___ am a student.

3b. ___?___ do you live? I live on State Street.

(Lado, Examination in Structure, Form C, p. 5)

4. (Choose correct alternative)

Is (it, there) ten o'clock yet?

Items 1a, 1b, 2a, and 2b supply Finnish versions, as if to suggest the "thoughts" which are to be rendered, while the context alone is strong enough, in items 3a, 3b, and 4 to suggest the proper completion. Items 1a, 1b, 3a, and 3b require active recall, while the remainder do not. There are many who dislike the multiple-choice grammar items, since it is difficult to make "likely" alternatives, and much can be said against presenting ungrammatical forms and constructions at any stage of language learning.

Any of these item types may be chosen, depending upon the particular structure point or lexical item to be tested. It will be noted that it is difficult, as always, to separate grammar from lexicon in test items, particularly if the items which require recall rather than mere recognition, For example, in item 1a. the individual might not recall the verb donner, despite its commonness. The item might be changed to

Give (donner) me the pen-knife. ____?_____ le canif.

Furthermore, in many of these items one is testing foreign language decoding as well as encoding; nevertheless, it may generally be assumed that decoding is more facile than encoding, and hence generally at a lower level on a scale of difficulty.

The format of any of these items can almost e carried into tests in which connected discourse is involved, thus enabling one to test for almost any level of mastery of the vocabulary, grammar, and idioms of a foreign language, Andrus (1942), for example, reports

much success with a completion test in which parallel passages of English and French were given; certain words and phrases were deleted from the French text and corresponding parts of the English passage underlined and numbered to agree with the numbers replacing the omitted word or phrase in the French text. The so-called "interlinear exercise" developed by the CEEB (CEEB, June 1954) to measure English composition might be used at the upper levels of foreign language ability to measure performance in writing in the foreign language. The interlinear exercise presents a connected paragraph which contains at certain points a number of grammatical blunders, improperly chosen, and infelicitous or awkward expressions. The text is printed with wide spaces between the lines, and the student is instructed to "treat it as though it were a first draft of a composition of your own, and revise it so that it conforms with standard formal English". Experienced graders can achieve a high degree of reliability in marking the papers, using guide sheets showing the trouble spots in the text and examples of acceptable and inacceptable corrections.

For the direct, diagnostic testing of active knowledge of foreign language items, it is necessary to use recall items in which the stimulus for the eliciting of a foreign language words is either (a) a defining phrase or a synonym, (b)an English word of phrase, or (c) a picture. For example, the word Bleistift in German might be elicited either by the German definition Ein Ding, das aus Holz gemacht ist, mit dem man schreibt, or by the English word pencil or by a picture of a pencil.

Several other types of items are useful in testing knowledge of foreign language grammar. There is a whole series of possible types, exemplified in many textbooks, in which the student is given some definite task in manipulating linguistic forms. Practically all of these can be made highly objective because of the inherently all-or-none character of linguistic structure. For example, students can be asked to change tense, person, number of verbs; to convert statements into interrogative forms, etc. Nevertheless, one should be careful

to avoid favouring students trained in the terminology of formal grammar and upsetting the chances of students who may have learned good grammatical habits without learning formal terminology. Hence, special grammatical terminology should be avoided as far as possible; one is interested in testing ability to encode in the foreign language rather than the ability to talk about its grammar. (For example, avoid the type of item recommended by Coleman in the book edited by Hawkes, Lindquist, and Mann, 1936, p. 321, in which the test maker formulates several grammatical rules and then asks the student which rule is illustrated by each of the series of sentences. This sort of grammatical exercise, which has little to do with knowledge of language, is justifiably avoided in contemporary modern language teaching, and hence in contemporary achievement examining.)

For purposes of illustration, here are a few samples of special grammatical tasks taken from Lado's tests in English structure:

Supply the proper interrogative word(s):

____ is my pencil? On the table.

____ Shirts do you want? Two, please.

Which expresses permission?

You (Should, may, ought to) use my pen.

Convert to negative form.

She sings well.

Give the correct verb form.

She (SING) a beautiful song last night.

Rearrange;

the, on the corner, house, big. I like ___?__?__?__?__.

1 2 3 4

For reasons stated earlier, there are a number of arguments against tests of free composition. They depend as much on what the student may happen to have to say as upon his knowledge of foreign language structure and lexicon, and they present numerous difficulties with respect to administrative feasibility, scoring reliability, etc. The topic set for free composition must be such that students cannot easily be coached for them and prepare a composition in advance; yet,

these topics must lead to compositions which are within the student´s experience and vocabulary. For the sake of completeness, however, should be mentioned that quality scales in written composition exist for several languages (French, German, and Spanish); this work is extensively reported in Henmon (1929, Chapter III). These scales are to be used as standards of reference in judging any given composition. For example, here are qualities 0, 8, and 14 on German scale A:

Quality 0

Die Knob sie moschend eine Hous sie in ein baum est. Sie arbeitet ich seehe drei Knabe. Sie sind in einer Gross Baum.

Quality 8

Auf die Leiter (?) is alle drei Knaben sind um zwölf Yahrealt. Sie sind sehr klug, weil sie so schön ein Haus bauen können. Ein Knabe hat der Hammer in der Hand und ist fleissig am Nägel schlagen, Diesses Haus hat viele Fenster, und ich nehme das es ein Summer Haus ist. Vielleicht werden die Knaben darin wohnen weil das Wetter gut ist. Das wird schön sein, Die Vögel werden sie amussieren, und die Laube wird die Sonne von ihnen palten.

Quality 10

Heinrich,Karl, und Georg sind die Kinder eines Zimmermannes. Der Vater sprach oft mit seine Söhne von seiner Arbeit.Manchmal haben die Kinder dem Vater geholfen; sie konnten ihm die Nägel bringen, oder den Hammer halten.

Einst gab der Vater den Knaben einige Bretter und eine Leiter. Georg sagt er möchte ein Fogelhaus bauen. Heinrich wollte eine Scheune bauen. Nach vielem Plaudern wählte Karl eine Idee die den Brudern auch gefiel. Eine Leiter wurde gegen einen Baum getragen, und die Arbeit was begonnen. Oben, unter grosze Ächste, wurde ein kleines Spielhaus gebaut.

Jeden Tag, nach der Schule, haben die Kinder da gelesen, gesungen oder geschrieben. Auch luden sie die Mutter und der Vater ein ihn zu besuchen.

It might be commented, incidentally, that the type of material exemplified by Quality Scale 8 might make the basis for a good "interlinear exercise."

One major difficulty with quality scales, at least as applied to foreign language materials, is the multi-dimensionality of the objects to be rated. What if the spelling of a composition is impeccable, but its vocabulary and syntax put it very low? Does one srike a balance, or does one weight one factor more than another? These are difficult questions to anticipate or to resolve.

Tests of Auditory Comprehension

Here we shall consider tests emphasizing the ability to decode spoken foreign language stimuli: It has often been alleged that this type presents insuperable obstacles; in actuality, it presents no more difficulty for the test constructor than tests of reading comprehension. Many item types feasible in tests of reading vocabulary and comprehension can be converted to tests of auditory comprehension by the simple device of presenting the stimuli in spoken form, leaving the options in printed form.

One of the difficulties alleged to be associated with tests of auditory comprehension is that on the one hand, spoken language stimuli are too unstandardized if they are left to be read by the person administering the test, but on the other hand, the voice of a speaker on a phonograph or a tape recording will be too unfamiliar, either in register or in dialect, to the student who has never heard the voice of any foreign speaker other than his instructor. In my opinion, this difficulty has been exaggerated; if students have difficulty with a recorded, unfamiliar voice, it reflects upon the failure of the instruction to provide the necessary variety of models. By all means, auditory comprehension tests must be given by means of recordings of high fidelity in rooms with appropriate acoustical properties. There can be a "warm-up" period on the recording, in which the native speaker´s voice starts in English and then gives very simple materials in the foreign language - materials so simple that nearly any student will gain confidence by realizing success with them.

As in other types of foreign language achievement tests, auditory comprehension tests measure different kinds of performances. They may test comprehension of the spoken language without any references to the printed form of the foreign language, or they may test the ability to match spoken utterances with printed materials in the foreign language; these performances may be quite different. They may tests knowledge of lexicon or of grammar and usage. The amount of retention required may vary from practically nothing, as the tests were a simple phrase is to be comprehended, to a great deal, as in tests which present long discourses with questions to be answered thereon. They may require either the recognition (selection) of correct answers or the supplying (recall) of the correct answers. It is difficult to choose a basis for ordering these types in our discussion. IN all cases the auditory stimulus is an utterance (of some length, short or long) in the feeing language, but the type of response to be made by the examinee varies.

1. Writing from dictation: (a) Foreign language orthography.

The subject writes a spoken sentence in the foreign language orthography. The average memory span places a limit upon the amount that can be dictated continuously with sufficient time for the subject to execute his response. This type obviously calls into play the subject´s knowledge of foreign orthography; it could not be used, for example, in a course in Chinese where the writing system is not being taught. It is of no particular use when interest is primarily in comprehension of the spoken language. The technique involves dome difficulties in scoring, but it can be made objective, though not mechanical.

2. Writing from dictation: (b) writing in phonetic or phonemic transcription. This kind of task has not appeared, to my knowledge, in any formal achievement test, but the technique may be in use by some foreign language teachers. Scoring can be subjective, but not mechanical.

3. Writing from dictation: (c) Response to fast dictation. In general, this type exists only in theory; it would be feasible only if students were taught a foreign language shorthand. However, one variant is possible; if numbers are dictated (e.g., zwei tausend vier hundert sechs und siebzig), the examinee should

be able to keep up with the dictation. This type has been exploited, using an artificial language test, in an aptitude test developed by the writer (Carroll, 1941, Artificial Language Numbers Test).

4. Translation from spoken utterance: (a) Single words or short phrases. In effect, this is a test of auditory passive vocabulary. The word is presented, the English equivalent to be written by the examinee. Except for the fact that scoring may be present some difficulties, and cannot possibly be mechanical, this type has much to recommend it, since it can be directly scaled, and used diagnostically. Occasionally a certain amount of foreign language context may have to be provided in view of the possible multiple English meanings for foreign language forms.

5. Translation from spoken utterance: (b) Sentence or short paragraphs. This type of test can tap knowledge of grammar and supplies the answer, it does not lend itself readily to objective scoring, and it cannot be mechanically scored. If the stimuli are too long, there is too much reliance on memory span.

6. Following directions from spoken utterance. This type is suggested by certain intelligence tests in which the examiner asks the subject to perform certain tasks. Perhaps this could be done in a foreign language as a group paper-and-pencil test. To my knowledge, this type has not been tried in any formal foreign language utterance could be paced too rapidly, but the disadvantage that the vocabulary of paper-and-pencil test directions ("circle the star," "cross out the triangle," etc.) is not likely to be within usual foreign language vocabularies unless special ingenuity is shown in constructing the test.

7. Indicating the truth or falsity of foreign language statements. This has been an extremely popular variety which seems to be easily constructed and highly reliable and valid if care is exercised in construction. Many examples can be found in the work of Rulon (1944) in German and Russian, and Sandri and Kaulfers (1946) in Italian. It was used extensively in Army Language Proficiency Tests in a number of languages. This type of test has the advantage

that foreign language orthography is involved in no way; it depends solely on comprehension of the foreign language sentenced, -- it being assured that anyone who comprehends the sentence will immediately perceive its truth or falsity.

Examples: (from Rulon's German test)

Voices	Anser	(Translation)
Eine Bibilothek ist eine Verkaufstelle.	F	A library is a place where things are sold.
Viele Deutsche trinken Bier.	T	Many Germans drink beer.
Wir schlafen in ein Koffer.	T	We sleep in a trunk.
Bauern leben in Schlosser.	F	Farmers live in castles.
Es schneit im Winter.	T	It snows in winter.
Berufsbeziehungen findet man auf Familiennamen.	T	One finds occupational terms in family names.
Siebmal fünf is fünf und zwanzig.	F	7 x 5 = 25
Können Sie den Mond berühren?	F	Can you touch the moon.

By keeping the items short and limiting the crucially difficult aspect of each item to one element (the others being relatively easy), this type of test can probably be made scalable; it is possible that structural and lexical knowledge can be tested separately.

8.Answering questions (one-word answers). (Since we are here interested in auditory comprehension rather than oral production or writing skill, it will be assumed that the answers are written in English. Bovée(1948) provides numerous examples of items of this type; in what he calls his "audio recognition test of typical thought units," 40 questions are to be dictated orally; they are so formulated that "a single-word answer in either French or English would give unmistakable evidence of comprehension."

Examples: (answers to be written in English)

Quel est votre nom?

De quoi a-t-on besoin pour acheter un chapeau?

9.Multiple choice: (a) Choosing correct translation in English.

This can be done either with single word stimuli or short phrases, or with sentences. In effect, these correspond to tests of reading comprehension and reading vocabulary discussed earlier. This form of test was employed by Sandri and Kaulfers (1946) in the aural comprehension test in Italian. It seems probable that if both spoken and written forms of the language have been taught, auditory and reading vocabulary tests will correlate highly.

10. Multiple choice: (b) answers to questions, completing definitions, etc. The options are printed either in English or the foreign language; usually it is the former, in order to be sure that one is testing auditory comprehension of the spoken language stimulus rather than the reading of the options. Some prefer to have options in the foreign language in order to reduce the tendency to translation. This type has been very popular and is easy to construct. Scoring can be objective and mechanical.

Examples:

Voice	Options	
1. Ein Mann, der Fleisch verkauft	1. baker 2. butcher 3. doctor	(Agard-Dunkel Tests of Aural Comprehension, Lower level)
2. Se cultivan las flores en un..	1. garden 2. box-car 3. coal-mine	(Agard-Dunkel Tests of Aural Comprehension)
3. Où va-t-on pour prendre un train?	1. En voiture 2. Au guichet 3. À la gare 4. En chemin de fer 5. À la guerre	(From Barnard-Yale Aural Test, Sample Form, Part II, #5)

11. Multiple choice: (c) Choosing "associations" with stimulus word or sentences. Here the options do not necessarily have to be translations of the stimuli; they may have any kind of association with the stimulus, however remote, as long as it is closer than those of the wrong alternatives. The options may be either in English or the foreign language. This interesting technique, with options in the foreign language, was used by Buchanan in his "Spanish Aptitude Test " (Henmon, 1929, p. 309), from which the following examples are taken:

Voice (Spanish)	Options (printed) [Translations in brackets]			
leer [read]	casa [house]	libro [book]	esclavo [slave]	osar [dare]
agua [water]	beber [drink]	nombre [name]	luz [light]	ayer [yesterday]
después de despertarse, mi hermano miro su reloj [After waking, my brother looked at his watch.]	atreverse gastar comercial levantarse		[dare] [waste] [commercial] [rise, get up]	

12. Multiple choice: (d) Choosing pictures represented by or associated with foreign language stimuli. The foregoing two types suffer from the disadvantage that the options must be in English or in a foreign language orthography; some will object to the necessity for translating and others will object to the use of the foreign language orthography. Both objections can be met when the options are pictures. It seems clear that an infinite variety of language patterns can be tested pictorially; it is not necessary that the picture represent what is said, and thus one is not limited to the use of directly "picturable" words and concepts. All that is needed is to utilize associations which may be clearly suggested by a picture.

As has been mentioned in connection with reading and writing tests, the preparation of pictures may require special talents, and is a time-consuming process. Nevertheless, it has proved a highly successful technique. Examples may be found in a number of Lado's tests of English in a foreign language, in several of Rulon's tests in German and Russian, and elsewhere. Here are examples (pictures are described verbally because reproductive facilities are lacking):

Voice	Picture Options
"The boy likes milk chocolate."	A. [Happy boy holding glass of chocolate milk.] B. [Unhappy boy rejecting glass of chocolate milk] C. [Happy boy holding a big bar of milk chocolate.]

(This example from Lado, 1950)

Here is another example, from one of Rulon's German tests. A series of 10 items use the same four pictures; four of the items are given here.

(picture) man and wife sit at home; wife knitting; man reading newspaper A	(picture) a letter is being put into a mailbox B
(picture) airplane about to leave from an airport; persons embarking C	(picture) woman at grocery counter, talking to clerk D

(Voices) (Answers)

1. Kann man zu dem Bestimmungsort noch die genaue Nummer hinzufügt, spart man zu mindest einen halben Tag. (B)

2. Ganz mühelos, als sässe man zu Hause in einem gemütlichen Sessel, kann man heute die grössten Entfernungen in kürzester Zeit überwinden. (C)

.....

4. "Habe ich Ihnen schon erzählt, dass ich gestern einen Brief von meinem früheren Verkäufer hatte, per Luftpost? Er sehnt sich so nach Hause und nach seinen Eltern." (D)

5. "Sie sind der letzte auf meiner Liste. Wenn ich mit meinen Besorgungen hier fertig bin, kann ich nach Hause gehen und mich den ganzen Abend ausruhen." (D)

[Translations:]

1. If one adds the exact number to the place of destination, one saves at least half a day.

2. Without any effort, as if one were sitting at home in a comfortable easy chair, one can nowadays master the greatest distances in a very short time.

4. "Did I tell you that I had a letter from my former salesman, via air-mail? He has such longing for home and his parents."

5. "You are last on my list. After I get through with my purchases here, I can go home and rest all evening."

---

13. Multiple choice form for a dictation test. This type of test has been popular on several occasions during the last twenty years. It was used originally in one of the subtests of the Lundeberg-Tharp Audition Tests (Cole and Tharp, 1937, pp. 345 f.), and it has recently been utilized in the auditory tests developed by Nelson Brooks and his committee (Kellenberger, 1954). In this form of test the stimulus (preferably a recorded voice) reads one of the options in each item, the student being required to indicate which one it is. Here are examples of items in French, German, and Spanish presented by Cole and Tharp (1936, p. 345) as representative of those in the Lundeberg Tharp Audition Tests:

French	German	Spanish
_____ nous avons	___ wenig	___ chocolatera
_____ nos savons	___ pfennig	___ chocolate era
_____ nos savants	___ wenn ich	___ choca la tiera
_____ nous savons	___ wenn nicht	___ choque ladera

And here are some more recent examples. from Nelson Brook's (Barnard-Yale Conference) French test; the options are marked with an (x) are those which are read aloud by the examiner.

1. (x) Le soleil se levait lentement.
( ) Le soleil s'est levé lentement.
( ) Le soleil se leve lentement.
( ) Le soleil s'etait levé lentement.
( ) Le soleil se leva lentement.

2. ( ) A-t elle peur de cela?
( ) A-t-elle perdu celle-là?
(x) A-t-elle perdu cela?
( ) A-t-elle peur de celle-là?
( ) A-t-elle perdu ceux-là?

- This type of test probes the ability of the examinee to hear fine distinctions in phonology, as well as to recognize the orthographic representations of what he hears. Some question might be raised as to whether this ability to hear fine phonological distinctions is actually involved in the understanding of speech in normal conversational situations, where the context would be expected to prevent misunderstanding. There are two ways of finding out whether this objection is a valid one. First, can native speakers of the foreign language perform this type of item satisfactorily? If native speakers have no trouble with it, it probably represents a kind of performance which has been naturally acquired in the process of learning to use the language in everyday situations. Otherwise, we would have to infer that the fine phonological distinctions of a language are essentially useless, and that situation in which they are crucial (i.e., are the sole carriers of important differences in meaning) almost never arise. Such as conclusion is difficult to accept, but we must suspend judgment until suitable evidence is at hand. A second line of argument against the objection stated above would be the finding that items of this type discriminate well and correlate with other phases of language achievement. The high item-test coefficient reported by Brooks (Kellenberger, 1954) would seem to support the notion that this type of items is a reliable test of something which students learn in foreign courses.

If it is necessary or desirable to eliminate any foreign language orthography, pictorial materials can be substituted for the options. Use of pictures may put a limit on the kinds of distinctions which can be employed - for example, it would probably be extremely difficult to prepare a distinctive picture for each of the five options in the item about the sunrise, above (the first example from Brook's test). Nevertheless, with a little ingenuity one could doubtless develop numerous items exemplifying this technique.

14. Identification of correct usage. There is no reason why knowledge of "correct usage" cannot be tested in an auditory test as well as in the printed test. Indeed, the argument could be made that the auditory test is more realistic, spoken language being the primary form of language.* Sandri and Kaulfers (1946) include such a test in their auditory comprehension test of Italian. They present spoken sentences in pairs; one of these is correct, the other incorrect, and the student indicates which it is on his answer sheet. Such a test probably tests at a rather high level of ability in a language. Examples from the Sandri-Kaulfers test of aural comprehension in Italian:

State which is correct.

Read by a voice: 1. a. I alberi sono alti.
b. Gli alberi sono alti.
25. a. Io ho piu che venti lire.
b. Io ho piu di venti lire.
50. a. Io vorrei fargli un regalo.
b. Io piacerai fargli un regalo.

15. Auditory paragraph comprehension tests. In all the types of auditory comprehension tests discussed so far, the foreign language auditory stimuli have been relatively short - seldom beyond a single sentence. (An exception are the pictorial tests of Rulon, where some of the auditory stimuli may be 40 - 50 words long. It has been thought that while a person might be able to perform quite well in comprehending short sentences in the foreign language, he might not do as well in maintaining comprehension of continuous discourse, because he could not take the time he needs

---

* Gray (1938) found that U.S. pupils in grades II to VIII were able to detect errors in spoken language (English) more readily than they could in the printed form.

to decode the foreign language stimulus. The very fact that the beginning learner often wants a foreign speaker to "speak more slowly, please" suggests that his decoding speed in slower than it has to be if he is to keep up with normal rates of speech.* Therefore, there seems to be some validity in the notion that it is necessary to test comprehension of passages of connected discourses longer than a sentence or two. On the other hand, it is probable that other types of auditory comprehension tests, using short foreign language stimuli, will correlate highly with tests using connected discourse, making it possible to dispense with the latter. In some ways, this would be fortunate, for the paragraph comprehension tests present a number of difficulties in construction.

In general, auditory paragraph comprehension tests follow the pattern of reading comprehension tests. A paragraph or two I presented auditorily (preferably by a recorded voice), after which the examinee is tested by any of the standard item-types possible with paper-and-pencil tests: true-false statements, multiple-choice questions, completion questions, etc. In general, these items are in English (i.e. the native language of the examinees); this seems preferable to using the foreign language in these questions, because the point of testing is to measure comprehension of the spoken passage.

Examples of auditory paragraph comprehension tests are to be found in the work of Sandri and Kaulfers (1946) for Italian, Agard and Dunkel (1948) in French, German, Spanish, and Russian, Villareal (1947) for Spanish, and the Barnard-Yale Aural Test in French (Kellenberger, 1954). Agard and Dunkel, for example have two varieties of this test form. In some of their tests they present anecdotes, averaging 1 ¼ minutes in duration, often with and old-world literary flavor, with subsequent 3-option multiple choice question wholly inn English based on the anecdotes. In their "upper level" tests they present a dialogue between a man and a woman speaker; this dialogue lasts nearly 5 minutes, occupying one side of a 12" 78-RPM phonograph record. Then a series of about 15 multiple-choice questions, printed in the answer booklet, is presented.

---

*Note that it does not demonstrate that decoding speed is slower. The learner may actually be requesting a sharper definition of word boundaries, for example.

The most serious defect in paragraph comprehension tests that I have examined (and listened to) is that the questions are not properly designed. It is often true that a person can answer a series of questions quite accurately (well beyond chance) without even hearing the paragraphs on which the questions are based, much less understanding them. He is guided either by general information or by clues afforded by the questions themselves. Also, the questions in a given group are likely to be spuriously correlated, in that the options are parallel and the examinee tends to answer the questions on the basis of some consistent notion about the stimulus paragraph; if his notion happens to be correct, he gets most of the questions correct, but if his notion happens to be incorrect he gets more questions incorrect than he would be likely to by mere chance. Finally, it has been the writer´s experience that the correct comprehension of a single word in what was otherwise a welter of confusion enabled him, sometimes, to answer a considerable number of questions correctly. These defects are so serious in the Chicago test of aural comprehension that a great deal more doubt has been cast, in my mind, upon the conclusions of the Agard-Dunkel Investigation (1948). [This investigation was supposed to discover whether new-type courses employing oral-aural skills produced greater achievement than traditional courses emphasizing grammar and translation. The results were inconclusive, possibly because of deficiencies in the criterion measurement.]

In order to avoid the defects mentioned above, it is recommended that (a) the paragraph be "topical", in the sense that they will deal with particular things, particular people, or particular situations at particular times; in this way questions which can be answered from general information will be minimized; (b) the questions be constructed in such a way that parallelism between questions, and other clues, are avoided; (c) the questions be made as independent of each other as possible. In short, considerable care and item-writing skill has to be exercised in order to succeed with this type of examination.

Tests of Oral Production

Tests of reading, writing, and auditory comprehension can all be administered as group paper-and-pencil tests, and the foregoing survey of test types was restricted very largely to a consideration to group paper-and-pencil tests. Except for one or two special cases , or save for the case where one has a battery of recording machines available, tests of oral production must be administered individually. There has been very little success in objectifying oral production tests; the examinee´s production must be evaluated by trained persons, in some cases preferably by more than one judge.

Productions may be rated with respect to any one of a number of dimensions. e.g., accuracy of pronunciation, correctness of grammar, choice of words, etc.

The first problems in developing tests of oral production is to decide what kinds of foreign language responses one wants to elicit, and then to discover appropriate stimuli for eliciting such responses. Test procedures differ in the extent to which they attempt to control the response.

(a) Response controlled very little. In this type one merely asks the examinee to discourse for a short while (e.g., 2 or 3 minutes) on a topic which is assigned to him on the spot. [Presumably there is no point in announcing a topic in advance and thus allowing examinees to prepare and memorize a "speech", since in this case all one could reliably gauge would be accuracy of pronunciation. Furthermore, one would be testing the motivation of the subject in preparing his "speech."] One difficulty is in the selection of the topic; some might be so banal as to be of little interest to the subject, e.g. "What I Do Every Day," while others may tax his vocabulary, e.g. "A Visit to a Factory," or lie outside his experience. Another difficulty occasionally is the fact that one must have a variety of topics; otherwise, the topic leaks out to examinees who are still waiting to be tested.

Once the examinee has made his speech, there remains the problem of rating his production. This can be done either by establishing a number of rating scales (e.g. for vocabulary, grammar, fluency, originality, etc.)or by comparing the output

with the points on a pre-established quality scale. Some of the problems associated with quality scales have been alluded to earlier.

But there is an even more serious considerations. If we ask a group of examinees to discourse on a given theme in their native language, wide variations in performances will be noted; this is attested by the experience of speech teachers, and by a number of psychological investigations. There is a good likelihood, I think, that performance in a foreign language will reflect speaking ability in the native language.

In view of these difficulties, testing of oral production by this method is not recommended.

(b) Moderate degree of control of responses. Greater success will doubtless be attained if one attempts to elicit oral productions with a greater degree of control of content. The problem is similar to that encountered in tests of writing ability, where it was pointed out that we must put "thoughts" in the subject´s mind if we are to be sure that we are measuring control of the language rather than ideational fluency. Hence, we must consider what stimuli we might use for eliciting responses.

(1) The controlled interview. Rulon (1922) made use of a controlled interview situation. The examinee was interviewed individually; he was brought in and told (in the foreign language) that he would be interviewed in the foreign language. Questions were then asked, such as "When did you get up this morning? Why so early? What time did you have breakfast? Do you like this course?" etc. The effectiveness of the questions in eliciting fluent answers in English had been confirmed in pre-tests.

The productions were recorded for later rating. Rating was in terms of a pre-established quality scale, recorded on a series of phonograph records. Materials were provided for the judges to practice rating interviews.

One objection to a controlled interview in which the foreign language is spoken by the examiner is that measurement of aural comprehension is, as it were, confounded with the measurement of oral production. If the examinee cannot understand the questions, he cannot be expected to answer them. [It is conceivable that a person might have developed facility in oral production without a corresponding profanely in auditory comprehension.]

(2) Controlled conversation with interpretation. A procedure which may be somewhat superior to the controlled interview is what might be called the controlled conversation. Agard and Dunkel (1948, p. 59) describe their test as follows: "Part III, the Conversation, consists of a directed exchange of remarks between a student and a native speaker whose voice is recorded on a phonograph disc. The student is asked to imagine that he is in the company of a friend whose native language is French, German, or Spanish, as the case may be. The friend speaks to him, and immediately afterward another voice on the record directs the student in English what to reply to his friend. For example, the friend may say, "Como esta usted?", whereupon the English word says: "Tell him that you're fine and ask him how he is." Pauses are provided in the record while the student makes his contributions. which are rated by the examiner according to the following scale:

2. Expresses ideas accurately.

1. Partially incorrect; conveys the correct idea but has one or more errors of grammar; conveys almost the correct idea, having one or two errors of vocabulary.

0. Only small part of idea conveyed; wrong idea conveyed; wrong idea conveyed; not understandable; no utterance made.

Agard and Dunlek point out that "the remarks of the foreign friend serve only to provide the illusion of a real conversation, but they do not have to be accurately understood before a correct response can be made." In effect, this type of test is an oral translation exercise, -- but it is really more than translation because the examinee has to manipulate grammatical structures in the light of the situation (e.g., change number. person, tense in verbs, etc.). I believe more work should be done on developing this promising form of test.

A slightly different type of test, along similar principles, is that employed by Sandri and Kaulfers (1945) for an oral fluency scale in Italian, and by Kaulfers (1942) for an oral fluency scale in Spanish.(Actually these two are highly similar.) In this test, the examiner tells the subjects that he is to imagine himself in a foreign country; he is to give the response he would make under various conditions. For example, in Part I ("Securing Essential Services"), the examinee is asked:

How would you tell an Italian:

(1) (a) to speak English?

(2) (b) to open the window?

.

.

(5) to find out where the man went?

.

.

(29 (b) if he knows where the man went?

This pattern is followed in a number of different kinds of situations, and with increasingly difficult questions. This form makes possible a wide variety and sampling of responses within a relative short time, and it should make possible a fairly objective evaluation of responses. By imposing a time-limit on the subject's response to each item, the standardization of the procedure in increased. The materials published by Sandri and Kaulfers for Italian and Spanish can be easily adapted for other languages.

(3)The picture description test. Agard and Dunkel (1948, p. 56) also worked with what they called a Picture Series, in which the examinee was presented a series of simple pictures, each of which could be described with a simple sentence such as "The man is waiting for the train." or "The mouse is eating the cheese." Answers are rated by the examiner.

This form of test would be particularly appropriate where it is desirable or necessary to avoid the use of the native language in the testing. The picture would be pre-tested for clarity and explicitness. It may be necessary to give the subjects some idea of the form of response required: Agard and Dunkel had two sample pictures "with printed answers which would be normal if expected in English."

(c) High degree of control of responses. In contrast with some of the tests discussed earlier, certain kinds of tests provide stimuli which lead directly to specific foreign language responses. Some of the following ideas have apparently never been tried.

(1)A picture naming task. Such tests have been used in English to measure what the writer has called the "Naming factor" (Carroll, 1941). A series of pictures of common objects would be presented, and the examinee would be asked to name them as rapidly as possible. Response would be measured in terms of accuracy and speed. It might be necessary to obtain control measurements on speed of naming in English.

(2) Controlled association test. As is done in certain kinds of psychological testing, the subject could be asked to respond to single stimuli as rapidly as possible so that response latency could be measured. The following variations come immediately to mind:

Stimulus	Response (in foreign language)
English word	Corresponding (foreign language word
Foreign word	Opposite word in foreign language; Species or genus, etc.

Tests of Pronunciation

Attention has been focussed on one particular aspect of oral production, namely accuracy of pronunciation. The traditional method was to ask the subject simply to read a passage aloud; the examiner then attempted to mark every error. This cumbersome method is now being replaced by more objective and reliable techniques.

The principle on which the newer techniques are base is that examples of the subject´s pronunciation of each aspect of the phonology of the language, or a sample thereof, must be deliberately elicited. Evan (1937) found that "a direct oral test as the word, in which the judges rate a single words in each sentence, is almost as good as a measurement of mere accuracy of pronunciation as a longer connected paragraph in which judges attempt to mark every error." What Evans meant by "almost as good" I don´t know (having read only an abstract of her thesis), but I assume that the word test, as a sample, can be made more reliable

and valid as the length of the sample is increased. Lado, at the University of Michigan, has constructed a picture test which is designed to elicit a series of English words exemplifying all the phonemes which give particular give particular trouble to native speakers of Spanish.

Indeed, Lado (see Hill, 1935, p) has claimed that it is possible to measure pronunciation ability by a written test which can be administered by mail if necessary: A typical item in the test presents three pictures, e.g., a picture of a ball, a picture of rain, and a picture of a cake, accompanied by skeleton printed words like b_ll, r_n, c_ke; the subject is to identify the option which as a dissimilar sound. Thus, the test is somewhat similar to Thurstone´s Sound Grouping Test (Thurstone, 1939, a paper-and-pencil test in which the subject is required to eliminate the odd rhyme in groups of words. It is somewhat disconcerting to find that even native speakers of English do not do uniformly well on Thurston's test; indeed, the test tends to correlate with tests of reasoning. A question might be raised, therefore, as to whether native speakers of English would do uniformly well on Lado´s test.

This completes the survey of item types which have been used in measuring foreign language achievement. We must now examine several other problems in constructing foreign language achievement tests.

V. The problem of content

The problem which has continually dogged efforts to devise valid foreign language achievement tests is that of the content which should be included. It is rarely that this problem has arisen in respect to language structure (phonology, grammar, syntax; it most often arises with respect to lexicon and vocabulary.

There should not be any great problem in cases where teste are being constructed as achievement examinations for particular courses of training for here the clear solution is to use the vocabulary and grammar which has been taught in the course.

It is where one has the task of constructing an achievement examination which will apply equally well for a whole gamut of foreign language courses, or which will be valid for "testing knowledge of X language" regardless of the training received, that particular trouble is caused for the test constructor. Most test constructors have had recourse to frequency counts, which exist for most of the European languages commonly taught in the American schools but not for languages like Chinese, Japanese, or Burmese. Even the frequency counts give us trouble, for most of them are based on the literary, written language rather than upon samplings of spoken language. This is quite in order for the construction of tests of reading comprehension, but it does not suffice for tests of spoken language skills. Even when one is concerned only with the word counts of written materials, different results will be obtained depending upon the texts which are includes in the sample to be counted. The very high frequency words in word-counts turn out to be largely the "function words," like the, will, of, act, in English; likewise in other languages.

Perhaps this problem has been exaggerated, however. What with the drawbacks of the existing frequency counts in specific languages, perhaps more attention should be paid to the Semantic Frequency List prepared by Helen S. Heaton (1940). This is an attempt to construct a composite word list which could display the commonest "concepts" in the various European languages. This list might provide a standard to which foreign language tests of structure could be limited. It is always possible to talk about certain nearly universal concepts such as man, woman, boy, girl, day, sun, week, month, walking, running and their likes, using these as a vehicle for the testing of knowledge of the structure and lexicon which one might find, for example, in the principal

parts of a relatively infrequent verb.) Then, where vocabulary as such is to be tested, one may resort to frequency lists to get a first approximation to the probably difficulty of the vocabulary items. It should be remembered that the construction of a vocabulary test even in English, where frequency counts of all sorts are available, it is protracted task, if one counts the time spent in item analysis to determine difficulties, revisions to reflect item-analysis data, etc.

One thing to be avoided, in all probability, is the use of literary or archaic linguistic items. Agard and Dunkel (1948) admit that some of the anecdotes they used in their auditory comprehension tests failed to represent modern colloquial speech.

Still another idea which perhaps has not been sufficiently exploited, is to make a frequency count, not of words used in elementary tests, but of the topics which form their subject-matter from the content point of view. For example, one elementary textbook in German conversation (Goedsche, Wie geht´s? N.Y. Crofts 1938) has its first few lessons on the following topics: greetings between students, making acquaintances, family relationships, time, at tea, and sport. By inspection of other textbooks it might be found that here is a common core of topics which run through a number of texbooks; these could then be used as the basis of achievement examinations.

VI. The Problem of Scaling.

Most of the standardized achievement tests, for example, those of the Cooperative Test Service and even the tests of the Agard-Dunkel investigation are scaled only in terms of percentiles attained by groups with varying amounts of formal training. Such norms may be of use to teachers in judging whether their students are keeping pace with normal progress in language courses, but they are of almost no use in determining what a score on one of these tests actually means. For none of these standardized tests have I ever seen any information which would help in gauging what kinds of

of scores would signify near-native proficiency, what scores would signify minimal ability to conduct routine affairs in a foreign country, etc. Nor do I know of any published reports (except in a thesis by Villareal, 1947) about the administration of these tests to native speakers of the foreign language involved, but this is probably because the English elements in the tests make them inapplicable to native speakers. (The only approach to this has been made in so-called Inter-American test constructed by Manuel, 1950).

It is nevertheless imperative to develop means to obtain quasi-absolute standards for test scores. Questionnaires which inquire of job applicants how fluently they read, write, understand, and speak foreign languages are mute evidence that such standards are needed. How can they be obtained?

A few workers in the field have provided a certain amount of evidence. Shane (1933) was a pioneer in the absolute measurement of vocabulary. His work, which anticipated that of Seashore and Eckerson (1940) in English vocabularies, estimated the average size of active and passive French vocabularies in Florida high school classes. Sandri and Kaulfers (196) offered a system for interpreting scores on their auditory comprehension in Italian:

> 0-100 A. Cannot understand the spoken language.
>
> 101-150 B. Can catch a word here and there and occasionally guess the general meaning through inference
>
> 151-200 C. Can understand the ordinary questions and answers relating to the routine transactions involved in independent travel abroad
>
> 201-225 D. Can understand ordinary conversation on common, non-technical topics, with the aid of occasional repetition or periphrastic restatements
>
> 226-250 E. Can understand popular talks, talking-pictures, ordinary telephone conversations, and minor dialectal variations without obvious difficulty, as well as detect departures from normal usage

The basis for this system was not adequately explained by Sandri and Kaulfers, but it is a step in the right direction. Unfortunatelee it assumes that foreign language achievement is a "unitary trait.". We should assume, on the contrary, that it consists of a number of different aspects, for each of which it would be necessary to construct a scale.

There are two lines of attack on the problem of scaling: First, it is possible to arrange the performances in a test along a difficulty scale (analogous to the mental maturity scale of such a test as he Binet intelligence test), and it is possible to locate an individual's limen on this scale, -- i.e., the point where he has, say, a 50% probability of passing the performances. By expert judgments to the difficulty and importance of the performances at points of the scale, it is possible to gain some idea of the meanings of the scores. For example, if a certain score implies that the individual who gets it has a 50% probability of knowing a group of words which are deemed of rather commonplace usefulness in the language, such as score can be regarded as reflecting somewhat low ability in the language, regardless of what the norms in college classes might suggest.

A second approach is to attempt to obtain a series of scaled scores, or a series of normative values, on groups of native speakers of the language in question. Possibly grade norms could be obtained, so that it would be possible, for example, to say that such-and-such a score on a French language achievement test represents the average achievement of he 3rd grade (or its equivalent) in France.

Unfortunately, most foreign language tests contain such a large freight of English that they are inapplicable to non-bilingual native speakers of the foreign language. The technique mentioned in the last paragraph therefore could be applied only with those tests which incorporate no English elements (beyond the instructions, which could be easily translated). Other kinds of tests could then be calibrated against the tests which are free of English.

Possibly some comparative idea of the level of achievement of typical groups of American language-learners could be obtained by administering language tests which have been developed and standardized in foreign countries. For example, French normative data are available on at least two verbal tests, that of Bonnardel (1951) mentioned earlier, and the vocabulary test developed by Binois and Pichot (undated). Further bibliographic research or correspondence with foreign scholars would probably disclose analogous tests in Spanish, German, Portuguese, Italian, and other languages.

VII. Recommended Procedures in Developing Foreign Language Achievement Tests.

The survey of item types made in preparation for this memorandum makes it possible to outline recommended formats for achievement tests measuring the various language functions. First, a format for tests using English will be given: in most cases, use of English leads to a more convenient, more easily constructed, and probably more reliable and valid form. On the other hand, non-English forms are needed fort use with individuals of heterogeneous language backgrounds, and for use with native speakers of the language in question for purposes of calibration of achievement standards.

TESTS OF FOREIGN LANGUAGE ACHIEVEMENT (Form A, for Use with Native Speakers of English)

1. Test of Reading Comprehension and Speed

a. Vocabulary test: 100 multiple choice items; foreign language words and phrases (context to be held to the minimum necessary to specify meaning intended; usually no context necessary), 5 options in English for each item, In this test, a deliberate attempt is made to probe the extent of the individual's vocabulary in the foreign language; the items will range from the easiest to the most difficult. Work-limit-test.

b. Test of Reading Comprehension: paper-and-pencil test, 100 True-False statements (consisting of one or more sentences). This test would be designed to measure knowledge of structural characteristics of the language; the vocabulary would be limited to high-frequency items, or, where that is not possible, glosses in English would be supplied. The statements could occasionally consist of two or three sentences in order to take advantage of such things as pronoun antecedents, which might confuse some less able examinees. If possible, a list of structure points should be set up; each of the true-false statements should be constructed to turn on one of these structure points. Work-limit test.

c. Tests of Reading Speed: A passage of simple reading material should be altered so that a word in every other sentence or two makes nonsense. The subject is introduced to read this material as fast as possible, crossing out the nonsense words (the latter being as obvious as possible).A time-limit test, in order to measure speed; score is the number of words underlined within the time limit.

2. Tests of Writing and Grammar.

a. Multiple choice test of grammar: 50 items similar to items 2a and 2b illustrated under our discussion of writing tests (p. 21). Work-limit test; objectively scored. An attempt should be made to sample structure points widely.

b. An "interlinear exercise": This type of test has been described previously (p. 23). The subject is asked to edit a passage of connected discourse. Work-limit test; scoring by trained raters.

3. Tests of Auditory Comprehension.

a. True-False statements. A test similar to that which Rulon (1944) prepared for Russian and German; about 50 statements recorded phonographically or on tape. Objective scoring.

b. Multiple choice: pictures associated with spoken stimuli. This also follows the format of tests of this type prepared by Rulon (1944) for German and Russian. Objective scoring.

b1. (Id pictures are too cumbersome). A multiple choice form in which the options are printed in English; the types labeled "multiple-choice (b)" or "multiple-choice (c)" (see pages 30-31) are judged of mots general usefulness. This test can be objectively scored.

c.Following directions; An attempt should me made to construct a test of this type in order to test the ability to follow a lengthy discourse. The speech could start slowly, then increase gradually to "normal" rate and to "fast" speech, in order that speed of auditory comprehension may be calibrated.

4.Tests of Oral Production

a.Controlled conversation (interpretation). It is recommended that a test modeled closely along the lines of that constructed by Sandri and Kaulfers (1945) should be developed. This type asks the examinee such questions as how he would tell an Italian to speak English, to open the window, to find someone to repair his car, etc. (see p. 40).

b. Test of pronunciation. A distinct set of printed words or phrases exemplifying the phonological distinctions of the language is to be read aloud by the examinee, whose response is to be evaluated by the examiner.

TESTS OF FOREIGN LANGUAGE ACHIEVEMENT (Form B, Use with Persons who do not Speak English; Can Also be Used with Native Speakers of English)

1.Tests of Reading Comprehension and Speed.

a. Vocabulary Test: Paper-and-pencil test, same as corresponding test in Form A, but multiple choice options in the foreign language, consisting usually of phrases of less vocabulary difficulty than the lead words.

b. Test of Reading Comprehension. Same as in Form A (T-F statements).

c. Test of Reading Speed. Same as in Form A.

2. Tests of Writing and Grammar,

a. Multiple choice test of grammar. Similar to that in Form A, but omit English cue, and make options such that only one is correct usage of grammar.

b. Interlinear exercise. Same as in Form A.

3. Tests of Auditory Comprehension.

a. True-False statements. Same as in Form A.

b. Multiple choice. pictures associated with spoken stimuli. Same as in Form A.

b1. (If pictures are too cumbersome). same as in Form A, but with options printed in the foreign language.

c.Following directions. Same as in Form A.

4. Tests of Oral Production

a.Picture description Test. This will be similar to the Picture Series developed by Agard and Dunkel (1948)

b. Test of pronunciation. Same as in Form A.

Appendix A

List of Available Achievement Tests in Foreign Languages

(Note: To save space, citations of tests are considerably abbreviated. Where possible, a reference is given to its citation in Buros's series of mental measurement yearbooks; 40 denotes an entry in the Nineteen Forty Mental Measurements Yearbook; 49 denotes an entry in The Third Mental Measurements Yearbook; and 53 denotes an entry in The Fourth Mental Measurements Yearbook.)

ENGLISH AS A FOREIGN LANGUAGE

Test	Type	Remarks
English examination for foreign students, including a test of non-verbal reasoning. Educ. Testing Service, 1947. See Buros 53:233.	Omnibus	Reading Comprehension, Aural Comprehension, Pronunciation, Composition.
English language test for foreign students. 1951. Robert Lado. See Buros 53:234.	Pronunciation, Grammar, Vocabulary (no spoken material.)	Reviewed very favorably.
Test of Aural Comprehension in English as a Foreign Language. 1946. Robert Lado. See Buros 53:235.	Auditory comprehension.	A good test; useful as a model.
[Villareal, Jesse J.] Test of Aural Comprehension of English for Native Speakers of Spanish. Contained in Ph.D. Thesis, Northwestern Univ., 1947.	Auditory comprehension.	Useful; carefully constructed.

Note: For more information on tests of English as a foreign language, see the surveys by Lado (1950; 1953-54).

FRENCH

Columbia Research Bureau Aural French Test. 1930. Seibert and Wood. World Book Co. See Buros 40:1347.	Auditory comprehension.	T-F questions, parts would be useful.
ACE French Reading Test. 1937-39. Cheydleur, Hermon, Walker. Coop. Test Service. See Buros 40:1346.	Reading comprehension, vocabulary.	Competently done.
American Council Alpha French Test. 1926-27. World Book Co. See Buros 40:1342.	Vocab., grammar, Silent reading, Composition (with quality scale.)	Highly regarded. Item analysis data in Hermon (1939)
American Council Alpha French Test: Aural Comprehension. 1933. Rogers and Clarke. Teachers Coll. Bur. Publications. See Buros 40:1343	Auditory comprehension.	An early attempt. Some useful materials complete test not recommended

FRENCH (continued)

American Council Beta French Test. Greenberg and Wood. 1926-27. World Book Co. See Buros 40:1344.	Printed vocabulary, comprehension, grammar.	Useful materials. Item analysis data in Wood (1927).
American Council French Grammar Test. 1927. Cheydleur. World Book Co. See Buros 40:1345.	Printed grammar.	Useful, but sampling of content could be improved.
[Bovée, A.G. Tests of French reading comprehension.] Contained in:Bovee,1947. (see bibliography).	Reading comprehension.	Supposed to be comparable to Thorndike-McCall reading test in English.
[Bovée, A.G. Test of auditory comprehension. Contained in: Bovee,1948 (see bibliography).	Auditory comprehension	Should be useful. Not completely objective.
Cohen French Test. 1945-49. S.W. Cohen, Australian Coun. for Educ. Res. See Buros 53:236.	Vocabulary Silent Reading Grammar Aural Comp.	Reviewer feels it was carelessly edited.
CEEB Achievement Test in French Reading. See Buros 53:237.	Vocabulary, grammar, reading comp.	Not available to public.
Columbia Research Bureau French Test. 1926. Meras,Roth,and Wood. World Book Co. See Buros 40:1348.	Vocabulary, Reading Comp,, Grammar	Generally satisfactory.
Cooperative French Comprehension Tests. 1942-47. See Buros 49: 180; 53:238	Vocabulary, Reading Comp.	Technically, very competent.
Cooperative French Test: Elementary and Advanced Levels. 1939-41. See Buros 40:1349-50;49:181.	Vocabulary, Reading Comp.	Technically, very competent.
Cooperative French Tests: Lower and Higher levels. 1942-47. See Buros 49:182; 53:238.	Vocab.,Reading Comp.,Grammar, Fr.Civilization	Technically competent.
Examination in French Grammar. 1944-45. USAFI. See Buros 49:183.	Grammar.	Item construction often questionable.
Examination in French Reading Comprehension. 1944-45. USAFI. See Buros 49:184.	Reading Comp.	Item construction often questionable.
Examination in French Vocabulary. 1944-45. USAFI.See Buros 49:185.	Vocabulary(wrds. always in context.)	Item construction often questionable.
French I and II:Achievement Examinations for Secondary Schools. 1951. W.W.Cook. Educ. Test Bureau. See Buros 53:239.	Omnibus.	Incompetently done Not recommended.
French Grammar Test: Dominion Tests. 1940-41. Univ. of Toronto. See Buros 49:186.	Grammar	Poorly edited. Not recommended.

FRENCH (continued)

French Reading: Dominion Tests. 1940-41. Univ. of Toronto. See Buros 49:187.	Reading comp. (paragraphs)	Not recommended
French Recognition Vocabulary Test 1948. E.R.Ryden. Purdue Univ. See Buros 53:240	Printed vocabulary	Generally satisfactory.
French Vocabulary Test:Dominion Tests. 1940-41. Univ. Toronto. See Buros 40:1353; 49:188.	Printed vocabulary	Generally satisfactory.
Graduate Record Examinations: Advanced French Test. 1939-51. ETS. See Buros: 53:241.	Not known.	Not publicly available.
Iowa Placement Examinations: French Training, Series FT1,Rev. 1924-26. See Buros 49:189.	Vocab.,grammar, reading comp.	Completion items involve subjective scoring.
Lundeberg-Tharp Audition Test in French. 1934. J.B.Tharp, Ohio State Univ.,See Buros 40:1354.	Auditory comprehension	Useful; generally satisfactory.
Miller-Davis French Test. 1935. Kansas State Teachers College. See Buros 40:1355.	Omnibus.	Not recommended. Tries to cover too much ground.
Standard French Test.1929. Sammartino and Krause. Bloomington, Ill.: Pub.Sch.Pub.Co. See Buros 40:1356.	Vocab.,grammar, comprehension	Generally satisfactory.
A standardized French Grammar Test. 1951. T.S.Percival. Univ. of London. See Buros 53:242.	Grammar.	Satisfactory.
A standardized French Vocabulary Test. 1951 T.S.Percival. Univ. of London. See Buros 53:243.	Vocabulary.	Generally satisfactory. Needs more editing.
Test B.V. C-8. Contained in Bonnardel,1951. (see bibliography)	Vocabulary.	Designed for native French speakers.
Test de vocabulaire. Binois and Pichot, Centre de Psychologie Appliquee, France.	Vocabulary.	Designed for native French speakers.
Univ. of Chicago Aural Comprehension Tests in French. Lower and Upper Levels. Available from Veteran's Testing Service, 5741 Drexel Ave.,Chicago 37,Ill.	Auditory comprehension.	The parts based on short sentences or questions are generally satisfactory; the questions based on connected discourse and dialogues are poorly constructed and edited.

GERMAN

American Council Alpha German Test. 1926-27. Henmon, Morgan, Hinz, Purin, Rossberg. World Book Co. See Buros 40:1357.	Vocab., grammar, reading comp., composition.	Highly regarded. Item analysis data in Henmon (1929).
ACE German Reading Test. 1937-38. Appelt and Henmon. Coop.Test Service, ETS.	Reading Comp.	Competently done.
CEEB Achievement Test in German Reading. See Buros 53:244.	Vocab., grammar, Reading.	Not available to the public.
Columbia Research Bureau German Test. 1926-27. Purin and Wood. World Book Co. See Buros 40:1359.	Vocab., grammar, Reading.	Generally satisfactory.
Cooperative German Tests. (Various levels and forms) See Buros 40: 1360; 49:190; 53:245.	Vocab., grammar, reading.	Highly competent technically.
Examination in German Grammar: Lower Level.1945. USAFI. See Buros 49:191.	Grammar.	Generally satisfactory.
Examination in German Reading Comprehension:Lower Level.1945. USAFI. See Buros 49:192	Reading comp.	Too literary; vocab. not well controlled.
Examination in German Vocabulary: Lower Level. 1945. USAFI. See Buros 49:193	Vocabulary.	Competent.
German I and II:Achievement Examinations for Secondary Schools. 1951. W.W.Cook. Educ.Test Bureau See Buros 53:246.	Omnibus,	Not recommended.
Graduate Record Examinations: Advanced German Test. 1939-51. ETS. See Buros 53:247.	Not known.	Not available to public.
Lundeberg-Tharp Audition Test in German. 1929. J.B.Tharp, Ohio State Univ. See Buros 49:194.	Pronunciation, Aural comprehension. (Not recorded)	Useful; generally satisfactory.
Rulon, P.J. Aural comprehension test in German Contained in Rulon, P.J. et al., Report on contract test constructed for the ASTD, ASF, Contract No. W-19-073 AST(SCI)-26, Comprehension of spoken German, Term 6. Harvard University., March 1944.	Auditory comprehension. (identify pictures described orally) (not recorded)	Recommended. (Higher level)
Rulon, P.J. et al.] German Interview Rating Scale (Series A). RCA Records, 1944. Records ND3-MC-3463 to 3470, and 3488-3490.	Oral production (Individual examiner)	Useful as a possible model; this material is too specific to the ASTP situation.

GERMAN (continued)

[Rulon, P.J., et al.] Oral German Comprehension test. RCA Records, 1944. Discs ND3-MC-3473 to 3480.	Auditory comp. (T-F statements) (phonograph rec.)	Generally satisfactory.
Univ. of Chicago Aural Comprehension Tests in German. Lower and Upper Levels. Available from Veterans Testing Service, 5741 Drexel Ave., Chicago 37, Ill.	Auditory comp. (with phonograph records)	The parts based on short sentences or questions are generally satisfactory; the questions based on connected discourse and dialogues are poorly constructed and edited.

ITALIAN

College Entrance Examination Board Achievement Tests in Iatlian Reading. See Buros 53:249.	Vocab., grammar, Reading comp.	Not available to public.
Cooperative Italian Test. 1947. See Buros 40:1362; 49:199.	Vocab., grammar, reading, culture.	Apparently not quite as competent as Coop. tests in other languages. Needs editing.
Examination in Italian Grammar: Lower Level. 1945. USAFI. See Buros 49:200.	Grammar.	(No review available)
Examination in Italian Reading Comprehension, Lower Level. 1945. USAFI. See Buros 49:201.	Reading Comp.	(No review available)
Examination in Italian Vocabulary: Lower Level, 1945. USAFI. See Buros 49:202.	Vocabulary.	(No review available)
[Sandri-Kaulfers Oral-Fluency Rating Scale in Italian.] Contained in Sandri and Kaulfers, 1945 (see bibliography)	Oral production. (Individual examiner)	Excellent; the best scale of its type I have seen.
[Sandri-Kaulfers Aural Comprehension Scale in Italian.] Contained in Sandri and Kaulfers, 1946 (see bibliography)	Auditory comprehension. (for local recording.)	Excellent; in general, the best scale of its type I have seen.

RUSSIAN

[Rulon, P.J., et al. Aural Comprehension test in Russian.] Contained in: Rulon, P.J. et al. Report on contract test constructed for the ASTD, ASF, Contract No. W-19-073 AST(SCI)-26: Comprehension of Spoken Russian, Term 6, Harvard Univ., March 1944.	Auditory comp. (identify pictures described orally)	Recommended for upper level testing.

RUSSIAN (continued)

[Rulon, P.J. et al.] Russian Interview Rating Scale. RCA Records, 1944. Records ND3-MC-3455 to 3462 and 3492-94.	Oral production (quality scale)	Useful as a possible model; material too specific.
[Rulon, P.J. et al.] Oral Russian Comprehension Test. RCA Records 1944. Records ND3-MC-3471 to 3472 and 3481-86.	Auditory comp. (T-F statements)	Generally satisfactory.
University of Chicago Aural Comprehension Tests in Russian. Lower and Upper Levels. Available from Veterans Testing Service, 5741 Drexel Ave., Chicago 37, Ill.	Auditory comp.	(Same remark as for Univ. of Chicago French and German tests.)

SPANISH

American Council Alpha Spanish Test. 1926-28. Buchanan, Crawford, Keniston, Henmon, World Book Co. See Buros 40:1371.	Vocab., grammar, Reading, composition.	Generally satisfactory.
College Entrance Examination Board Achievement Tests in Spanish Reading. See Buros 53:259.	Vocab., grammar, reading comprehension.	Not available to public.
Columbia Research Bureau Spanish Test. 1926-27. Callcott and Wood. World Book Co. See Buros 40:1372.	Vocab., grammar, reading comp.	Generally satisfactory.
Cooperative Spanish Test: Lower and Higher Levels. 1948-51. See Buros 40:1373; 40:1374; 53:260.	Comprehension, grammar, civilization.	Competent.
Examination in Spanish Grammar: Lower Level. 1944. USAFI. See Buros 49:208.	Grammar.	Generally satisfactory.
Examination in Spanish Reading Comprehension. 1944. USAFI. See Buros 49:209.	Reading comp.	Reviewed somewhat unfovorably.
Examination in Spanish Vocabulary: Lower Level. 1944. USAFI. See Buros 49:210.	Vocabulary.	No review available. Probably has disadvantage that vocab. is tested always in context.
First Year Spanish Test. O.H. Patterson. 1945. Purdue Univ. See Buros 53:261.	Not known.	No review available.

SPANISH (continued)

Furness Test of Aural Comprehension in Spanish. 1945-51. E.L. Furness. Banks Upshaw Co., Dallas 1, Texas. See Buros 49:213; 53:262.	Auditory comprehension. (Recordings available: phono, tape, wire.)	Generally satisfactory, some defects.
Graduate Record Examinations: Advanced Spanish Test. 1946-51. See Buros 53:263.	Not known.	No review available.
Iowa Placement Examinations: Spanish Training: Series ST1, Revised. 1924-25. Bur. Educ. Res. and Serv., State Univ. of Iowa. See Buros 49:212.	Vocab., grammar, reading.	Fairly satisfactory.
Kansas First Year Spanish Test. 1947. M.M.Miller. Kansas State Teachers College. See Buros 53:264.	Not known.	No review availab...
[Kaulfers, W.V. Oral-fluency test in Spanish.] Contained in: Kaulfers, 1944 (see bibliography)	Oral production.	Excellent, for its type. Usable as it stands.
Lundeberg-Tharp Audition Test in Spanish. 1929. J.B.Tharp, Ohio State Univ., See Buros 49:211.	Auditory comprehension.	Generally satisfactory.
[Manuel, H.T., et al.] Tests of Language Usage: Active Vocabulary and Expression: Cooperative Inter-American Tests. 1950. ETS. See Buros 53:176.	Vocabulary and grammar -- parallel English and Spanish editions.	Not favorably reviewed; probably not appropriate for testing Spanish achievement of English speakers.
Spanish I and II; Achievement Examinations for Secondary Schools. 1951. W.W.Cook. Educ. Test Bureau. See Buros 53:265.	Not known.	No review available. (parallel tests in French and German unfavorably reviewed.)
Stanford Spanish Tests. 1927. Espinosa and Kelley. Stanford Univ. Press. See Buros 53:266.	Grammar, vocab., reading comp.	Very favorably reviewed. Tests have continued in demand.
Univ. of Chicago Aural Comprehension Test in Spanish. Lower and Upper Levels. Available from Veterans Testing Service, 5741 Drexel Ave., Chicago 37, Ill.	Auditory comprehension.	See remarks for Univ. of Chicago tests in French and German.

BIBLIOGRAPHY

Agard, F.B., and Dunkel, H.B. An investigation of second-language learning. Boston: Ginn, 1948. vii, 344 p.

Anderson, I.H., and Fairbanks, G. Common and differential factors in reading and hearing vocabulary. Journal of Educational Research, 1937, 30, 317-324.

Andrus, Lawrence. Reporting a test. Modern Language Journal, 1942, 26, 368-374.

Binois, R., and Pichot, P. Test de vocabulaire. Published by the Centre de Psychologie Appliquée, 15, rue Henri Heine, Paris XVIe, France. (undated)

Bonnardel, R. Étude d'une épreuve de compréhension du vocabulaire, le test B. V. C-8. Le Travail Humain, 1951, 14, 77-89.

Bovée, Arthur G. A study of the relationship between visual thought comprehension in English and in French. French Review, 1947, 21, 120-123.

Bovée, Arthur G. The relationship between audio and visual thought comprehension in French. French Review, 1948, 21, 300-305.

Buchanan, M.A., and McPhee, E.D. An annotated bibliography of modern language methodology. Toronto: Univ. of Toronto Press, 1928. (Publications of the American and Canadian Committees on Modern Languages, Vol. 8)

Buros, Oscar K.(Editor) The Nineteen Forty Mental Measurements Yearbook. Highland Park, N.J.: The Mental Measurements Yearbook, 1941.

Buros, Oscar K.(Editor) The Third Mental Measurements Yearbook. New Brunswick, N.J.: Rutgers Univ. Press, 1949.

Buros, Oscar K.(Editor) The Fourth Mental Measurements Yearbook. Highland Park, N.J.: The Gryphon Press, 1953.

Carroll, J.B. A factor analysis of verbal abilities. Psychometrika, 1941, 6, 279-307.

Carroll, J.B. Foreign language teaching: the state of the art. Staff Research Memorandum, Training Aids Laboratory, Air Force Personnel and Training Research Center, Chanute Air Force Base, Ill. 5 May 1954. Mimeographed.

Cole, R.D., and Tharp, J.B. Modern Foreign Languages and their Teaching. (New York, D. Appleton-Century Co., 1937.)

College Entrance Examination Board. Questions set at the Examination of 1928. Boston: Ginn & Co., 1928.

College Entrance Examination Board. English composition, a description of the English composition test of the College Entrance Examination Board. June 1954.

College Entrance Examination Board. Foreign Languages, a description of the College Board tests in French, German, Latin, and Spanish. Princeton, N.J.: April 1954.

Committee on Resolutions and Investigations appointed by the Association of Modern Language Teachers. Report f teachers of the middle states and Maryland. Modern Language Journal.,1917, 1, 250-261.

Coutant, Victor. Evaluation in foreign language teaching. Modern Language Journal, 1948, 32, 596-599.

Dorcus, Roy M., Mount, G.E., and Jones, M.H. Construction and validation of foreign language aptitude tests. Personnel Research Branch Research Report 993. 30 June 1952. 27 p.

Doyle, H.G. Review of: Fiske, T.S., The work of the College Entrance Examination Board, 1901-1925. (Boston, Ginn, 1926) Modern Language Journal, 1927, 10, 568-570.

Dunkel, Harold B. Second-language learning. Boston:Ginn, 1948. vi,218p.

Dyer, H.S. The validity of certain objective techniques for measuring the ability to translate German into English. J. educ. Psychol., 1946, 37, 171-178.

Eaton, Helen S. Semantic frequency list for English, French, German, and Spanish: a correlation of the first six thousand words in four single language frequency lists. Chicago: Univ. Chicago Press, 1940. x+1, 441 p.

Eurich, Alvin C. Minnesota Speed of Reading Test for College Students. Minneapolis, Minn.: Univ. of Minnesota Press, 1936.

Evans, Marjorie Katherine. The measurement of French pronunciation. Unpublished MA thesis, Ohio State U.,1937. pp. 49. (briefed in Coleman ABMLT iii, 590 (p. 365)

Fichen Clarence E. Intercorrelations of part scor[illegible] in foreign language tests. Doctor's diss., U. Wisc. 1937. (Briefed in Coleman, ABMLT, ii, 859)

Fiske, Thomas S. Examination to test competence in the use of the English language. Pp. 18-21 in The Thirty-Fourth Annual Report of the Secretary, 1934. New York: College Entrance Examination Board, 1934. (See Coleman, ii, #862.)

Fuess, Claude M. The College Board, its first fifty years. New York: Columbia Univ. Press, 1950. 224. p.

French, John W. The description of aptitude and achievement tests in terms of rotated factors. Chicago: Univ. Chicago Press. 1951. x. 278 p. (Psychometric Monograph No. 5)

Gibbons, H. The ability of college freshmen to construct the meaning of a strange word from the context in which it appears. J. exper. Educ., 1940, 9, 29-33.

Gray, H.A. Recorded sound in the field of achievement testing. J. educ. Res., 1938, 31, 608-615.

Greene, H.A., Jorgensen, A.N., and Gerberich, J.R. Measurement and evaluation in the secondary school. N.Y.: Longmans, Green, 1954. xxii, 690 p.

Hall, Ernest J. Oral examinations in Spanish for undergraduates. Hispania, 1936, 19, 461-466. (Coleman, ii, 873.)

Handschin, C.H. Tests and measurements in modern language work. Mod. Language Journal, 1920, 4, 217-225. (Briefed in Buchanan & McPhee, p380)

Hawkes, Herbert E., Lindquist, E.F., and Mann, C.R. (Editors) The construction and use of achievement examinations; a manual for secondary school teachers. Boston: Houghton Mifflin Co., 1936, x, 497 p.

Henmon, V.A.C. Achievement tests in the modern foreign languages. N.Y.: Macmillan, 1929. xxvi, 363 p. (Vol 5, Publications of the American and Canadian Committees on Modern Languages.)

Hill, A.A. (Editor) Report of the 4th Annual Round Table Meeting on Linguistics and Language Teaching. Georgetown Mono. Series on Languages and Linguistics, No. 4, September 1953.

Holt Spoken Language Series. N.Y. Henry Holt & Co. (various years)

Jordan, A.M. Measurement in education. N.Y.: McGraw-Hill Book Co., Inc., 1953. Chapter 8, pp. 207-224.

Josselson, H.H. The Russian word count and frequency analysis of grammatical categories of standard literary Russian. Detroit: Wayne Univ. Press, 1953. 274 p.

Kamman, James Foster. A comparison of factor patterns in a native language and an auxiliary language. Ph.D. Thesis, 1953, U. Illinois. Microfilm Abstracts, 1954, 14, 406.)

Kaplan, H. and Berkhouse, R.G. Survey of the literature on measurement of oral fluency in foreign languages. Amer. Psychologist, 1954, 9, 403. (Abstract) (review of oral production tests in foreign languages presented at Sept 1954 APA convention)

Kaulfers, W.V. Objective tests and exercises in French pronunciation. Modern Language Journal, 1937, 22, 186-188.

Kaulfers, W.V. Modern Languages for Modern Schools. N.Y.: McGraw-Hill, 1942. pp. 525.

Kaulfers, W.V. Wartime developments in modern-language achievement testing. Mod. Language Journal, 1944, 28, 136-149.

Kellenberger, Hunter (Editor). Committee reports, 1954 Northeast Conference on the Teaching of Foreign Languages. Brown Univ., Division of Modern Languages, 1954.

Lado, Robert. Survey of tests in English as a foreign language. Language Learning. 1950, Vol. 3,(no. 1-2) 51-66.

Lado, Robert. Linguistic science and language tests. Language Learning. 1950, Vol. 3 (no. 3-4), 75-82.

Lado, Robert. Testing control of the structure of a foreign language. Language Learning. 1951-52, Vol 4(no. 1-2.), 17-35.

Lado, Robert. Materials and tests in English as a foreign language: a survey. Language Learning, 1953-54, 5(1-2), 48-55.

Learned, William S., and Wood, Ben D. The student and his knowledge. Bull. 29, Carnegie Foundation for the Advancement of Teaching. N.Y.: Carnegie Foundation for the Advancement of Teaching. 1938. pp. xx,406.

Manuel, Herschel T. Tests of Language Usage: Active Vocabulary and Expression: Cooperative Inter-American Tests. Cooperative Test Division, Educational Testing Service, 1950.

Monroe, W.S., De Voss, J.C. and Kelly, F.J. Educational tests and measurements. Boston: Houghton-Mifflin, 1917.

Odell, C.W. Educational measurements in high school. N.Y. Appleton-Century-Crofts, 1940.

Poston, L.S., and Clark,R.E. French Syntax List. N.Y.: Henry Holt & Co., 1943.

Pressey, S.L., and Pressey, L.C. Introduction to the use of standard tests. Yonkers-on-Hudson, World Book Co., 1923. vi, 263 pp.

Rogers, Agnes L., and Clark, Frances M. Report on the Bryn Mawr test of ability to understand Spoken French. Modern Language Journal, 1933, 17, 241-238. (Coleman, ii, 874).

Ross, Lawrence W. Pronunciation quiz for French. High Sch. J., 1937, 20, 96-97. (briefed in Coleman, ii, 875.)

Ruch, G.M. and Stoddard, G.D. Tests and measurements in high school instruction. Yonkers, N.Y., World Book Co., 1927.

Rulon, P.J. Report on contract test constructed for the ASTD, ASF, Contract No. W-19-073 AST(SC-1)-26: Report on Scales for Measuring Ability to Speak German and Russian, Term 5. Harvard Univ.,December 1943

Rulon, P.J. Report on contract test constructed for the ASTD, ASF, Contract No. W-19-073 AST(SC-1)-26: Comprehension of Spoken Russian, Term 6. Harvard University, March 1944.

Rulon, P.J. Report on contract test construction for the ASTD, ASF, Contract No. W-19-073 AST(SC-1)-26: Comprehension of spoken German, Term 6. Harvard University, March 1944.

Rulon, P.J. Report on Contract Test Constructed for the ASTD, ASF, Contract No. W-19-073 AST(SC-1)-26: Comprehension of Spoken Russian, Items Proposed for Use in War Department Tests. Harvard Univ.,March'44.

Sammartino, Peter. Improvement curves in the comprehension of printed French and in the acquisition of French vocabulary. Unpublished doctor's diss., N.Y. Univ.,1931. (Briefed in Coleman, ABMLT, ii,868.)

Sandri, Luigi, and Kaulfers, W.V. An oral-fluency rating scale in Italian. Italica,1945, 22, 133-144.

Sandri, Luigi, and Kaulfers, W.V. An aural comprehension scale in Italian. Italica, 1946, 23, 335-351.

Schenck, Ethel A. Studies of testing and teaching in modern foreign languages, based on materials gathered at the University of Wisconsin by the late Professor Frederic D. Cheydleur. Madison, Wisconsin: Dembar Publications, Inc., 1952, vi, 72 pp.

Seashore, R.H., and Eckerson, Lois D. The measurement of individual differences in general English vocabularies. J. educ. Psychol., 1940. 31, 14-38.

Shaeffer, Rudolph F. What kind of tests for oral-aural courses? German Quarterly, 1948, 21, 94-101, 153-157.

Shane, Milton Lanning. Measuring the extent of French vocabulary of high school students. (Briefed in Coleman, ii, 535).French Review,1933, 7, 108-124.

Smith, Francis Prescott, and Campbell, Helen S. Objective achievement testing in French: recognition versus recall tests. Mod. Language Journal, 1942, 26, 192-198.

Stalnaker, J.M. and Kurath, W.A. A comparison of two types of foreign language vocabulary test. J. of Educ. Psychol., 1935, 26, 435-42. (Briefed in Coleman, ii, 853).

Symonds, Percival M. Measurement in secondary education. N.Y.: Macmillan 1927.

Thurstone, L.L. Primary mental abilities. Chicago: Univ. of Chicago Press, 1938. ix, 121 p. (Psychometric Monograph No. 1)

Villareal, Jesse J. A Test of the Aural Comprehension of English for Native Speakers of Spanish. (Northwestern Univ. dissertation, 1947.) See Northwestern Univ. Summaries of Dissertations, Vol XV, June-Sept 1947, pp. 77-82.

Werner, Heinz, and Kaplan, Edith. Development of word meaning through verbal context: an experimental study. Journal of Psychol., 1950, 29, 251-257.

Wittenborn, J.R., and Larsen, Robert P. A factorial study of achievement in college German. J. educ. Psychol., 1944, 35, 39-48.

Wood, Ben D. New York experiments with new-type modern language tests. N.Y.: The Macmillan Co., 1927, xxii, 339 p. (Publications of the American and Canadian Committees on Modern Languages, Vol. 1)