This is not the first conference ever called on language testing, nor will it be the last one. Language testing has a long history, and much experience and wisdom have accumulated. Most test constructors are aware of the various kinds of tests that can be built; they are knowledgeable about such matters as item construction, item analysis, reliability, validity, standardization, norms, and so forth. Most testers can recite all the arguments pro and con the use of objective tests, the use of "translation" in testing, and the use of incorrect linguistic forms. Language testing has reached such a stage of professionalization that a whole book on the subject is about to appear—or perhaps has already appeared. A speaker who is asked to talk about the "theory of language testing" is surely not expected to go back to elementary principles, as if the audience were not already adequately indoctrinated. The work papers which you have already had a chance to see are impressively sophisticated on many aspects of testing. All I can expect to do, therefore, is to readdress the attention of the audience to certain basic and fundamental problems and points of view, some of which may have been lost sight of in the heat of enthusiasm for technical detail.

Reprinted by permission from *Testing*, pp. 31–40, Center for Applied Linguistics, 1961.
Dr. Carroll, former professor of educational psychology at Harvard University, is now with the Center for Psychological Studies, Educational Testing Service, Princeton, N.J.

Let us speak first of the *purpose of testing*. The purpose of testing is always to render information to aid in making intelligent decisions about possible courses of action. Sometimes these decisions affect only the future design or use of the tests themselves, in which case we are dealing with solely experimental uses of tests. Sometimes the decisions have to do with the retention or alteration of courses of training, as when one decides that poor test results are due to ineffective training. Most often, the decisions have to do with the management of the educational careers of individuals. Different stages of the training cycle call for different kinds of tests. Before training begins, one may wish to predict learning rate or ultimate success in training—and a test which will do this validly is called an aptitude test. At the start of training we may also wish to give *pretests* to ascertain the status of the individual's skill or knowledge before training, and as a basis for measuring the true effects of training when pretest scores are eventually compared with posttest scores. Educators make all too little use of pretests, although foreign language teachers are probably frequently right in their assumption that their students start at a virtual zero point. At various points in a particular training course, and especially at the end, achievement tests are given to ascertain progress and to diagnose learning difficulties. The content and design of such tests can be tailored to suit the purposes of the designers of the course. Somewhat different problems arise when there is a problem of assessing achievements of learners in a variety of courses which ostensibly have the same subject-matter but which actually vary considerably in content and rationale. Such examinations are often called *external examinations* and are exemplified by the College Entrance Examination Board Achievement tests in a number of secondary school subjects, including several foreign languages. Note that although the construction of the College Board examinations is supervised by a nationally representative committee of teachers and subject-matter authorities who come to agreement on a common content on which the examination is to be built, there is still much discrepancy, in some cases, between what a student has been exposed to and what the examiners assume he has been exposed to. It is small wonder that a proposed external examination on English proficiency, designed for the testing of candidates from many countries and courses, will have to face the fact of profound differences in the kinds of preparation that these candidates will have had. Of course, after a program of external examinations has been conducted over a sufficient number of years, the nature of the test program gets to be known to the teachers, who may actually begin to shape their teaching in the direction of the examination. But if an external examination in English is to be held on a world-wide basis, natural forces will hardly be sufficient for insuring a modicum of uniformity in course content. A sensible step, it seems to me, would be for the universities who are to use the results of a world-wide test to announce the kind of product they seek and the kind of training course content they would expect candidates to have mastered at a minimum. American universities should try

to get together to specify what kinds and levels of English language proficiencies they desire in foreign students. If this were done, it would undoubtedly have a beneficial influence upon English training courses abroad and hence upon the preparation of students. Incidentally, the development of specifications for the desired product would be a necessary step in the design of the examination or examinations.

Nevertheless, there is a further step that should be taken. The mere announcement of desired standards could all too easily be done without adequate forethought and without taking stock of the actual experiences of foreign students. There is need for surveys of the kinds of linguistic situations faced by these students and the success or failure of students at various levels of English proficiency in meeting these situations. For example, what kinds of English mastery are required for the foreign student to comprehend reading matter and lectures in the several academic disciplines? What standard of proficiency in English pronunciation is the minimum required for foreign students to be understood by American students and teachers? What are typical social situations in which foreign students must engage? In which language skills is there most deficiency in failing or dropout cases? Perhaps these analyses have all been undertaken, but if so, I am unaware of them.

In the present case, the external examination in English proficiency takes on some of the characteristics of an aptitude test, at least in the sense that it would be designed to help predict success in collegiate subject-matter courses. After a preliminary test has been designed and administered, its scores should be kept on file and carefully compared with students' performance in their subject-matter courses in American universities. Only on the basis of such information will it be possible to establish rules and guides for using examination results to make decisions concerning the admissibility of foreign students to American universities. This requirement may entail selection and placement of a few foreign students who are below standard. But because the external examination is so largely an *aptitude test*, it should be designed and constructed like an aptitude test; i.e., it should be subjected to external validation. External validation in this case, would be solely against the criterion of *having sufficient English to operate in given situations*. Thus, measurements would be taken of students' ability to comprehend lectures and reading material, and ability to be understood in specified social settings. Obviously, all these procedures would ideally entail a not inconsiderable program of validation research, but I believe the outcome would be worth the effort.

The kinds of performance that should be tested on the proposed external proficiency examination should, then, be dictated by, and flow from, the specifications of the performance abilities desired in candidates to be selected by the examination. These specifications may be divided up into several levels. For example, there is already the suggestion that students entering engineering schools do not have to be as proficient in English as those entering liberal arts

colleges and majoring in literature and the humanities. Some students may not have to be as proficient in speaking as others.

The specification will refer, then, to a series of at least logically independent kinds of performances. For example, speaking ability and reading ability are logically quite different kinds of performances, because one can exist without the other; this is true even though they may be rather highly correlated in a given group of people simply because this group of learners had common training experiences which led them to perform equally well, on the average, on both types of performance. Logically different kinds of language performances have sometimes been identified by using a grid in which different kinds of mastery are displayed against different aspects of the language structure which one is testing. Thus, one could theoretically obtain measures of ability in each cell of the following grid:

	Language Aspect			
Skill	*Phonology or orthography*	*Morphology*	*Syntax*	*Lexicon*
Auditory comprehension				
Oral production				
Reading				
Writing				

In practice, however, it would be foolish to attempt to obtain these sixteen different measures, for this would be carrying the process of analysis too far. It is unlikely that ability varies in precisely these sixteen independent ways. As far as morphology, syntax, and lexicon are concerned, the important aspect of ability which should be tested is the individual's knowledge of the "facts" of the language itself. Since language is recognized as primarily a vocal phenomenon, one would prefer to test these in spoken form, but in the case of English as a second language for foreign students, we can agree to use the expedient of written testing because use of written language is taught to these students and is expected of them in American colleges. One can assume that all the "facts" of the language, such as the constancies of its grammar and the meanings of its words and grammatical structures, form a more or less continuous spectrum of frequency and utility—from the most frequent and useful items to the

rarer and only occasionally useful items. One of the main jobs of the language test is to determine how far towards the end of this spectrum the examinee can demonstrate a substantial and useful degree of knowledge. It is a matter of convenience, however, to separate structural and lexical aspects, partly because structural and lexical items may be learned in somewhat different ways or in different contexts. Further, it can be argued that mastery of structure is more essential than, and in a sense prerequisite to, mastery of lexicon.

For the sake of clarity let us list the separate aspects of language competence which might be considered in drawing up specifications for a proficiency test. Let us begin with a large category which we will call:

1. *Knowledge of structure* (morphology and syntax). The *knowledge* aspect is emphasized here because we will first be concerned with what the individual has learned, not with how rapidly or facilely he can use it. This aspect will therefore be measured by a "power" test, with a liberal time-limit. Many types of test items exist for testing knowledge of structure. The items would be chosen so that each one focuses on a single "structure point," but there would be a sufficient sampling of such structure-points, with a range from some of the more frequent of them to some of the less frequent, to yield a reliable score and a suitable score distribution. Specification of level of knowledge would if possible be based upon the level of difficulty reached. We will next list:

2. *Knowledge of general-usage lexicon* (vocabulary and "idiomatic phrases"). Here again, *knowledge* would be emphasized, rather than speed, and the test would be composed in such a way as to give reliability and a wide-ranged score distribution. If possible, specification of level of knowledge could depend upon level of word rarity attained, rarity being measured with reference to such compilations as the Thorndike frequency tables.

It might be desirable, however, to draw up separate specifications for vocabulary knowledge in certain semantic areas. Thus, let us list:

2a. *Knowledge of lexicon in designated specialized areas*. The subtests of the Michigan Vocabulary Profile Test, a commercially available vocabulary test, might be used as a guide. (These are: human relations, commerce, government, physical sciences, biological sciences, mathematics, fine arts, and sports. Obviously not all of these categories would be suitable for drawing up specifications for foreign students' vocabulary knowledge.)

Only in matters of phonology and orthography does it make sense to use the separate cells under "phonology or orthography" in our grid; the resulting categories may be labeled as follows:

3. *Auditory discrimination* (of phonemes, allophones, and suprasegmentals)

4. *Oral production* (of phonemes, allophones, and suprasegmentals)

5. *Reading* (in the sense of converting printed symbols to sound, i.e., mastery of word pronunciation and stress patterns)

6. *Writing* (in the sense of converting sounds to printed symbols, i.e., spelling)

The four skills of listening, speaking, reading, and writing must also be regarded as integrated performances which call upon the candidate's mastery of the

language as a whole, i.e., its phonology, structure, and lexicon. It is worthwhile to specify the level of competence desired in each of them, independently of essential "language fact" mastery, because each involves elements of quickness of response. In fact, the specification could well be partly in terms of rates, i.e., rate at which material of some set standard of difficulty could be heard and understood, rate of speaking in a standard interview situation, speed of silent reading attained under conditions where comprehension was to be tested, and speed of written composition. Hence we have:

7. *Rate and accuracy of listening comprehension*
8. *Rate and quality of speaking, as in an interview situation*
9. *Rate and accuracy of reading comprehension*
10. *Rate and accuracy of written composition*

The work of Lado and other language testing specialists has correctly pointed to the desirability of testing for very specific items of language knowledge and skill judiciously sampled from the usually enormous pool of possible items. This makes for highly reliable and valid testing. It is the type of approach which is needed and recommended in the first six categories of language proficiency specification listed above—that is, where knowledge of structure and lexicon, auditory discrimination and oral production of sounds, and reading and writing of individual symbols and words are to be tested. I do not think, however, that language testing (or the specification of language proficiency) is complete without the use of the approach recommended for categories 7 through 10, that is, an approach requiring an integrated, facile performance on the part of the examinee. It is conceivable that knowledge could exist without facility. If we limit ourselves to testing only one point at a time, more time is ordinarily allowed for reflection than would occur in a normal communication situation, no matter how rapidly the discrete items are presented. For this reason I recommend tests in which there is less attention paid to specific structure points or lexicon than to the total communicative effect of an utterance. For example, I have had excellent success in ascertaining levels of audio-lingual training by a listening comprehension test in which auditorily-presented sentences of increasing length and rapidity are to be matched with an appropriate picture out of four presented. The examinee is not concerned with specific structure-points or lexicon, but with the total meaning of the sentence, however he is able to grasp it.

Indeed, this "integrative" approach has several advantages over the "discrete structure point" approach. It entails a broader and more diffuse sampling over the total field of linguistic items and thus depends less upon the specifics of a particular course of training. It thus may lend itself somewhat more effectively to the problem of an external examination in which the examiner does not ordinarily know, in detail, what was covered in any particular course of training. Furthermore, the difficulty of a task is subjectively more obvious than in the case of a "discrete structure point" item. Thus, when the tasks of an "integrative"

approach test are arranged in the order of their difficulty for a typical class of examinees, the interpretation of performance relative to a subjective standard may be easier. Finally, the "integrative" approach makes less necessary the kind of comparison of language systems upon which much current language test is premised. The important question is to ascertain how well the examinee is functioning in the target language, regardless of what his native language happens to be. Indeed, the overconscientious use of the bilingual comparison axiom could lead to different tests of English proficiency for each native language group, and a virtual impossibility of establishing common standards of English language proficiency across native language groups.

I should like to summarize the major point of this paper up to this point. It is this: that an ideal English language proficiency test should make it possible to differentiate, to the greatest possible extent, levels of performance in those dimensions of performance which are relevant to the kinds of situations in which the examinees will find themselves after being selected on the basis of the test. The validity of the test can be established not solely on the basis of whether it appears to involve a good sample of the English language but more on the basis of whether it predicts success in the learning tasks and social situations to which the examinees will be exposed. I have attempted to suggest what might be the relevant dimensions of test performance, although I have not attempted to link them with collegiate learning and social situations—that is a task for college foreign student advisers and others who are familiar with the matrix of foreign student experiences.

Having made my major point, I wish to devote a little space to certain subsidiary issues which I consider important to testing.

THE CONTROL OF EXTRANEOUS VARIABLES OR INFLUENCES

In some ways, a good test is like an experiment, in the sense that it must eliminate or at least keep constant all extraneous sources of variation. We want our tests to reflect only the particular kind of variation in knowledge or skill that we are interested in at the moment. Suppose that we have an auditory comprehension test in which the examinee hears a question, and then has to select one of four *printed* answers. Is this a test of auditory comprehension? Not necessarily: it could be a test of reading skill for an individual who understands the question but cannot read the answer. Again, suppose that we have a written test of grammar: the printed instructions tell the examinee to "convert" a positive statement to a negative question. Is it a test of grammar? Not necessarily: It can be a test of vocabulary for the individual who knows the grammatical facts but who does not know such words as "convert," "positive," and "negative." Or suppose we have an oral production test in which the individual must make his response into a microphone: for many examinees, this might be a test of experience with a microphone, or even a personality test. Some of these

examples are familiar and obvious; nevertheless, if one scrutinizes the tests currently being used for assessing English proficiency abroad, one finds many instances where the test response may depend, for many people, on some quite incidental fact or circumstance, such as failure to understand instructions through lack of a sufficient number of sample items. It is true, of course, that one cannot control *all* extraneous sources of variation, and there are situations in which it is actually desirable or necessary to test two or more things simultaneously. Nevertheless, the problem needs constant attention from test constructors. Otherwise we tend too often to put our examinees in double jeopardy or multiple jeopardy, and to reduce our chances of making useful diagnoses of learning difficulties of areas of ignorance on the part of the examinees.

THE PROBLEM OF SAMPLING

If one happens to be dealing with a test of some very well circumscribed area of knowledge or skill, such as knowledge of the order of an alphabet, one can test for the presence of the total range of that knowledge. In most cases, however, the items that can be tested in any reasonable amount of time constitute only a small sample of all those that might be tested. It is important to be conscious of this fact, and to define as carefully as possible the total area from which one is sampling. Ideally, one should have a list of all the possible items which one might cover, and draw a sample by random sampling techniques, but this is rarely possible. Some makeshift approaches are available, however, and should be used whenever feasible. One of these, for example, is to sample from actual texts—e.g., to construct items deliberately on the grammatical structures found at the top of every fifth page of a text. Ingenuity will suggest other methods to help avoid the biases to which one is subject if the materials of a test are gathered solely on the basis of one's free associations.

THE DISJUNCTIVE FALLACY

This is the fallacy that one must do either one thing, or the other. Some writings on testing seem to assume that because a certain technique, e.g., the objective test time, will not do what one desires it to do at a given moment, e.g., test skill in constructing sentences, it is totally to be abandoned for all purposes. Different objectives may require different techniques, and there is no reason why a variety of techniques should not be used in one examination.

THE PROBLEM OF SCORE INTERPRETATION

Nothing is more frustrating in the area of testing than to be given a test score, even in percentage or percentile terms, without a ready means of interpreting this score in terms of some immediately practical consequence. Unfortunately, such scores are the customary way of reporting performance on objective tests,

and considerable effort is required to establish proper interpretations. *Percentage* scores are frequently misleading because their interpretation depends upon the intrinsic difficulty levels of the items on which they are based; *percentile* scores cannot be properly interpreted unless one is thoroughly familiar with the nature of the norming or comparison group. Contemporary test makers are turning to two types of score interpretation. One is to make use of the fact that test tasks can be arranged in difficulty level, and to report the difficulty level which the examinee can pass some set percentage of the time, say 75 percent. Thus, one might report that a given examinee has a vocabulary knowledge such that he knows 75 percent of the words near the rank of 10,000 in order of frequency, and one might then proceed to list sample words at this difficulty level. The other is to make greater use of what have been called expectancy tables. These are tables showing, for any given score level, the chances that an examinee has of succeeding in a variety of possible future courses of action. For example, the table might show that the chances are only three out of ten that an examinee with a given score level would be able to pass the freshman year in a college. Obviously, both types of score interpretation require considerable research in order to establish them on a firm basis, but even so, careful consideration should be given to these possibilities in developing a worldwide testing program in English.