DOCUMENT RESUME

ED 079 432 TM 003 037

AUTHOR	Carroll, John B.
TITLE	Foreign Language Testing: Will the Persistent Problems Persist?
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	Jun 73
NOTE	20p.; Paper presented at the ATESOL Conference (Dublin, Ireland, June 26-29, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Feedback; Language Instruction; Language Tests; *Second Language Learning; Test Construction; *Testing; *Testing Problems; *Test Validity

ABSTRACT

The persistent problems in foreign language testing are considered under four headings: (1) validity, (2) scope, (3) efficiency, and (4) the problem of how tests relate to the wider context of instruction. The first consists of insuring that the measurements and assessments obtained reflect what they are intended to reflect. The problem of scope consists of insuring that all the varied components of foreign language competence and skill are measured. The problem of efficiency encompasses the obtaining of the best assessments possible within the limits of time and resources available for the construction and administration of the assessments. An example of how test relate to the wider context of instruction is the degree to which testing either enhances instruction or distorts it through undesirable feedback effects from the tests. It is concluded that, with continual attention to the criteria for good test construction and to the need for new research on testing procedures, it should be possible to effect a gradual net improvement in the quality and effectiveness of foreign language tests. (Author/CK)

FOREIGN LANGUAGE TESTING: WILL THE PERSISTENT PROBLEMS PERSIST?

John B. Carroll

Educational Testing Service

Princeton, New Jersey, USA

[For presentation at the ATESOL Conference, Dublin, Ireland, June 26-29, 1973]

If we take a historical view of the development of foreign language testing, we can conclude that the present "state of the art" is demonstrably superior to that of, say, 1929, when one of the volumes commissioned by the American and Canadian Committees on Modern Languages (Henmon, 1929) considered methods of testing achievement and proficiency in foreign languages. At that time, objective testing methods had only recently been introduced in the foreign language field, and they were applied in a rather primitive fashion as judged by present-day standards. Most of the tests that became available were of the pencil-and-paper variety and tested only skills with the written language. Up to the time of World War II there was little experience with tests of the spoken language. It was not until Word War II that test makers (mostly in defense establishments) made serious efforts to develop comprehensive tests of foreign language achievement and proficiency, and the fruits of their labors did not have an influence in civilian circles until after the war. At least in the United States, large-scale efforts toward the development of standardized tests did not take place until the era of the NDEA Title VI funding activities, resulting, for example, in the development of several series of foreign language tests sponsored by the Modern Language Association of America--the MLA Cooperative Tests and the MLA Proficiency Tests for Teachers and Advanced Students.

The development of tests in English as a Foreign Language was spearheaded by the efforts of such people as Charles C. Fries and David Harris in the early 1950's, but the formats and procedures of such tests did not become perfected until the 1960's. Even now the so-called TOEFL test (Test of English as a Foreign Language) produced by ETS is under continual research scrutiny.

Without going further into the details of the history of foreign language testing in the U.S. and abroad, one can say that there has been an enormous increase in the sophistication with which foreign language tests are constructed and used. Gradually, throughout the world, there has developed an increased concern for the preparation of carefully designed measurement procedures in the foreign language field. I do not mean to give the impression that a "carefully designed" meacurement procedure is necessarily an objective test, for a measurement procedure can be characterized as "carefully designed" as long as there is attention to the measurement properties of the test--its reliability, validity, norming, standardization, and so on. It could well be a subjectively graded test and still meet these requirements. As we are now fully aware, some aspects of language proficiency cannot be measured by solely objective techniques. This is generally true of the more active and productive aspects of competence--speaking, writing, pronunciation, and fluency.

As against the purely paper-and-pencil tests that were introduced in the 1920's and 1930's, we now have a full range of tests purporting to measure not only competence and skill in written language but also competence and skill in spoken language, at various levels of ability and in various languages. We also have several well-validated tests of foreign language aptitude--not only for speakers of English but also for speakers of a few other languages. We are continually exploring the usefulness of new kinds of test formats--such as the "cloze procedure."

These developments have been stimulated by a number of factors. First, student enrollments in foreign languages have increased markedly, making it desirable and in fact necessary to increas the efficiency of testing procedures. Secondly, the foreign language profession has been faced with the need to develop more accurate and comprehensive measures because such measurements have become crucial in many important educational and real-life situations--for the selection, placement, and guidance of students and for the evaluation of competence of all those who use or are likely to use foreign language skills in their work. When people's wages or salaries depend, at least in part, on the level of attainment they can show on a foreign language test, it behooves test makers to provide valid and accurate tests and assessment procedures. Thirdly, there have been marked changes in language teaching, with a shift of emphasis towards the real-life language skills, and it has been a challenge to test makers to provide measurements of such skills, not only of the traditional reading, writing, listening and speaking skills, but also tests of cultural knowledge, "communicative competence," and the like. Fourth, there have been many scientific developments bearing on foreign language testing: developments in the theory and practice of educational and psychological measurement, in scientific linguistics, in the psychology of language (psycholinguistics), and in educational research. Fifthly, funds and resources for the development of tests and measurements have become more plentiful and available than previously. Finally, with the preparation of a number of excellent treatises on the techniques of test construction, such as those of Lado (1961), Valette (1967), Davies (1968), and Clark (1972), we have many more trained specialists in this field.

Yet, despite the relatively advanced "state of the art" that I have described, it cannot be said that foreign language testing is a completely perfected art. There remains much to be done in the way of research, the development of a wider variety of tests, and the training of teachers and others concerned with the construction and use of tests. But even if all the research and development that we can now foresee were to take place, I would claim that there are certain persistent problems that will probably never be resolved. It is simply in the nature of things that we will never be able to satisfy all our ideals and requirements in the field of foreign language testing. I want to spell out why I believe this to be so.

The persistent problems in foreign language testing may be considered under four headings, which I will first identify before entering into a detailed discussion of them:

(1) The problem of validity--that is, making sure that the measurements and assessments we obtain reflect what we want them to reflect.

(2) The problem of scope--that is, making sure that we measure or assess all the varied components of foreign language competence and skill.

(3) The problem of efficiency--that is, obtaining the best assessments we can obtain within the limits of time and resources available for the construction and administration of the assessments.

(4) The problem of how tests relate to the wider context of instruction--for example, the degree to which testing either enhances instruction or, contrariwise, distorts it through undesirable feedback effects from the tests.

Validity and Realism

Under this heading I include problems of reliability, or accuracy of measurement, but in fact, the problem of reliability is never as severe as the prouiem of validity, because there are well known procedures for increasing the reliability of measurements to as close to perfection as we desire. Reliability is thus merely a technical problem, but validity, or relevance, or realism--whatever we may call it, is in the last analysis a judgmental problem. No matter how high the reliability of an assessment or score, we always wonder whether that assessment reflects what we want it to reflect. There is no purely "scientific" way to find out. We shall always have to depend upon whatever logic or reason we can muster to help us to decide.

Take what might appear to be a simple case: appraising the validity of a "reading comprehension" test in English as a second language, a test, let us suppose, that consists of a number of paragraphs that the examinee is to read, with accompanying multiple-choice questions based on these paragraphs. The test purports, we are told by its makers, to measure the extent to which the student can read and understand English with a sufficiently deep comprehension to allow him to profit from books, newspapers, and other printed materials that he might encounter in his university studies. Let us further suppose that from a technical standpoint, the test is about as well designed and well researched as one might wish--that the internal consistency reliability coefficient is .96 as determined in a representative sample of examinees, and that every item shows a significant and reasonable correlation with the total test score. The scores in a representative sample of examinees are distributed approximately like a normal distribution. But does this test measure "reading comprehension"? Does it reflect the student's ability to read and understand English?

If the test is "valid," high scorers will have the desired ability, and low scorers will be found not to have that ability. But how can we know this?

There are, of course, certain procedures that might suggest an answer. We can "validate" the test against performance in the university. But even if the correlation with university performance is high, this will not demonstrate that the test measures our desired "reading comprehension ability," for it may be measuring some general intellectual factor, quite apart from reading comprehension ability, that manifests itself both in our test and in university performance. Or if the correlation is low, this will not demonstrate that the test does not measure reading comprehension, because university performance calls upon many more factors and abilities than reading comprehension ability. It will only demonstrate that what is measured by the test and what is required for successful university performance are largely independent. The conclusion we will have to draw is that "validating" a test against external criteria is not a sure guide to whether it is measuring what we want it to measure, although the accumulation of data may indeed give clues as to what it is measuring.

The problem is somewhat like that of accepting and rejecting scientific hypotheses. Just as we can often find sufficient evidence to reject a hypothesis, but can never find sufficient evidence to accept a hypothesis beyond doubt, we can never accumulate sufficient evidence to prove that a test is valid beyond doubt, although we can often find evidence to indicate that a test is invalid and that it does not reflect what we want it to reflect.

We could, for example, find that our reading comprehension test is too much dependent upon knowledge of vocabulary--that is, that students can get high scores on the test by knowing a great many words but without really understanding the meaning of the passages given on the test. We might do this, for example, by discovering that the questions could be answered equally well if we were to delete all but the content words in the passages (eliminating all clues to grammatical structure).

Or we might find that our reading comprehension test is more a test of logical inference. This could be the conclusion if we were to find that students scored just as well on the questions even without being able to read the passages on which the questions were based.

Or we might find that the "reading comprehension" test is for many students a test of "test wiseness," i.e., the ability to indicate answers with the paraphernalia of separate answer sheets that many tests employ. We might draw this conclusion if we were to find that scores improved markedly after giving students training in the item format and the use of the answer sheet.

These are merely examples of some possible sources of *invalidity*, but even after eliminating these possibilities, we would still not know whether the test is truly a test of reading comprehension as we conceive it. Of course, we should always consider possible sources of invalidity, and attempt to eliminate them, but my point is that even after eliminating them we are not in a position to conclude that the test is valid beyond doubt. However, we are generally in at least a *better* position to draw this conclusion than we would be if we had not considered the possible sources of invalidity.

In the end, the only possibility of deciding on the validity of the test would come from an analysis of what competences, skills, and discriminations are *essential*, necessary, and sufficient for successful performance on the test, and similarly, what *lacks* of competence, skill, and disrimination are, if you will, necessary and sufficient for poor performance on the test.

Experts in measurement will recognize, perhaps, that I am trying to convey in simple terms what we mean by the *construct* validity of a test. In the case I have been describing, the issue is the validity of the test as a measure of the construct of "reading comprehension." This construct is itself very difficult to define, but I will attempt to define it as the ability on the part of the student to capture, from a printed text, all the essential meanings that have been put there by the author of the text, as represented by the lexical, grammatical, and rhetorical devices used by the author.

Even this construct is not adequately defined, for one should add at least a footnote concerning the relative difficulty and complexity of the text material under consideration. A student might be able to understand a simple narrative but not understand a highly reasoned argument.

Even with the addition of remarks about the difficulty and complexity of the textual material, the construct of "reading comprehension" is of course only one particular aspect of foreign language competence. I have been using it only as an example. Similar constructs could be defined for other aspects of foreign language competence, such as listening comprehension, pronunciation accuracy, compositional skill, and so forth. Judging the validity of tests of those competencies would require the same kind of searching examination and analysis as for the reading comprehension test.

One of the most difficult constructs to measure is what Jakobovits (1970) has called "communicative competence," the ability to use the foreign language in real-life situations with appropriate selections of registers, and native-speaker-like intuitions about contextual meanings. One might imagine--and Jakobovits has actually proposed this--the construction of a series of situational tests where the student would have to perform as a producer and receiver of language, as in buying a list of items in a grocery store or native market, obtaining information about routes for getting to a specified location, or discussing matters concerning international trade balances. Within limits, such tests might be successful, but I can see many problems with them. Successful performance in such situations will often depend upon many factors other than language competence, and it would be necessary to sample a rather wide variety of situations in order to make anything like a complete test of the individual's "communicative competence." Also, such tests would be expensive and rather impractical to administer on any wide scale. I myself feel that it will be more efficient and useful in the long run to obtain good assessments of basic language competences, by more traditional means, for I think that as far as the foreign language teacher is concerned, the teacher is responsible only for developing in the student those basic competences on which successful use of language depends. The FL teacher cannot be held responsible if a student possessed of all the basic language competences fails to use them in practical situations.

I am similarly conservative about the cry that we need to replace our present tests with "criterion-referenced" tests. The notion of criterion-referencing of test results is a good one, and in fact I have urged that tests be criterion-referenced as much as possible. But what needs to be better understood is that "criterion-referencing" does not necessarily imply a radical departure in the types of tests that are employed. It has more to do with how test results are interpreted and used than with how the test itself is constructed. What exactly is "criterion referencing"? It means interpreting the test score with reference to the kinds of language behaviors and competencies that are implied by that test score--rather than with reference to a comparison with the scores of others in a norm group. Thus, we could say that a test is "criterion-referenced" if it is possible to say that such-and-such a score means that, for example, the individual can read material of such-and-such difficulty, or can converse fluently about such-and-such kinds of topics. In fact, we can take a conventional "norm-referenced" test and through certain research operations find out what the scores on that test mean in behavioral terms, thus converting it into a "criterion-referenced" test. I have done this for some of the MLA Advanced Proficiency tests (Carroll, 1967) by relating their scores to performance on an "absolute" scale of proficiency as provided by an interview procedure. On the other hand, I must admit that in order for a test to be amenable to criterion-referencing, it must contain materials that will make it possible to refer it to meaningful criterion behaviors. But most well-constructed foreign-language proficiency tests do contain enough material relevant to language competencies for this purpose.

The Problem of Scope

By the problem of "scope" I refer to the question of whether foreign language tests can provide a properly comprehensive measurement of the various skills and competencies that are embodied in the concept of "full foreign language competence," and whether it is possible to differentiate those skills from one another in order to provide profiles of relative attainments in different skills.

In recent work on the comprehensive measurement of language skills, we have the paradox that the more we attempt to measure different language skills, and the better our measurements of those skills, the higher the correlations among the skills, and thus the more they appear to converge toward the measurem ιt of a single all-embracing skil_. If we start with a good test of reading comprehension skill, and then add measurements of listening comprehension, of speaking skill, and of writing ability, we often find that these skills tend to correlate highly, especially if we adjust for unreliability through statistical techniques. What does this mean? It is tempting to conclude that there is indeed only one basic foreign language skill--that we can epitomize as simply "knowledge of the structure and lexicon of the language." But perhaps it is incorrect to draw this conclusion.

In the first place, we find these high correlations generally only where the instruction itself has been broad and comprehensive, covering all these skills adequately. Correlations among different tests are subject to variation depending upon the types of instructional treatments. For example, we could expect a low correlation between reading and listening comprehension tests if instruction focused on reading skills and gave little attention to speaking and listening skills; such a pattern of instruction could produce a situation where there is no significant and reliable variance in the listening comprehension test, with the consequence that there could be no reliable covariance with a reading comprehension test.

In the second place, high correlations among various tests of skill should not necessarily preclude the separate consideration and profiling of those skills. Even with a high correlation between reading and listening tests, for example, it is still possible to look at the relative levels of proficiency on these tests: frequently we find that the average level of proficiency on a reading test is high, while the average level of proficiency on a listening test is relatively low.

Furthermore, we must remember that "reading comprehension" and "listening comprehension" are what I have called "integrated" language skills (Carroll, 1968), because they depend (or should depend) on a wide variety of detailed competences in particular aspects of the language--its phonology, spelling, grammar, lexicon, and so forth. As I have already indicated, a score on a reading comprehension test may mask wide variations in the student's abilities in these various aspects--he may be good in his knowledge of vocabulary but deficient in knowledge of grammar. I think it is unfortunate that present-day tests of language proficiency have retreated somewhat from the older concept of separately measuring knowledge of grammar and knowledge of vocabulary. Language skills can be broken down even further: in the realm of vocabulary we could separately measure vocabulary knowledge in different domains--such as general, literary, and scientific. In the realm of grammar we could separately measure knowledge of morphology, of major sentence types, of noun compounding, etc. Even though such measures might still be rather highly correlated, they would give more diagnostic information on which to base instructional planning.

Recently there has been much interest in the use of the so-called "cloze technique" in foreign language tests. This is a technique, you may already know, where, for example, every fifth word in a passage is deleted and replaced by a standard-sized blank; the student is asked to try to guess what word has been deleted, either by filling it in--in a "free response" mode, or by selecting it from a number of choices--in the "multiple choice mode." Let me point out that the cloze technique measures an "integrated language performance," rather than separate competences in vocabulary, grammar and so on. Among the blanks to be filled in on a "clt.. passage, some of the answers will depend upon a knowledge of grammar, others will depend mostly on a good knowledge of vocabulary, and still others will depend .argely upon careful attention to the total meaning of the passage. Although the "cloze" technique may be an effective way of measuring certain integrated language performances, it will not easily lend itself to diagnostic interpretation.

Adequate profiling of different language skills will require not only careful differentiation of those skills, but also control of the ease and difficulty of the test materials. It is necessary to provide a range of difficulty in each of the skills so that the exact level of the individual's competence can be appraised. Ease or difficulty should be recognized as dependent upon one or more of the following:

(a) The inherent ease or difficulty of the material itself, in terms of linguistic simplicity or complexity, or ease of learning.

(b) The relative frequency or rarity of the linguistic item under consideration, e.g., a particular word or grammatical construction. Other things being equal, a word or grammatical construction that occurs rarely will show itself to be more difficult than one that occurs frequently.

(c) The amount of emphasis that has been given to the item in the instruction. Other things being equal, items not emphasized in instruction will show themselves to be more "difficult," from a test construction standpoint, than items that have been emphasized and given much attention in the instruction.

Adequate profiling of language skills will also depend upon a broader conception of language skills than is represented by most presently-available tests. Those tests are primarily measures of language knowledge--by inference, measures of "competence," but they fail to consider adequately such performance factors as speed and fluency. The giving of a reading comprehension test under a time-limit yields a score that represents some unknown combination of competence in the language and reading speed. Listening comprehension tests do not ordinarily take advantage of the possibility of varying the rates at which texts are spoken and thus determining the rate at which the student can comprehend them.

Tests of language production could be improved by giving more thought to performance components such as the ability to retrieve lexical items quickly, the ability to manipulate syntactic elements flexibly, and the ability to encode meanings effectively. Recent work in psycholinguistics contains many examples of linguistic tasks that might well be adapted to use in foreign language tests.

It is along these lines that I believe foreign language tests could be further improved in order to provide a more comprehensive picture of language skills. Although the problem of comprehensiveness will probably never be adequately resolved, we can at least put our best efforts into it.

The Problem of Efficiency

The problem of efficiency has several related aspects: first, efficiency in test construction activities; second, efficiency in the actual use of testing time; and third, efficiency in the scoring or marking of tests and the interpretation of test results. By "efficiency," I mean to include factors of cost in terms of time, money, and resources.

In many situations throughout the world, there are demands for the construction of foreign language tests in multiple alternate forms and for repeated administrations. Because of security factors, tests get very rapidly "consumed"; a test once used is sometimes practically unusable for another occasion. Under these conditions it is difficult to devote the care to test construction that would be ideal. There seem to be practical limitations on the number of items or test tasks that can be constructed. Increasingly, testing organizations have had to employ large staffs of test constructors, and they are developing large pools or "banks" of pretested and validated items from which new tests can be constructed. This appears to be the only way in which they can meet the demands for new tests. I myself can give no suggestion as to any other way of operating. The demands for the construction of ever "new" tests will constitute one of the persistent problems in the foreign language testing field. My only advice is that it should be recognized as such, and adequate means should be mobilized to meet it head on.

Efficiency in the actual use of testing time and testing facilities will also continue to be a persistent problem. In order to obtain more reliable, valid, and comprehensive measures of foreign language achievements and proficiencies, increasing amounts of testing time will be needed, often with the use of special facilities such as tape recorders and other electromechanical equipment. Some have even proposed that tests should be administered at special computer consoles or terminals, but so far, cost factors have generally precluded such developments. Even without these developments, efficient use of the present technology of test administration--with separate answer sheets, specially printed test forms, tape cassettes, etc.--requires much care and attention. Test administrators need to be carefully trained, and test locations need to be carefully selected from the standpoint of comfort, lighting, acoustic conditions, and so forth. Test constructors need to take account of these logistic factors by constructing the tests so that any adverse effects can be minimized. (For example, they need to use care in the writing of test instructions both to students and to test administrators, and in selecting types of test formats that will be least likely to result in difficulties in test administration.)

Finally, there is the matter of efficiency in the scoring, grading, or marking of the completed tests. Large testing organizations like ETS have had to develop advanced technologies for this--involving the use of electronic equipment to score tests. But these technologies are in the main developed only for tests of an objective type where the responses can indeed be scored by machine. For tests in which subjective grading of the quality of responses must be employed, as for example in the evaluation of speaking and writing tests, it has been necessary to develop staffs of qualified graders and to maintain controls on the standards of accuracy and quality in grading. Again, achieving efficiency in grading test results will remain as a persistent problem in the foreign language field as in other fields; in fact, the problems are particularly acute in the foreign language field because of the great importance of subjectively scored tests in this field.

I have been speaking, of course, of wide-scale administration of external examinations and proficiency tests such as the TOEFL test. But on a smaller scale, some of the same problems exist in the case of teacher-made tests to be given in the classroom. In order to keep up with the demands, teachers need to maintain their own "item banks," and they need to be aware of whatever technologies of test construction and administration can be applied in their situations.

Testing in the Context of Instruction

No matter how much we may improve instruction, the kinds of tests that are used to evaluate students can often have adverse effects on students' learning. This is particularly true of external examinations, but it may also be true for teacher-made tests. It is only natural for students to shape their learning efforts so as to be maximally successful on tests, and if the tests measure objectives that are in some ways different from those of the instruction, students will work towards those objectives and pay less attention to achieving other objectives. The nature of external examinations will often shape the behavior of the teachers themselves. We sometimes complain that teachers do nothing but "teach for the tests."

In Japan, for example, it has been observed that the important tests for university entrance tend to be of an excessibly formalistic character, emphasizing certain reading and writing skills to the exclusion of speaking and listening skills. Often they demand primarily an ability to comment on the literary qualities of English classics, or to gloss highly specialized forms of literary expression. Small wonder that Japanese students preparing for the university pay little attention to the acquisition of speaking and listening skills. Teachers of English in Japanese secondary schools make a point of teaching so that their students will have the best chance of succeeding on these examinations.

Of course, this kind of situation is by no means unique to Japan; similar situations are to be observed in countries throughout the world, wherever those who prepare external examinations are for one reason or another unable or unwilling to make their examinations cover all the objectives and skills that should be covered if the examinations are to be sufficiently comprehensive.

This matter of the relation between instruction and examining is certainly one of the persistent problems of foreign language testing. To some extent it is a political problem, in the sense that the power structures underlying the creation of external examinations are distant from, and relatively inaccessible to, those who are doing the instruction at the local level. Bodies of external examiners are often excessively traditional and conservative in their outlook, and insensitive to the trends taking place in foreign language teaching and testing. But it is also true that factors of logistics and cost impede the creation of realistic and valid external examinations. For example, the administration of adequate tests of listening and speaking skills may appear to be unfeasible from an administrative point of view, because such tests may require special equipment, special testing conditions, and special procedures for grading the results. One suspects, however, that the inadequacies of external examinations are often due simply to lethargy and ignorance on the part of examining bodies. Examining authorities do not understand that the creation of adequate examinations requires considerable outlays for technically trained staff personnel and for all the technological aids for test development and administration--such as computers, test scoring machines, tape recording equipment, etc. An adequate technology for the creation of highly valid and comprehensive tests is now available, but that technology is not being adequately utilized.

If examinations can have adverse effects on instructional efforts, they can also have beneficial effects. The solution for the problem of "teaching for the tests" is to make better tests. The salutary effects of more comprehensive external examinations were well demonstrated when in the U.S. the College Entrance Examination Board was induced, by pressure from the foreign language teaching profession, to include tests of listening skills in its foreign language examinations. The introduction of these tests provided much more motivation for both teachers and students to orient their efforts toward greater attention to the spoken foreign language. And while we do not regard the TOEFL test (Test of English as a Foreign Language) as anything near perfect, the fact that it does include tests of listening skills has surely had at least some beneficial effects on the nature of TEFL instruction throughout the world.

Conclusion

Let me summarize my remarks by reminding you that I have outlined four aspects of foreign language testing in which there have been persistent and often frustrating problems. Neither I nor anybody else has final solutions for these problems; these problems will continue to exist and to challenge our best efforts. Like keeping a house in order, the job of developing adequate foreign language tests will never be done, because it has to be done again and again as new generations of students come into our courses and as the requirements and objectives of foreign language instruction change. I have, however, tried to suggest some ways in which that job can be done better and more efficiently with each new undertaking. With continual attention to the criteria for good test construction and to the need for new research on testing procedures, it should be possible to effect a gradual net improvement in the quality and effectiveness of foreign language tests.

* * * * *

References

Carroll, John B. *The foreign language attainments of language majors in the senior year: A survey conducted in U.S. colleges and universities.* Cambridge, Mass.: Harvard Graduate School of Education, 1967. (ERIC Document ED 013-343.)

Carroll, John B. The psychology of language testing. In Alan Davies (Ed.), *Language testing symposium: A psycholinguistic approach.* London: Oxford University Press, 1968. Chapter 4, pp. 46-69.

Clark, John L. D. *Foreign-language testing: Theory and practice*. Philadelphia: The Center for Curriculum Development, 1972.

Davies, Alan (Ed.) *Language testing symposium: A psycholinguistic approach.* London: Oxford University Press, 1968.

Henmon, V. A. C. *Achievement tests in the modern foreign languages*. New York: Macmillan, 1929.

Jakobovits, Leon A. *Foreign language learning: A psycholinguistic analysis of the issues*. Rowley, Mass.: Newbury House, 1970.

Lado, Robert. *Language testing*. London: Longmans, Green, 1961. (Reprinted by McGraw-Hill, New York, 1964.)

Valette, Rebecca M. *Modern language testing: A handbook*. New York: Harcourt, Brace, Jovanovich, 1967.