TESTING ENGLISH as a Second Language

David P. Harris

Testing English as a Second Language

David P. Harris
Georgetown University

McGraw-Hill Book Company

New York • St. Louis • San Francisco
London • Toronto • Sydney • Mexico • Panama

Preface

Although there are now a number of very excellent textbooks on the methods of teaching English as a second language, we have lacked a short, concise text on the testing of ESL, a subject about which both classroom teachers and trainers of teachers have shown an increasing concern. It is hoped that this little book will help to meet the need by providing just about the right amount of material for the testing component of teacher-training courses and that at the same time it will prove useful for home study by teachers whose formal training slighted this important subject.

The twofold objective of the book is to enable the ESL teacher both to improve his own classroom measures and to make sound assessments of standardized tests which he may from time to time be asked to select, administer, and interpret. In the opening chapters he is introduced to the general purposes and methods of language testing and is asked to consider the chief characteristics of good educational measures. A series of six chapters then describes specific techniques for testing each of the major components of English as taught to speakers of other languages, after which attention is directed to the step-by-step process whereby the ESL test is constructed and administered, and the results interpreted. The final chapter offers procedures for calculating a few basic test statistics which will aid the teacher–test writer in evaluating the soundness of his tests and the performance of his students. As with the rest of the book, the final chapter assumes no previous training in tests and measurement and no knowledge of advanced mathematics.

In the preparation of this book, the writer drew from "sources too numerous to mention," though a small number are identified in the footnotes and in the list of selected references which appears in the back. In addition, he included material from two of his own earlier writings: the *English Testing Guidebook*, Parts I-II, prepared for the International Cooperation Administration in 1961; and the article "The Testing of Student Writing Ability," which appeared in *Reflections on High School English*, edited by Gary Tate and published by the University of Tulsa in 1966. The writer wishes to express his appreciation to Professor Tate and the University of Tulsa for permission to incorporate portions of this article in the present work.

The writer is deeply indebted to Dr. Edith Huddleston of the National Institute of Mental Health, Professor Betty W. Robinett of Ball State University, Professor Leslie A. Palmer of Georgetown University, and Mr. John Upshur of the English Language Institute, University of Michigan, for their careful reading of parts or all of the manuscript and for their extremely valuable comments and suggestions. For the material added during revision, the writer is, of course, entirely responsible.

David P. Harris

Contents

1 Purposes and Methods of Language Testing

TEACHER-MADE VERSUS STANDARDIZED TESTS

In any consideration of educational testing, a distinction must be drawn between the rather informal, teacher-made tests of the classroom and those formal, large-scale, "standardized" instruments which are prepared by professional testing services to assist institutions in the selection, placement, and evaluation of students.

Classroom tests are generally prepared, administered, and scored by one teacher. In this situation, test objectives can be based directly on course objectives, and test content derived from specific course content. Inasmuch as instructor, test writer, and evaluator are all the same individual, the students know pretty much what is expected of them—what is likely to be covered by the test questions and what kind of standards are likely to be applied in the scoring and the interpretation of results. And since the scoring will be done by only one person, the standards should remain *reasonably* consistent from paper to paper and test to test. Moreover, it is very likely that the teacher's ultimate evaluation of his students will be based on a number of tests and other measures, not just one. Therefore a single

bad test performance by a student need not do irreparable damage to his final standing, nor, probably, will one inadequate or ineptly constructed test prevent the teacher from making a reasonably sound final judgment. Finally, since the number of students to be tested is relatively small, the teacher is not limited to quickly scorable item types but may, if he wishes, make full use of compositions and short-answer techniques (see below).

Obviously, few if any of the above conditions apply to the standardized test, designed to be used with thousands and sometimes hundreds of thousands of subjects throughout the nation or the world, and prepared (and perhaps administered, scored, and interpreted) by a team of testing specialists with no personal knowledge of the examinees and no opportunity to check on the consistency of individual performances.

Even though this book has been designed primarily for the classroom teacher, we shall deal throughout with both types of testing. For although the teacher's primary testing concern will be in improving his own classroom measures, he will quite probably need at some time or other to make use of standardized tests, and it is therefore important that he know how to select and evaluate such instruments as well. And, in turn, learning more about the techniques and research findings of the professional testers will help the classroom teacher to write better tests himself.

THE PRINCIPAL EDUCATIONAL USES OF LANGUAGE TESTS

Before we can even begin to plan a language test, we must establish its *purpose* or *function*. Language tests have many uses in educational programs, and quite often the same test will be used for two or more related purposes. The following list summarizes the chief objectives of language testing; the categories are not by any means mutually exclusive, but they do indicate six different *emphases* in measuring student ability or potential.

1. To determine readiness for instructional programs. Some screening tests are used to separate those who are prepared for an academic or training program from those who are not. Such selection tests have a single cutoff point: examinees either "pass" or "fail" the test, and the degree of success or failure may not be deemed important.

2. **To classify or place individuals in appropriate language classes.** Other screening tests try to distinguish *degrees of proficiency* so that examinees may be assigned to specific sections or activities on the basis of their current level of competence. Such tests may make no pass-fail distinctions, since some kind of training is offered to everyone.

3. **To diagnose the individual's specific strengths and weaknesses.** Diagnostic screening tests generally consist of several short but reliable subtests measuring different language skills or components of a single broad skill. On the basis of the individual's performance on each subtest, we can plot a *performance profile* which will show his relative strength in the various areas tested.

4. **To measure aptitude for learning.** Still another kind of screening test is used to predict future performance. At the time of testing, the examinees may have little or no knowledge of the language to be studied, and the test is employed to assess their potential.

5. **To measure the extent of student achievement of the instructional goals.** Achievement tests are used to indicate group or individual progress toward the instructional objectives of a specific study or training program. Examples are progress tests and final examinations in a course of study.

6. **To evaluate the effectiveness of instruction.** Other achievement tests are used exclusively to assess the degree of success not of individuals but of the instructional program itself. Such tests are often used in research, when experimental and "control" classes are given the same educational goals but use different materials and techniques to achieve them.

For simplicity, the foregoing six categories can be grouped under three headings: *aptitude* (category 4 above), *general proficiency* (categories 1 to 3), and *achievement* (categories 5 and 6). These three general types of language tests may be defined in the following manner:

An aptitude test serves to indicate an individual's facility for acquiring specific skills and learnings.

A general proficiency test indicates what an individual is capable of doing now (as the result of his cumulative learning experiences), though it may also serve as a basis for predicting future attainment.

An achievement test indicates the extent to which an individual

has mastered the specific skills or body of information acquired in a formal learning situation.

Not all measurement specialists use this three-way division of tests or interpret the terms aptitude, proficiency, and achievement precisely as we have done above. Our three categories do, however, seem to lend themselves well to the classification of language tests and will be of value in helping us in succeeding chapters to differentiate among the principal testing objectives.

Actually, our concern in this book will be almost entirely with measures of proficiency and achievement. For although some successful attempts at developing general language aptitude tests have been made,[1] this area of testing is still relatively new, and no aptitude measures specifically for learners of English as a second language could be said to have passed the experimental stage. Valid English aptitude measures would be of inestimable value to both educational institutions and international-exchange agencies in this country, for if English-learning potential could be accepted as a substitute for current proficiency and achievement, it would then become economically feasible to admit non-English-speaking students to academic or technical-training programs that would include short-term, intensive English language components. Let it be hoped, therefore, that some of the current experimentation will soon bear fruit.

THE PRINCIPAL LANGUAGE-TESTING TECHNIQUES

Translation

Translation was formerly one of the most common teaching and testing devices, and it remains quite popular today in many parts of the world. However, with the spread of the new "linguistically oriented" methods of instruction and measurement, translation has lost much of its appeal in this country. In the first place, translation is in reality a very specialized and highly sophisticated activity, and one which neither develops nor demonstrates the basic skills of listening, speaking, reading, and writing. Indeed, the habit of

[1] Especially significant is John B. Carroll and Stanley M. Sapon's *Modern Language Aptitude Test,* Form A (New York: The Psychological Corporation, 1955-1958), a test designed for native speakers of English learning modern foreign languages. Adaptations in other languages are now in experimental use.

translating is now felt to *impede* the proper learning of a foreign language, for one of the first objectives in modern foreign-language instruction is to free the learner from native-language interference—to teach him to react in the target language without recourse to his mother tongue. To be sure, modern language departments often include advanced-level courses in translation, but here translation is treated as a creative activity which follows, and depends upon, fairly complete mastery of the target language. In an achievement test for translation courses, there of course would be very good reasons for having the examinees translate.

Secondly, translation is extremely difficult to evaluate. Is a "good" translation one that captures the tone and mood of the original by substituting the idiom of the second language, or is translation only "good" when it approaches a literal, word-for-word rendering of the original? The criteria and standards for judging translations depend so much on individual taste that the translation test tends to be a highly unreliable kind of measure, and particularly when large numbers of examinees require several scorers.

Dictation

Dictation is another testing device that retains some of its former popularity in certain areas. Dictation is undoubtedly a useful pedagogical device (if used in moderation) with beginning and low-intermediate-level learners of a foreign language, and the responses that such students make to dictations will certainly tell the teacher something about their phonological, grammatical, and lexical weaknesses. Other types of tests, however, provide much more complete and systematic diagnosis, and in far less time. As a testing device, then, dictation must be regarded as generally both uneconomical and imprecise.

Composition

A composition test allows the examinee to compose his own relatively free and extended written responses to problems set by the examiner. In foreign-language testing these responses may consist of single paragraphs or may be full essays in which the student is rated not only on his use of the grammatical structures and lexicon of the

target language but also on his ideas and their organization. Grades for such "free-response" tests may also take into account the examinee's employment of the graphic conventions—spelling, punctuation, capitalization, paragraphing, and even handwriting.

If composition tests are somewhat less frequently employed in foreign-language courses now than formerly—at least in this country—the principal reason is probably the growing popularity of the audio-lingual method of teaching, not the long-standing objections of the educational-measurement specialists. At least in advanced-level courses, such tests remain one of the favorite forms of measurement for the very understandable reasons that they are an easy type of test to construct and appear to measure certain high-level abilities better than do the objective techniques.

The chief difficulties in using and assessing compositions as a measurement device are (1) eliciting the specific language items that the test writer particularly wishes to test and (2) finding a way to evaluate these free responses reliably and economically.

Composition tests will be treated at some length in our chapter on the testing of writing, and therefore a detailed discussion of the pros and cons will be deferred to that chapter. It should be stated at the outset, however, that in view of recent research it no longer appears necessary to adopt an either-or approach to the subject: there are unquestionably many language-testing situations in which the use of free-response techniques is highly inefficient, just as there is a narrow range of measurement objectives that may best be attained through the use of carefully prepared and scored compositions.

Scored Interview

Roughly parallel to the composition as a measure of students' written language is the scored interview as a device for assessing oral competence. Both are classed as free-response tests in which the subjects are allowed to express their answers in their own words in a relatively unstructured testing situation. The chief differences between these two devices, in addition to the obvious one that compositions call for writing and interviews call for speaking, are that in interviews (1) the examiner must provide a large number of cues throughout the performance and (2) the evaluation is generally made during the actual production of the responses, and there is no

way for the examiner to reexamine the performance later in order to check the accuracy of his ratings.[2]

Most teachers who use the interview test do so not out of any strong conviction that it is the best of all possible techniques, but simply because they have no better way of assessing the oral competence of their students. Most of the weaknesses that we noted in our brief discussion of the composition apply to the interview as well. In our chapter on the testing of speaking, we shall deal at some length with this and alternative methods of measuring the oral abilities.

Multiple-choice Items

Multiple-choice or *selection* items types were developed to overcome a number of the weaknesses of the composition test that we noted earlier. Because of the highly structured nature of these items, the test writer can get directly at many of the specific skills and learnings he wishes to measure, and the examinee cannot evade difficult problems as he often can with compositions. As these items generally can be answered fairly rapidly, the test writer can include a large number of different tasks (that is, individual items) in the testing session. Finally, inasmuch as the examinee responds by choosing from several possible answers supplied by the test writer, scoring can be done quickly and involves no judgments as to degrees of correctness. Because of these virtues, multiple-choice tests tend to have superior *reliability* and *validity*, two important test characteristics which we shall examine in some detail in Chapter 2.

In its "classic" form, the multiple-choice item consists of (1) a *stem* or *lead*, which is either a direct question or an incomplete statement, and (2) two or more *choices* or *responses*, of which one is the *answer* and the others are *distracters*—that is, the incorrect responses.

To walk through water is to _____

A. wade C. crouch
B. scold D. shrug

[2] An obvious exception is the interview that is tape-recorded. In most interview situations, however, the use of tapes is impracticable or undesirable because of its effect on the examinees.

The stem of this item is "To walk through water is to ———." The choices are the words marked A, B, C, D. The answer is choice A; the other choices are the distracters.

The very form of the multiple-choice item is the source of the most common objection to this testing method: the examinee does not have to think of his own answers; he "merely" makes choices. In our chapter on the testing of writing we shall treat this criticism in detail, citing a few of the many studies that give evidence that ability to *choose* the best of a number of given alternatives is actually quite highly related to ability to *compose* correct responses.

A more genuine disadvantage of multiple-choice tests is the very considerable skill and time that are required to prepare them. In deciding between compositions and selection methods, therefore, the classroom teacher must consider whether he wishes to put most of his effort into the preparation or into the scoring of his test. Fortunately, in many testing situations there is the possibility of another alternative—the short-answer test, which is a kind of compromise between the composition and selection types.

Short-answer Items

Short-answer tests combine some of the virtues of both multiple-choice and composition tests: the problems are short and highly structured, yet they provide the examinee with the opportunity to compose his own answers. As commonly used in language testing, short-answer items require the examinee either to complete a sentence or to compose one of his own according to very specific directions.

Directions--Complete each sentence by writing an appropriate form of the verb that is given in parentheses.

I wish I _____ (have) a new car.

Directions--Rewrite each statement to make it a negative question.

John knew the answer to the problem.

_____?

Short-answer items are extremely useful in informal classroom testing: they are relatively quick and easy to write and they require much less scoring time than would a composition. Their disadvantages for large-scale testing are, first, that they take longer to score than the multiple-choice types—an important factor when large numbers of papers are involved—and, second, that quite frequently

there are a number of possible right answers, some of which the item writer might not even have considered at the time he prepared the test. Thus, in the first item given above, though the most likely completion would be "I wish I *had* a new car," we would have to accept *had had* as quite acceptable in certain contexts. And would such forms as *might have* and *could have had* be considered wrong? The problem of having to make such value judgments about the examinees' responses is avoided in the multiple-choice item types.

THE LANGUAGE SKILLS AND THEIR COMPONENTS

Language exists in two forms, the spoken and the written. Had we been treating this subject a generation ago, we would probably have put writing ahead of speaking. But the "new" language teaching methods introduced during and immediately following the Second World War have led us to change our order of priorities, and this present-day emphasis on the spoken form of the language is now reflected in our testing as well as our teaching of second languages.

Two linguistic activities are associated with both speech and writing: an encoding and a decoding process. *Speaking* and *writing* themselves are the encoding processes whereby we communicate our ideas, thoughts, or feelings through one or the other form of language; and *listening* and *reading* are the parallel decoding processes by which we "understand" either a spoken or a written message. We may therefore say that language includes four skills, or complexes of skills: listening, speaking, reading, and writing. It is perhaps in this order that we originally learned our native language, and it is in that order that foreign languages are now very frequently taught.

If we are correct in referring to the above as *complex* skills, what do we identify as the components of each? First of all, a moment's thought will establish two very important elements shared by all four skills: *grammatical structure* and *vocabulary*. In addition to these, skill in both auditory comprehension and oral production depends in part on mastery of the sound system of English. Thus we may list *phonology* as a third component of two of our four skills. And parallel to phonology in the spoken form of the language is *orthography* in the written form. For convenience we may wish to treat the sound and graphic systems together as an "either-or"