A Guide to LANGUAGE TESTING

Development

Evaluation

Research

GRANT HENNING

A GUIDE TO

Language Testing

DEVELOPMENT • EVALUATION • RESEARCH

Grant Henning

University of California, Los Angeles

NEWBURY HOUSE PUBLISHERS, Cambridge

A division of Harper & Row, Publishers, Inc.

New York, Philadelphia, San Francisco, Washington, D.C.
London, Mexico City, São Paulo, Singapore, Sydney

Library of Congress Cataloging-in-Publication Data

Henning, Grant.
A guide to language testing: development, evaluation, research.

Bibliography.
1. English language—Study and teaching—Foreign speakers. 2. English language—Ability testing.
I. Title.
PE1128.A2H45 1987 428'.076 86-23446
ISBN 0-06-632277-4

Production coordinator: Maeve A. Cullinane
Designer: Carson Design
Compositor: Publication Services, Inc.
Printer: McNaughton & Gunn, Inc.

NEWBURY HOUSE PUBLISHERS
A division of Harper & Row, Publishers, Inc.

Language Science
Language Teaching
Language Learning

CAMBRIDGE, MASSACHUSETTS

Printed in the U.S.A.
63 22770

First printing: February 1987
4 6 8 10 9 7 5 3

Table of Contents

Dedication

To my students at UCLA and American University in Cairo, who inspired, criticized, and encouraged this work, and who proved to be the ultimate test.

Preface

The present volume began as a synthesis of class notes for an introductory course in testing offered to graduate students of Teaching English as a Second/Foreign Language. Chapters one through seven formed the point of departure for a one-semester course, supplemented with popular tests, articles on current testing techniques, and student projects in item writing and item and test analysis. To address the advent of important new developments in measurement theory and practice, the work was expanded to include introductory information on item response theory, item banking, computer adaptive testing, and program evaluation. These current developments form the basis of the later material in the book, chapters eight through ten, and round out the volume to be a more complete guide to language test development, evaluation, and research.

The text is designed to meet the needs of teachers and teachers-in-training who are preparing to develop tests, maintain testing programs, or conduct research in the field of language pedagogy. In addition, many of the ideas presented here will generalize to a wider audience and a greater variety of applications. The reader should realize that, while few assumptions are made about prior exposure to measurement theory, the book progresses rapidly. The novice is cautioned against beginning in the middle of the text without comprehension of material presented in the earlier chapters. Familiarity with the rudiments of statistical concepts such as correlation, regression, frequency distributions, and hypothesis testing will be useful in several chapters treating statistical concepts. A working knowledge of elementary algebra is essential. Some rather technical material is introduced in the book, but bear in mind that mastery of these concepts and techniques is not required to become an effective practitioner in the field. Let each reader concentrate on those individually challenging matters that will be useful to him or her in application. While basic principles in measurement theory are discussed, this is essentially a "how-to" book, with focus on practical application.

This volume will be helpful for students, practitioners, and researchers. The exercises at the end of each chapter are meant to reinforce the concepts and techniques presented in the text. Answers to these exercises at the back of the book provide additional support for students. A glossary of technical terms is also provided. Instructors using this text will probably want to supplement it with sample tests, publications on current issues in testing, and computer printouts from existing test analysis software. These supplementary materials, readily available, will enhance the concrete, practical foundation of this text.

Grant Henning

Chapter One

Language Measurement: Its Purposes, Its Types, Its Evaluation

There could be no science as we know it without measurement. Testing, including all forms of language testing, is one form of measurement. Just as we weigh potatoes, examine the length of a piece of cloth, count eggs in a carton, or check the volume of a container of milk, so we test reading comprehension or spelling to determine to what degree these abilities are present in the learner. There is potential for error when we weigh potatoes. For example, the scale might not work properly, or it may not be highly sensitive, so that we must settle for a rough estimate of the correct weight. Furthermore, the potatoes might be wet or dirty, or there might be a few yams mixed in. In either case our measurement may be inaccurate.

In the same way, tests of language abilities may be inaccurate or *unreliable* in the sense that repeated measures may give different results. These measures may also be *invalid* in the sense that other abilities are mixed in. Our test of reading comprehension on closer examination may turn out to be a test of grammar or vocabulary, or at least a few such items may be "mixed in." Tests, to be useful, must provide us with reliable and valid measurements for a variety of purposes.

1.1 Purposes of Language Tests

Diagnosis and Feedback

Perhaps the most common use of language tests, and educational tests in general, is to pinpoint strengths and weaknesses in the learned abilities of the student. We may discover through testing that a given student has excellent pronunciation and fluency of oral production in the language of interest, but that he or she has a low level of reading comprehension. On further testing, we might

find that a low or too highly specialized vocabulary is a major factor underlying low reading comprehension for this student. We might recommend suitable approaches for vocabulary expansion.

This use of tests, frequently termed *diagnostic testing*, is of value in that it provides critical information to the student, teacher, and administrator that should make the learning process more efficient. Without the specific information thus made available, the teacher might persist in teaching pronunciation to this student and fail entirely to address a weakness in the area of vocabulary.

Screening and Selection

Another important use of tests is to assist in the decision of who should be allowed to participate in a particular program of instruction. In every instructional program, teaching staff and facilities are limited in number and capacity. It becomes a matter of serious concern to find an equitable means of determining who should be allowed to participate when there are more applicants than spaces available. Such selection decisions are often made by determining who is most likely to benefit from instruction, to attain mastery of language or content area, or to become the most useful practitioner in the vocational domain represented.

Considerable controversy has arisen about the fairness of tests and the possibility that they may contain cultural or other biases against minority population groups when used for purposes of selection (Scheuneman, 1984). Some researchers seem to indicate that the effects of cultural bias, though present, may be small and actually in favor of minorities (Chen and Henning, 1985). However, most educators agree that some, though perhaps not entire, reliance must still be placed on test scores when screening or selection decisions are being made (Lennon, 1978). In order for such decisions to be fair, our tests must be accurate in the sense that they must provide information that is both reliable and valid.

In the area of language testing, a common screening instrument is termed an *aptitude test* (Carroll, 1965). It is used to predict the success or failure of students prospective in a language-learning program.

Placement

Closely related to the notions of diagnosis and selection is the concept of placement. In this case tests are used to identify a particular performance level of the student and to place him or her at an appropriate level of instruction. It follows that a given test may serve a variety of purposes; thus the *UCLA Placement Exam* may be used to assign students to levels as well as to screen students with extremely low English proficiency from participation in regular university instruction.

Program Evaluation

Another common use of tests, especially *achievement tests*, is to provide information about the effectiveness of programs of instruction. In this way the

focus of evaluation is not the individual student so much as the actual program of instruction. Therefore, group mean or average scores are of greater interest in this case than are isolated scores of individual students. Often one or more *pretests* are administered to assess gross levels of student proficiency or "entry behavior" prior to instruction. Following the sequence of instruction, one or more *posttests* are administered to measure postinstructional levels of proficiency or "exit behavior." The differences between pretest and posttest scores for each student are referred to as *gain scores*.

Frequently in program evaluation tests or quizzes are administered at intervals throughout the course of instruction to measure "en route behavior." If the results of these tests are used to modify the program to better suit the needs of the students, this process is termed *formative evaluation*. The final exam or posttest is administered as part of the process of what is called *summative evaluation* (Scriven, 1967).

Sometimes language programs may be evaluated by comparing mean posttest or gain scores of one program or partial program with those of other programs. Whatever the method of evaluation, the importance of sensitive, reliable, and valid tests is obvious.

Providing Research Criteria

Language test scores often provide a standard of judgment in a variety of other research contexts. Comparisons of methods and techniques of instruction, textbooks, or audiovisual aids usually entail reference to test scores. Even examination of the structure of language itself or the physiological and psychological processes of language use may involve some form of measurement or testing. If we are to learn more about effective methods of teaching, strategies of learning, presentation of material for learning, or description of language and linguistic processes, greater effort will need to be expended in the development of suitable language tests.

Assessment of Attitudes and Sociopsychological Differences

Research indicates that only from one-quarter to one-half of the variability in academic achievement is explainable in terms of cognitive aptitude (Khan, 1969). The importance of noncognitive factors in achievement is seldom more evident that in the field of language learning, where the level of persistence and application needed for significant achievement is enormous. Attitudes toward the target language, its people, and their culture have been identified as important affective correlates of good language learning (Naiman et al., 1978; Saadalla, 1979). It follows that appropriate measures are needed to determine the nature, direction, and intensity of attitudes related to language acquisition.

Apart from attitudes, other variables such as cognitive style of the learner (Witkin et al., 1977), socioeconomic status and locus of control of the learner (Morcos, 1979), linguistic situational context (Henning, 1978), and ego-permeability of the learner (Henning, 1979) have been found to relate to levels of language achievement and/or strategies of language use. Each of these factors in

turn must be measured reliably and validly in order to permit rigorous scientific inquiry, description, explanation, and/or manipulation. This is offered as further evidence for the value of a wide variety of tests to serve a variety of important functions.

1.2 Types of Language Tests

Just as there are many purposes for which language tests are developed, so there are many types of language tests. As has been noted, some types of tests serve a variety of purposes while others are more restricted in their applicability. If we were to consider examples of every specific kind of test, or even of language tests alone, the remainder of this text might not suffice. There are, however, many important broad categories of tests that do permit more efficient description and explanation. Many of these categories stand in opposition to one another, but they are at the same time bipolar or multipolar in the sense that they describe two or more extremes located at the ends of the same continuum. Many of the categorizations are merely mental constructs to facilitate understanding. The fact that there are so many categories and that there is so much overlap seems to indicate that few of them are entirely adequate in and of themselves—particularly the broadest categories.

Objective vs. Subjective Tests

Usually these types of tests are distinguished on the basis of the manner in which they are scored. An objective test is said to be one that may be scored by comparing examinee responses with an established set of acceptable responses or *scoring key*. No particular knowledge or training in the examined content area is required on the part of the scorer. A common example would be a multiple-choice recognition test. Conversely a subjective test is said to require scoring by opinionated judgment, hopefully based on insight and expertise, on the part of the scorer. An example might be the scoring of free, written compositions for the presence of creativity in a situation where no operational definitions of creativity are provided and where there is only one rater. Many tests, such as cloze tests permitting all grammatically acceptable responses to systematic deletions from a context, lie somewhere between the extremes of objectivity and subjectivity (Oller, 1979). So-called subjective tests such as *free compositions* are frequently objectified in scoring through the use of precise *rating schedules* clearly specifying the kinds of errors to be quantified, or through the use of *multiple independent* raters.

Objectivity–subjectivity labels, however, are not always confined in their application to the manner in which tests are scored. These descriptions may be applied to the mode of item or *distractor* selection by the test developer, to the nature of the response elicited from the examinee, and to the use that is made of the results for any given individual. Often the term *subjective* is used to denote *unreliable* or *undependable*. The possibility of misunderstanding due to ambiguity suggests that objective–subjective labels for tests are of very limited utility.

Direct vs. Indirect Tests

It has been said that certain tests, such as ratings of language use in real and uncontrived communication situations, are testing language performance directly; whereas other tests, such as multiple-choice recognition tests, are obliquely or indirectly tapping true language performance and therefore are less valid for measuring language *proficiency*. Whether or not this observation is true, many language tests can be viewed as lying somewhere on a continuum from natural-situational to unnatural-contrived. Thus an *interview* may be thought of as more direct than a *cloze* test for measuring overall language proficiency. A contextualized vocabulary test may be thought more natural and direct than a synonym-matching test.

The issue of *test validity* is treated in greater detail in chapter seven. It should be noted here that the usefulness of tests should be decided on the basis of other criteria in addition to whether they are direct or natural. Sometimes *cost-efficiency*, and statistical measures of reliability or predictive validity, are more reflective of test utility than the naturalness or directness of the testing situation. Sometimes tests are explicitly designed to elicit and measure language behaviors that occur only rarely if at all in more direct situations. Sometimes most of the value of direct language data is lost through reductionism in the manner of scoring.

Discrete-Point vs. Integrative Tests

Another way of slicing the testing pie is to view tests as lying along a continuum from *discrete-point* to *integrative*. This distinction was originated by John B. Carroll (1961). Discrete-point tests, as a variety of diagnostic tests, are designed to measure knowledge or performance in very restricted areas of the target language. Thus a test of ability to use correctly the perfect tenses of English verbs or to supply correct prepositions in a cloze passage may be termed a discrete-point test. Integrative tests, on the other hand, are said to tap a greater variety of language abilities concurrently and therefore may have less diagnostic and remedial-guidance value and greater value in measuring overall language proficiency. Examples of integrative tests are random cloze, dictation, oral interviews, and oral imitation tasks.

Frequently an attempt is made to achieve the best of all possible worlds through the construction and use of *test batteries* comprised of discrete-point *subtests* for diagnostic purposes, but which provide a total score that is considered to reflect overall language proficiency. The comparative success or failure of such attempts can be determined empirically by reference to data from test administrations. Farhady (1979) presents evidence that "there are no statistically revealing differences" between discrete-point and integrative tests.

Here again, some tests defy such ready-made labels and may place the label advocates on the defensive. A test of listening comprehension may tap one of the four general language skills (i.e., listening, speaking, reading, and writing) in a discrete manner and thus have limited value as a measure of overall language

proficiency. On the other hand, such a test may examine a broad range of lexis and diverse grammatical structures and in this way be said to be integrative.

Aptitude, Achievement, and Proficiency Tests

Aptitude tests are most often used to measure the suitability of a candidate for a specific program of instruction or a particular kind of employment. For this reason these tests are often used synonymously with intelligence tests or screening tests. A language aptitude test may be used to predict the likelihood of success of a candidate for instruction in a foreign language. The Modern Language Aptitude Test is a case in point (Carroll and Sapon, 1958). Frequently vocabulary tests are effective aptitude measures; perhaps because they correlate highly with intelligence and may reflect knowledge and interest in the content domain (Henning, 1978).

Achievement tests are used to measure the extent of learning in a prescribed content domain, often in accordance with explicitly stated objectives of a learning program. These tests may be used for program evaluation as well as for certification of learned competence. It follows that such tests normally come after a program of instruction and that the components or items of the tests are drawn from the content of instruction directly (Mehrens and Lehmann, 1975). If the purpose of achievement testing is to isolate learning deficiencies in the learner with the intention of remediation, such tests may also be termed *diagnostic tests.*

Proficiency tests are most often global measures of ability in a language or other content area. They are not necessarily developed or administered with reference to some previously experienced course of instruction. These measures are often used for placement or selection, and their relative merit lies in their ability to spread students out according to ability on a proficiency range within the desired area of learning.

It is important to note that the primary differences among these three kinds of tests are in the purposes they serve and the manner in which their content is chosen. Otherwise it is not uncommon to find individual items that are identical occurring in aptitude, achievement, and proficiency tests.

Criterion- or Domain-Referenced vs. Norm-Referenced or Standardized Tests

There is no essential difference between *criterion-referenced tests* and *domain-referenced tests*; but a third related category, *objectives-referenced tests*, differs from the other two in that items are selected to match objectives directly without reference to a prespecified domain of target behaviors (Hambleton et al., 1978). There exists such controversy between the advocates of *criterion-referenced tests* and the advocates of *norm-referenced tests* that greater description and explanation is warranted at this point (Ebel, 1978; Popham, 1978).

Characteristically criterion-referenced tests are devised before the instruction itself is designed. The test must match teaching objectives perfectly, so that any tendency of the teacher to "teach to the test" would be permissible in that attaining objectives would thereby be assured. A criterion or *cut-off* score is set in advance

(usually 80 to 90 percent of the total possible score), and those who do not meet the criterion are required to repeat the course. Students are not evaluated by comparison with the achievement of other students, but instead their achievement is measured with respect to the degree of their learning or mastery of the prespecified content domain. Consistent with a view of teacher or environmental responsibility for learning, the failure of a large proportion of the learners to pass part or all of the test may result in the revision of the course or a change in method, content, instructor, or even the objectives themselves.

When applied to the field of language measurement, these tests have both strengths and weaknesses. On the positive side, the process of development of criterion-referenced tests is helpful in clarifying objectives. Such tests are useful in ascertaining the degree to which objectives have been met, both in ongoing, formative evaluation and in final, summative evaluation. The tests are useful when objectives are under constant revision. The tests are useful with small and/or unique groups for whom *norms* are not available. Test security is considered less of a problem since students know in advance the precise content domain for which they are held responsible on the test. For the same reason, students' test anxiety is believed to be reduced with this type of test.

On the negative side, the objectives measured are often too limited and restrictive, as is frequently true when objectives must be specified operationally. Another possible weakness is that scores are not referenced to a norm, so students typically are unable to compare their performance with that of other students in the population of interest. Bright students, who easily attain the criterion level of mastery, may not be encouraged to reach higher standards of excellence. The very establishing of the criterion or cut-off score is in practice usually highly arbitrary. Until rate of success is compared with other students in other years and other settings (as in norm-referenced testing), it is difficult to know what is meant by reaching criterion. Techniques of estimating reliability and validity of such tests are only beginning to be developed, so it is not yet clear in most cases whether a given test is reliable or valid in any scientific sense (Popham, 1978).

Norm-referenced or *standardized* tests are quite different from criterion-referenced tests in a number of respects; although, once again, some of the identical items may be used under certain conditions. By definition, a norm-referenced test must have been previously administered to a large sample of people from the target population (e.g., 1,000 or more). Acceptable standards of achievement can only be determined after the test has been developed and administered. Such standards are found by reference to the mean or average score of other students from the same population. Since a broad range or distribution of scores is desired, items at various levels of difficulty are purposely included. Commensurate with a felt need to discriminate between low-achieving and high-achieving students, and consistent with a philosophy of learner responsiblity for achievement, the failure of a large number of students to pass all or a given portion of the test usually results in the revision of the test itself rather than revision of the program or dismissal of the teacher.

For purposes of language testing and testing in general, norm-referenced tests also have specific strengths and weaknesses. Among the strengths is the fact that

comparison can easily be made with the performance or achievement of a larger population of students. Also, since estimates of reliability and validity are provided, the degree of confidence one can place in the results is known. To the extent that the test is available, research using the test is readily replicable. Since acceptable standards of achievement are determined empirically with reference to the achievement of other students, it may be argued that such standards are fairer and less arbitrary than in the case of criterion-referenced tests. Since examinees are purposely spread on the widest possible range of performance results, it may be argued that more comparative information is provided about their abilities than in the case where only pass-fail information is available.

Norm-referenced tests are not without their share of weaknesses. Such tests are usually valid only with the population on which they have been normed. Norms change with time as the characteristics of the population change, and therefore such tests must be periodically renormed. Since such tests are usually developed independently of any particular course of instruction, it is difficult to match results perfectly with instructional objectives. Test security must be rigidly maintained. Debilitating test anxiety may actually be fostered by such tests. It has also been objected that, since focus is on the average score of the group, the test may be insensitive to fluctuations in the individual. This objection relates to the concept of reliability discussed in chapter six, and may be applied to criterion-referenced as well as to norm-referenced tests.

Speed Tests vs. Power Tests

A purely *speed test* is one in which the items are so easy that every person taking the test might be expected to get every item correct, given enough time. But sufficient time is not provided, so examinees are compared on their speed of performance rather than on knowledge alone. Conversely, *power tests* by definition are tests that allow sufficient time for every person to finish, but that contain such difficult items that few if any examinees are expected to get every item correct. Most tests fall somewhere between the two extremes since knowledge rather than speed is the primary focus, but time limits are enforced since weaker students may take unreasonable periods of time to finish.

Other Test Categories

The few salient test categories mentioned here are by no means exhaustive. Mention could be made of *examinations* vs. *quizzes*, *questionnaires*, and *rating schedules* treated more fully in chapter two. A distinction could be made between single-stage and multi-stage tests as is done in chapter nine. Contrast might be made between language skills tests and language feature tests, or between production and recognition tests.

At a still lower level of discrimination, mention will be made of cloze tests, dictation tests, multiple-choice tests, true/false tests, essay/composition/precis tests, memory-span tests, sentence completion tests, word-association tests, and imitation tests, not to mention tests of reading comprehension, listening comprehension,

grammar, spelling, auditory discrimination, oral production, listening recall, vocabulary recognition and production, and so on. Figure 1.1 provides a partial visual conceptualization of some types of language tests.

1.3 Evaluation of Tests

A consideration of the purposes and types of tests is only preparatory to the selection or development of tests for any stipulated use. When faced with the responsibility of having to choose or develop an appropriate test, we should take still further matters into consideration, including such information as *the purpose of the test, the characteristics of the examinees, the accuracy of measurement, the suitability of format and features of the test, the developmental sample, the availability of equivalent or equated forms, the nature of the scoring and reporting*

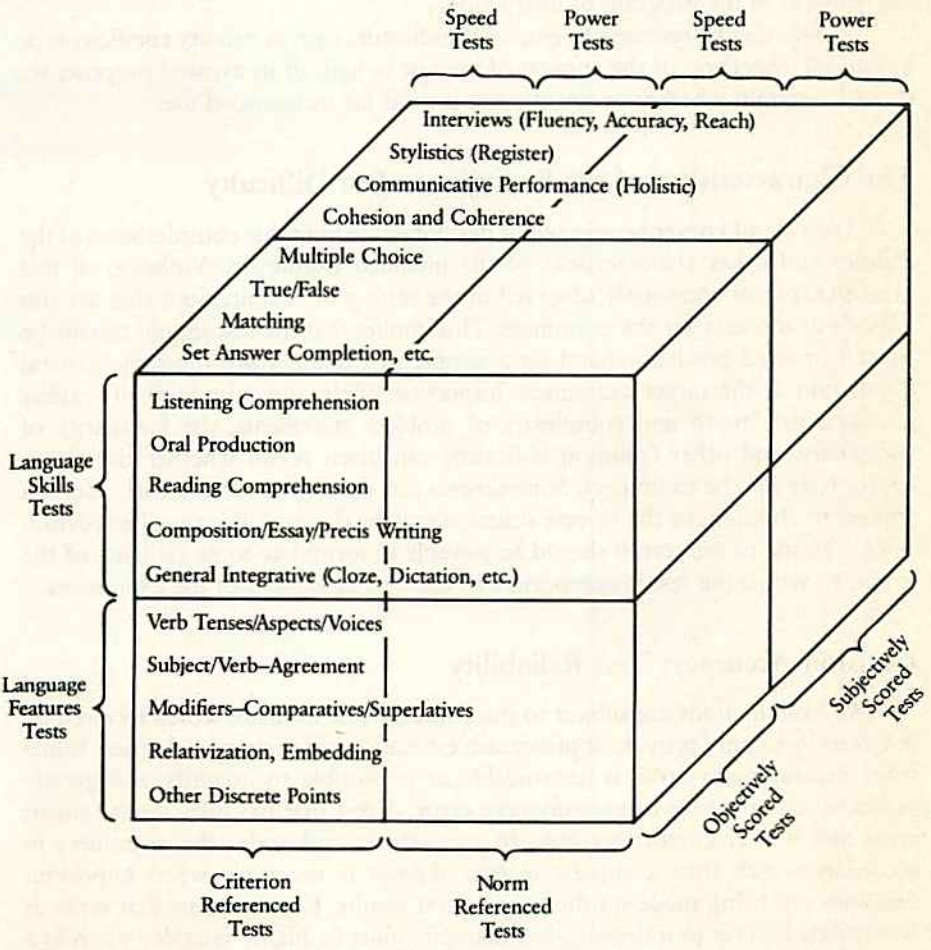


FIGURE 1.1 A partial conceptualization of types of language tests

of scores, the cost, the procurement, and *the political acceptability of the test.* These considerations are examined in greater detail in this section of the chapter.

The Purpose of the Test: Test Validity

Perhaps the first and foremost consideration in selecting or developing a test is, "What is the test going to be used for?" Most standardized tests come equipped with a manual reporting validity coefficients for a variety of uses. The overriding concern here is that the test should adequately measure what it is supposed to measure. Nearly all tests are valid for some purposes, but not for others. Validity is discussed in greater detail in chapter seven. At this point it is sufficient to ask ourselves, "Is the content of the test consistent with the stated goal for which the test is being administered?" If, for example, the test is to be an achievement test, accurately reflecting the extent to which students have mastered the content of instruction, the test itself must not contain material which was not encountered by the students in the program of instruction.

So, whether by recourse to empirical indicators such as validity coefficients or by careful inspection of the content of the test in light of its avowed purpose, we should ascertain whether or not the test is valid for its intended use.

The Characteristics of the Examinees: Test Difficulty

A test should never be selected or developed without due consideration of the abilities and other characteristics of the intended examinees. Violation of this principle is most commonly observed in the setting of examinations that are too difficult or too easy for the examinees. This implies that the test should usually be piloted or tried out beforehand on a sample of persons from the same general population as the target examinees. Inspection of the difficulty level of reading passages, the length and complexity of problem statements, the familiarity of vocabulary, and other common indicators can often reveal whether the test is appropriate for the examinees. Sometimes a test developed for use with adults is applied to children, or the reverse situation may be the case. By critical inspection or by piloting of the test, it should be possible to formulate some estimate of the extent to which the test is appropriate to the overall abilities of the examinees.

Decision Accuracy: Test Reliability

All examinations are subject to inaccuracies. The ultimate scores received by the examinees only provide approximate estimations of their true abilities. While some measurement error is unavoidable, it is possible to quantify and greatly minimize the presence of measurement error. A test that has little measurement error and that is found, therefore, to consistently rank-order the examinees in accordance with their comparative true abilities is necessary when important decisions are being made on the basis of test results. Examinations that serve as admissions criteria to university, for example, must be highly reliable; whereas a quiz used to determine which children may be rewarded with a longer recess

period between classes would be less critical. Both should be reliable, but the importance of the decision in the former situation requires greater reliability than the decision in the latter situation. When we consider the nature of the decision to be based upon the results of the test, and when we determine in this way the extent of reliability required, related decisions, such as the desired length of the test, can be made. Since test reliability is related to test length, so that longer tests tend to be more reliable than shorter tests, knowledge of the importance of the decision to be based on examination results can lead us to use tests with different numbers of test items. Here again, for standardized tests, accompanying test manuals usually report the reliability levels estimated. In developing our own tests, we should estimate test reliability using procedures described in chapter six.

Suitability of Format and Features: Test Applicability

The actual test format, or the way in which the author has chosen to measure the target ability, may be unfamiliar to the examinees. Consider the example of cloze tests rashly applied to students who have never encountered tests of this type or format before. Certainly their performance will suffer because of lack of familiarity with the test. Their results will not, however, be a reflection of their actual ability or underlying linguistic competence. Consider another example, that of a test that requires the use of tape-recording equipment which may not be available in the actual testing situation. The teacher or test administrator may choose to administer such a test using live voices; but, while such a decision may make the test applicable, it may also seriously alter the nature of what is being measured by the test. The actual choice of a test should be contingent on whether the format and features of the test may fairly be applied in the real testing situation.

The Developmental Sample: Test Relevance

Most widely used tests were originally normed on a sample of persons from a particular population. Often in the case of tests of English as a Foreign Language developed in the United States or in Britain, the developmental sample consisted of entering university students from diverse national backgrounds. One unfortunate artifact of this procedure is that such tests may turn out to be more reliable and valid as indicators of language proficiency for persons from one particular language background than for those from some other background. Or the tests may prove to be less reliable for persons from either of these hypothetical backgrounds than would tests of equivalent length developed exclusively for persons from one or the other of the language backgrounds. All this is by way of pointing out that it is important to consider the characteristics of the sample on which the test was developed before it is blindly applied to some sample from a different underlying population.

Just as we may speak of a sample of persons drawn from a particular population and the relevance of a test developed on that sample when it is applied to a sample from a different population, so we may speak of a sample of test items drawn from a particular domain of items and the relevance of this domain to

particular teaching objectives. It follows that certain items drawn from certain domains are irrelevant to certain specified objectives. By considering both the sample of persons on which a test was developed as well as the sample of items drawn from a particular domain, we can judge the relevance of a test for a particular group of people and for a particular stated objective. This notion is closely related to the concept of validity already mentioned. The distinction here is that focus is on the relevance of the examinees and/or the test item domain.

Availability of Equivalent or Equated Forms: Test Replicability

Usually when a testing program is established, there is the intention to test examinees repeatedly over a period of time. Perhaps there is a plan to administer a similar test at the end of every term of instruction or to apply the same test repeatedly as an admissions or a screening instrument. Such procedure is often necessary as an indication of trends over time, reflecting a need to compare examinees from one administration to another. In such circumstances one should choose or develop a test that has equivalent or equated forms, otherwise there will probably be a security breakdown and the test will no longer function as intended. Equivalent or alternative forms are so developed that a raw score on one is nearly equal in meaning to that on another, although the same exact items usually do not appear on the two tests. Equivalent forms are extremely difficult to develop and are therefore highly uncommon. More common are equated forms, which are accompanied by some conversion table whereby the test scores obtained on one form of the test may be converted or transformed to scores on the same scale as the scores of some other test. In this way scores may be compared across forms, even though the test forms may differ in length, difficulty, and so on. Procedures for developing equated or equivalent forms are discussed in chapter six.

Scoring and Reporting: Test Interpretability

Before a test is selected, developed, or applied, it is crucial to understand how the test is to be scored, how the scores are to be reported, and how the scores are to be interpreted. For some tests, such as the Foreign Service Institute Interview (Clark and Swinton, 1980), the scoring procedure requires extensive training and experience. Moreover, once scores have been recorded it is necessary to understand what the scores mean in order to communicate this information most efficiently to the examinees and to those who make decisions on the basis of scores obtained. In most instances a test whose scores can be more efficiently obtained, reported, and interpreted is to be preferred over a test that involves considerable delay, expense, and inconvenience for the same process.

Cost of Test Procurement, Administration, and Scoring: Test Economy

Ultimately one must consider the cost of the test. If the purchase price and the scoring services are too expensive, one may choose to develop one's own test. But

if the cost of test development is greater than the cost of purchasing a test already available, then one should usually opt for the latter. Economics of the test should be considered in terms of the cost of purchase or development, the cost of administration including duplicating and proctoring, the cost of scoring, and the cost of reporting, storing, and interpreting scores.

Procurement of the Test: Test Availability

Many tests employed in research literature were developed for one particular research situation and are not available for public use. Even popular standardized tests are not made available to every person for every purpose. Many of these tests will be released only to qualified persons or institutions after special application is made and permission is granted. The point here is that, even after a test is determined to be appropriate in every way including price, there is the possibility that the test is not available. In beginning a testing program one should determine which tests are actually available before final decisions are made.

Political Considerations: Test Acceptability

From the start, any test advocated must satisfy societal and institutional demands. In developing English tests for the ministry of education of one foreign country, the test developers learned that certain constraints existed from the outset (Henning et al., 1981). Any test developed was required to have an equal number of production and recognition items; furthermore, objective tests of reading, writing, and grammar had to be included. In accordance with existing classroom space, no more than 5 percent of the students could be failed at any given level. Some innovation was possible within these existing constraints. It is incumbent on the test developer or test selector to determine what the limits of acceptability are. Many excellent tests have been abandoned simply because they were not found acceptable in the eyes of teachers, parents, or administrators. One does well at this point to involve a wide spectrum of the community in the decision-making process.

These ten considerations in the selecting and developing of appropriate tests are summarized in the checklist of Table 1.1. One may use such a checklist by rating any given test on a scale of one to ten for each of the ten criteria cited. In this way a perfect test would obtain a combined rating of 100. Rough comparisons may be made among tests competing for selection.

1.4 Summary

This chapter has been concerned with testing terminology; introductory notions of *reliability* and *validity* have been presented. Six specific purposes of tests have been described. Categories of tests have been introduced, including *objective/subjective*, *direct/indirect*, *discrete-point/integrative*, *aptitude/achievement/proficiency*, *criterion-referenced/norm-referenced*, *speed/power*, and so on.

TABLE 1.1 A checklist for test evaluation

Name of Test _____
Purpose Intended _____

Test Characteristic	Rating (0 = highly inadequate, 10 = highly adequate)
1. Validity	_____
2. Difficulty	_____
3. Reliability	_____
4. Applicability	_____
5. Relevance	_____
6. Replicability	_____
7. Interpretability	_____
8. Economy	_____
9. Availability	_____
10. Acceptability	_____
	_____ Total

Finally, a checklist was presented for the rating of the adequacy of any given test for any given purpose using ten essential criteria.

Exercises

1. Give examples of what is intended by the terms *reliable* and *valid*.
2. List five common purposes for language tests and explain the purpose for which you have used a test most recently—either as examiner or examinee.
3. Considering the test categories presented in this chapter, how would you label a cloze test* employing a passage with every fifth word deleted, beginning from a randomized starting point, which was scored in such a way that only the exact original words of the passage were accepted, and students' scores were compared with those of other students to determine comparative language proficiency?
4. List three kinds of language tests that are not explicitly named in this chapter. Provide a brief description.
5. Distinguish between objective and subjective tests.
6. Distinguish between direct and indirect tests.
7. Distinguish between norm-referenced and criterion-referenced tests.
8. Distinguish between discrete-point and integrative tests.
9. Distinguish between speed tests and power tests.
10. Choose a particular test and rate it for adequacy for a given purpose by reference to the checklist of Table 1.1

*For additional information on cloze testing, consult Oller (1979).

Chapter Two

Measurement Scales

2.1 What Is a Measurement Scale?

Scales are most frequently thought of in conjunction with the measurement of weight. But scales are also used in psychological and educational measurement. If tests are the instruments of measurement, scales are the gauges of measurement. The magnitude of any measurement is indicated by some point or position on a scale, expressed as a number or a score. In this chapter we consider a variety of measurement scales, their purposes, and their characteristics.

2.2 What Are the Major Categories of Measurement Scales?

Table 2.1 summarizes the four major types of psychological measurement scales and their purposes with examples of how they are used. Each category appears in the research literature of education, psychology, and linguistics. Indeed, it is only insofar as these metrics have been applied that such behavioral and social fields of study can be termed sciences.

Nominal Scales

Nominal scales are used with *nominal* or *categorical* variables such as sex, native-language background, eye color, preferred teaching method, or phrasal and nonphrasal verbs occurring in speech. Such variables are not normally thought of as existing along a continuum as a matter of degree. They are either present or not

TABLE 2.1 Types, purposes, and examples of measurement scales

Type of Scale	Purpose of Scale	Example of Scale Use
Nominal	Counting frequency	Finding number of native speakers of French in an ESL class
Ordinal	Rank ordering	Ranking students according to frequency of spelling errors
Interval	Measuring intervals	Determining z-scores or standard scores on a grammar test
Ratio	Measuring intervals from a real zero point	Measuring height, weight, speed, or absolute temperature

present in any given situation or observance, and the task of the examiner is to count or tally their frequencies of occurrence. While such frequency counts may be made within a given category, the results are not additive across categories. Combining five native speakers of Tagalog with six brown-eyed persons would give a meaningless result. But the five Tagalog speakers in one class may be combined with the six Tagalog speakers in another class to provide a measure of the total number of Tagalog speakers in both classes; i.e., eleven.

Nominal data are often elicited in demographic questionnaires or sociolinguistic surveys requiring background information about the examinees. Such information can be of vital importance, but care must be taken in the manner in which it is handled. Only a limited range of statistical tests can be applied or inferences drawn from such data. (Cf. Tuckman, 1972, p. 229, for a table of appropriate statistical tests.)

Ordinal Scales

Ordinal scales, as the name implies, are used to rank-order examinees according to ability on some ability continuum, or according to frequency of a particular kind of error in writing, for example. Reference to an ordinal scale will, for example, provide us with information about which student is first, second, or third in achievement of specified objectives in a program of language instruction. It will not, however, shed light on the distance or interval between examinees. Thus we would not know if student number one is vastly superior to student number two or just slightly more capable. More to the point, we could not be sure whether the amount of achievement superiority exhibited by the first student with regard to the second student was the same as that of the second student with reference to the third student. Nor could we infer that the achievement of the first student was twice as great as that of the second student. Here again, the range of appropriate statistical tests applicable to ordinal data is limited (Tuckman, 1978).

Since tests measure performance from which we infer competence, and since the raw scores of all language tests present us with ordinal data, we here confront a major problem of language measurement and all psychological measurement. When a student gets a score of 68 on a 100-item vocabulary test, chances are this

student has greater knowledge of vocabulary than a student who scores 62 on the same test, provided the test is reliable. But since there is not a one-to-one correspondence between number of items passed and degree of underlying competence, still another student scoring 70 may be vastly more knowledgeable of vocabulary than the student scoring 68, while the student scoring 68 may be only slightly more knowledgeable than the student scoring 62—even if the particular items missed are similar for each student. The problem is usually solved by conversion of our ordinal scale to an interval scale, described on pages 18–21. Several kinds of ordinal scales are illustrated in Table 2.2.

Raw Scores or Obtained Scores

A raw score is merely the total number of correct items, the total score possible minus the cumulative penalties due to errors, or the numerical score of the examinee before any transformations are performed on the test score data.

Ordinal Ratings

An ordinal rating is a rank assigned to the examinee designating his or her status on some ability continuum being measured, with respect to other examinees on the same continuum. An ordinal rating may be inferred from the raw scores in Table 2.2, or it may be determined by direct examiner observation as in an interview.

Percentage Scores

Percentage scores may be easily obtained by dividing the raw score by the total score possible and multiplying the result by 100. In the example of Table 2.2, the percentage score of the examinee with a raw score of 5 was determined as follows:

$$\frac{5}{25} \times 100 = 20$$

Since 25 was the maximum possible, that figure was divided into the obtained raw score, and the result was multiplied by 100. While percentage scores have the advantage of adjusting the range of possible scores to extend from zero to 100, they do not add information nor alter the ordinal nature of the score scale.

TABLE 2.2 Ordinal scales for a 25-point dictation subtest for six students

Raw Scores	0	5	12	12	14	19
Ordinal Ratings	5th	4th	3rd	3rd	2nd	1st
Percentage Scores	0	20	48	48	56	76
Cumulative Percentage Scores	16.67	33.33	50.00	66.67	83.33	100.0
Percentile Ranks	8.33	25.00	50.00	50.00	75.00	91.67

Cumulative Percentage Distribution Scores

Since there were only six students in Table 2.2, each student accounts for one-sixth or 16.67 percent of the total number of students. By the same standard, consideration of the lowest-scoring three students accounts for 50 percent of the total number of students, cumulatively. Having arranged the students from lowest to highest according to raw score, we could simply assign a cumulative frequency distribution score to each student in accordance with the percentage of students accounted for at each point on the raw score continuum. Not only does computation of these cumulative percentage distribution scores not improve on the ordinal nature of the raw score scale, it adds arbitrary distortion by assigning different values to examinees who obtained the same raw score and may have equal ability (e.g., the two students in Table 2.2 with a raw score of 12 obtained cumulative percentage distribution scores of 50 and 66.67).

Percentile Scores or Percentile Ranks

Percentile scores or ranks, also termed *centile ranks* (Guilford and Fruchter, 1973), are obtained by adding the number of examinees scoring below a given examinee of interest to one-half the number of examinees obtaining the same score as this examinee, dividing this total by the total number of examinees, and multiplying by 100. In Table 2.2 the percentile score of the examinee with a raw score of 19 was obtained as follows:

$$\frac{5 + 1/2}{6} \times 100 = 91.67$$

Although percentile scores are extremely useful and considerably more meaningful than the other ordinal scales mentioned, they still do not provide us with equal interval measurement of the underlying competence of interest to the examiner.

Interval Scales

Interval scales have the profound advantage of providing all of the information of ordinal scales in terms of rank-ordering students and in addition a meaningful interval for considering comparative distances between scores. Interval scales are usually obtained by the transformation or normalization of ordinal scales. Inferences are drawn about the metric of underlying competencies measured based on the size and shape of the distribution of raw scores obtained.

The most common interval scales include *z-score*, *T-score*, *normal distribution area proportion*, *stanine*, and *I.Q.-equivalent scales*. In Table 2.3 the raw scores from Table 2.2 have been transformed into a variety of interval scale scores.

Table 2.3 illustrates interval scale computations. Such transformations imply the existence of an underlying normal distribution and a much larger sample of examinee scores. More information on normal distributions is provided in chapter three.

TABLE 2.3 Interval scales for a 25-point dictation subtest for six students

Raw Scores	0	5	12	12	14	19
z-Scores	−1.86	−0.96	0.30	0.30	0.66	1.56
T-Scores	31.4	40.4	53.0	53.0	56.6	65.6
Normal Distribution Area Proportions	.03	.17	.62	.62	.74	.94
Stanines	1	3	6	6	6	8
I.Q.-Equivalents (WISC)	72.1	85.6	104.5	104.5	109.9	123.4

z-Scores

For the computation of z-scores from raw scores, the following formula is used:

$$z = \frac{X - M}{s} \tag{2.1}$$

where, X refers to the raw score of a given examinee
M indicates the mean or average score
s refers to the *standard deviation* of the raw score distribution

The concept of standard deviation and how it is derived mathematically is further explained in chapter three. For the distribution of raw scores in Table 2.3, the mean is found to be 10.33, and the standard deviation is 5.55. Therefore the z-score of the examinee with raw score zero is determined as follows:

$$z = \frac{0 - 10.33}{5.55} = -1.86$$

Normalized standard scores or *z-scores* equal the number of standard deviations the raw score is found away from the mean score. Thus a raw score equal to the mean of raw scores would have a z-score of zero. z-scores have a number of important functions in addition to that of providing us with this particular interval scale. These functions are discussed below and on pages 20, 21 and 59.

T-Scores

T-Scores are designed to have a mean of 50 and a standard deviation of 10. Thus one can use the following formula for the computation of T-scores from z-scores:

$$T = 10z + 50 \tag{2.2}$$

In the case of the examinee in Table 2.3 with raw score zero, the T-score is determined as follows:

$$T = 10(-1.86) + 50 = 31.4$$

Normal Distribution Area Proportions

Normal distributions have the interesting characteristic that there is a direct correspondence between the magnitude of the z-scores and the area under the curve of the normal distribution. Since, as has been noted, a z-score of zero corresponds to the mean of the raw score distribution, we can readily see that it divides the distribution of scores in half. Therefore a z-score of zero can be said to be at a point above 50 percent of the test scores and below 50 percent of the test scores if the distribution is normal. In the same way, every z-score corresponds to a proportion of the distribution of scores represented. To obtain normal distribution area proportions, we have only to compare the z-scores and consult Table B in Appendix A for the corresponding proportions. Thus, for the student with raw score zero in Table 2.3, an area proportion of .03 was found in Table B to correspond to a z-score of -1.86.

Stanines

Stanine scores, derived from "standard nines," comprise a nine-point interval scale that is often used in test-score profiles, but is generally less sensitive and discriminating than T-scores for the same purpose. Stanine scores have a mean of 5 and a standard distribution of 1.96, except at the ends of the distribution where there is slight distortion due to the desire to maintain just nine points. To determine stanine score from any normal distribution of scores, one may assign a score of 1 to the lowest 4 percent, 2 to the next 7 percent, 3 to the next 12 percent, 4 to the next 17 percent, 5 to the next 20 percent, 6 to the next 17 percent, 7 to the next 12 percent, 8 to the next 7 percent, and 9 to the final 4 percent. Table 2.4 relates stanine scores to cumulative area proportions.

The stanine scores of Table 2.3 were easily obtained from Table 2.4 by taking the normal distribution area proportions of Table 2.3 and consulting Table 2.4 for the corresponding stanine scores.

I.Q.-Equivalent Scores

I.Q. or intelligence quotient scores are traditionally calculated by dividing *mental age* by *chronological age* and multiplying the result by 100. Thus an I.Q. score of 100 would be average in the sense that one's mental age is exactly equivalent to one's chronological age. I.Q. scores have fallen under disrepute for a number of reasons, including the fact that intelligence does not seem to follow a regular continuum throughout the possible range of scores and the fact that the

TABLE 2.4 Stanine scores and their normal distribution area cumulative proportions

Stanine Score	1	2	3	4	5	6	7	8	9
Normal Distribution Area Cumulative Proportion	0.00–.044	.046–.106	.107–.221	.222–.396	.397–.593	.594–.768	.769–.889	.890–.955	.956–1.00

standard deviation for I.Q. varies from test to test (usually from about 10 to 15). The Wechsler Intelligence Scale for Children (WISC) uses a standard deviation of 15 and is among the most widely used I.Q. measures. The same scale was used to compute the I.Q.-equivalent scores in Table 2.3. Such scores may be computed as follows:

$$\text{I.Q.-equivalent} = 15z + 100 \tag{2.3}$$

In the case of the student with a dictation raw score of zero in Table 2.3, the I.Q.-equivalent was computed in the following way:

$$\text{I.Q.-equivalent} = 15(-1.86) + 100 = 72.1$$

Ratio Scales

Ratio scales provide all of the information of ordinal scales in that they permit a rank-ordering according to magnitude of score. Such scales also provide all the benefits of interval scales in that every point on the scale is equidistant from the adjacent points. But ratio scales have an added feature: they join all measures to a real or absolute zero-point on the scale. Ratio scales are more typical of the physical sciences than of language testing in the behavioral and social sciences. If we find someone is 2 meters tall, we know that that person is 2 meters above a known zero-point. Furthermore, we know that such a person is exactly twice as tall as a one-meter-high table. In the realm of testing language proficiency, even after converting raw scores to an interval scale, we do not know that someone with a T-score of 60 is twice as proficient as someone with a T-score of 30. Nor can we be sure that someone with a raw score or an interval score of zero has no proficiency whatever.

Measures of height, weight, absolute temperature, speed, and time from a known starting point are said to be on a ratio scale. Time and speed or rate are the only ratio scales commonly occurring in the field of language or psychological measurement.

2.3 Questionnaires and Attitude Scales

Frequently in research in the behavioral sciences we have need of questionnaires designed to elicit attitudes and opinions of respondents. Such questionnaires or rating schedules usually employ some sort of artificial scale in order to gather information in the most efficient way. In this section we consider various kinds of these artificial scales of use in the construction of our own questionnaires. Of particular interest will be demographic, Likert, semantic differential, and general rating scales.

Demographic Scales

As the name indicates, with demographic scales we are mostly concerned with the elicitation of vital statistics or other personal information about the respondents. Table 2.5 presents examples of demographic items about language background.

As you can see from Table 2.5, there are many ways to elicit demographic information, even about a topic so limited as language background. Item 1 in the table will be recognized as using a nominal scale, since respondents are required only to list the number of languages known. Item 2 makes use of a four-point scale in the rating of each language skill. Item 3 involves a percentage scale. The selection of the scale depends upon the exact nature of the information sought.

TABLE 2.5 Sample demographic items regarding language background

1. Indicate the number of languages in which you are fluent. _____

2. Rate your mastery in your first foreign language.
   Name of language: _____

Level of mastery:	Poor	Fair	Good	Excellent
Listening				
Speaking				
Reading				
Writing				

3. If you are bilingual or multilingual, list the languages you know and indicate the percentage of time you usually use each language in any given day or week.

Name of Language	Percent
	100% total

4. Opposite each language you know indicate (X) the degree of confidence you would feel in a normal speaking situation.

Name of Language	Highly Confident	Somewhat Confident	Not Confident

5. List the languages you know and opposite each indicate the number of years you have studied or used that language.

Name of Language	Number of Years

Demographic items are often used to elicit other information such as age, level of education, father's occupation, income, place of birth, etc.

The Likert Scale

The Likert scale is a popular five-point scale used most commonly to elicit extent of agreement with some statement of opinion or attitude. The respondent is usually requested to circle the number or letters coinciding with his or her reaction to a statement. Table 2.6 provides examples of Likert scale questions designed to elicit attitudes toward the French language and people.

Notice that each of the questionnaire items in Table 2.6 represents an opinion or attitude about the French people, language, or culture. Respondents indicate the extent of their agreement by circling the letters corresponding to their reaction to each statement.

Scoring the Likert Scale

For scoring purposes, the letters SA to SD must be converted to numbers 1 to 5 respectively. It is not particularly important whether the scale runs from 1 to 5 or from 5 to 1; that is, whether SA is coded as 1 or 5, so long as the significance of the direction is borne in mind and the same procedure is consistently applied. The point here is that the item scores are usually intended to be additive so as to form a total attitude score.

Avoiding the "Halo" Effect

Some respondents, particularly those who are favorably disposed toward the subject of a given questionnaire, may tend indiscriminately to supply positive ratings for every item without due consideration of the import of each individual item. To avoid this, it is usually necessary to include negatively stated items or items for which the direction is reversed. Item 3 of Table 2.6 is an illustration of this procedure. This item, unlike the other items, is negatively stated. This is done purposely to avoid a "halo" effect. There is, however, a cautionary reminder necessary here. When scoring Likert scales with mixed positive and negative items, we must remember to invert the negative item scales so that their scores will be additive with those of the other items of the questionnaire.

TABLE 2.6 Likert scale items for the measurement of attitudes toward French

Instructions: Indicate the extent of your agreement or disagreement with the following statements. Do this by circling the letters corresponding to your opinion about each statement. (SA—strongly agree; A—agree; U—undecided; D—disagree; SD—strongly disagree)

1. I like to have French people for friends.	SA	A	U	D	SD
2. The French language is beautiful to hear.	SA	A	U	D	SD
3. Culturally I have little to gain from learning French.	SA	A	U	D	SD

Use of Filler Items

Sometimes it is necessary to disguise the purpose of a questionnaire in order to ensure that item responses will be valid. If respondents realize that we are measuring dogmatism or locus of control, this awareness may color their responses. To avoid this possibility, it is sometimes necessary to include filler items. These are items unrelated to the purpose of the questionnaire which are randomly dispersed among the actual items to prevent respondents from inferring the purpose of the questionnaire. In scoring, these items are disregarded.

Checking Reliability and Validity

When devising questionnaires for use in research, it is always desirable to check on the reliability and validity of such instruments. Reliability and validity are discussed in chapters six and seven, and procedures for their estimation are presented there. Suffice to note at this point that one common method for estimating reliability of a questionnaire is simply to administer it more than one time to the same group of persons over a period of time to determine whether their responses are consistent or not. Validity might be partially established by correlating each item with the total score and discarding items with low or negative correlations. In this way we can be certain that the items are homogeneous, measuring the same underlying trait.

The Semantic Differential

The semantic differential, as devised by Osgood, Suci, and Tannenbaum (1957), is a common scale for eliciting affective responses. Typically this scale consists of seven-point ratings on a series of bipolar continua. A series of antonymous adjectives such as *good/bad*, *hot/cold*, *friendly/hostile* are placed at extreme ends of the seven-point scale. The task of the respondent is, for each item, to mark the point on the scale that most nearly approximates his or her affective response to the subject being rated. An example of a semantic differential questionnaire, which might be employed in the rating of the recorded speech of a person using a nonstandard dialect of English, is provided in Table 2.7.

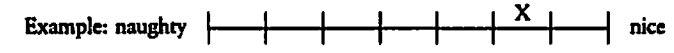
Notice that, in this instance, the adjectives were carefully selected to be descriptive of people—we did not, for example, include *wet/dry* or *shiny/dull*. It is most irritating for respondents to provide ratings on highly irrelevant characteristics. Osgood and his colleagues determined through factor analytic techniques that the various bipolar dimensions may be reduced to three major categories: *good/bad*, *active/passive*, and *strong/weak*. These categories were restated as *evaluative*, *activity*, and *potency* dimensions. Adjective pairs for inclusion in our questionnaires may now be selected to represent these important dimensions.

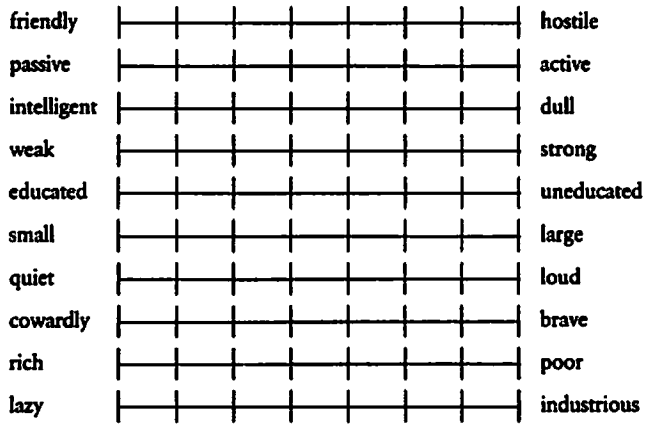
Scoring the Semantic Differential

As in the case of the Likert scale, responses to the semantic differential must be translated into numerical form for analysis. The numbers 1 to 7, possibly ranging

TABLE 2.7 A sample semantic differential scale for measuring affective response to nonstandard dialects of English

Directions: Indicate your attitude toward the person whose voice you hear on the tape. Place a mark (X) at the point between each pair of opposing adjectives that best represents your opinion about the person. Please mark each line between pairs only once.





from the negative to the positive characteristic, are assigned to each rating in accordance with the position of the mark on the scale. Here again it may be observed that the positions of positive and negative characteristics have been randomly reversed in Table 2.7 to avoid any "halo" effect. Notice that, unlike the items of the Likert scale, the items of the semantic differential are not additive. Each bipolar dimension must be evaluated separately, unless the dimensions are aggregated statistically as with factor analytic techniques.

Research Applications of the Semantic Differential

Aside from ascertaining attitudes toward speakers of nonstandard dialects in sociolinguistic research, the semantic differential has been used to identify and measure cultural stereotypes, cultural empathy, and proximity to native-speaker attitudes on the part of nonnative speakers of second or foreign languages.

In any of the various applications of this scale, care must be exercised not to alienate the respondents by employing irrelevant adjectives or insisting on *forced-choice* alternatives. With forced-choice alternatives the researcher compels the respondents to express a preference, when no preference may exist. An example of this would occur if we devised a semantic differential with only six points on the scale. Respondents would be forced to show preference toward one end of the scale or the other since no middle ground is available. Sometimes such procedure is warranted, but often it is found to alienate respondents and lose their cooperation. Bear in mind that we are usually already requiring judgments based on very limited information. To add further unreasonable constraints may antagonize the

respondents, who may in turn respond in some haphazard or antagonistic manner that will contaminate the data.

Other Common Rating Scales

A wide variety of other rating scales or schedules have been devised for a multitude of purposes. A few brief examples of such scales are presented in Table 2.8. Consider the various examples of rating schedules or scales presented in this table. Example A illustrates how we might elicit information about the frequency with which certain activities are performed.

Example B is a typical anxiety scale for use in gathering respondents' self-reports regarding level of test anxiety.

Example C presents a simple fluency rating scale. In this example a rating is supplied for the fluency of every person on the list. Of course, such a scale could be employed for rating a great number of other traits or characteristics.

Example D presents a typical classroom observation schedule. Each observed teacher activity is tallied in this way and an activity rate is calculated. The Flanders Interaction Process Analysis (1965) system works according to a similar principle. Only that method, called time-coding, seeks to ascertain what percentage of the time is devoted to each behavior. Also, student behaviors are coded in addition to teacher behaviors. Other coding systems, such as that of Bales (1950), seek to tabulate the frequency of occurrence of each behavior of each person in the classroom. This latter method is called sign-coding as opposed to the time-coding of the former method.

Example E illustrates another common five-point rating scale. In this instance it is being used to elicit ratings of teacher ability and performance.

Another use of an evaluative scale is illustrated in example F. Here we are shown a typical Need-Press Interaction questionnaire. This kind of questionnaire is used to determine whether the instructional press of an academic program is commensurate with the felt needs of the participants in the program. Areas of vast discrepancy between felt importance and perceived emphasis become the focus of curricular reform (cf. chapter ten).

2.4 Scale Transformations

Frequently it becomes desirable to transform one type of scale into another. Such transformations may be necessary in order to make test data more easily interpreted, such as when raw scores are transformed into percentage scores. Or the purpose of the transformation may be to make the test scores more amenable to computation and statistical manipulation, as when raw scores are transformed into normal distribution area proportions. Some statistical analyses assume a normal distribution of the data, and when the actual distribution fails to meet assumptions of normality, certain logarithmic, trigonometric, or other normalization transformations are applied. The two major categories of transformations are *linear* and *normalization*.

TABLE 2.8 Sample rating schedule items for various research purposes

A. Directions: Circle the number corresponding to the frequency with which you perform each activity on any average day.

	Never	Seldom	Occasionally	Often	Usually
1. Listen to English radio broadcasts	1	2	3	4	5
2. Read English books and newspapers	1	2	3	4	5

B. Directions: Indicate with a mark (X) approximately what percentage of the time you experience unpleasant anxiety when sitting for examinations.

5% _____ 25% _____ 50% _____ 75% _____ 100% _____

C. Directions: Rate each speaker for fluency on a scale of zero to 5. (0 = unable to speak; 1 = unable to complete utterances; 2 = halting, simple speech; 3 = frequent pauses to find words; 4 = fluent in most topics, hesitant in a few; 5 = as fluent as a native speaker)

Name	Fluency Rating					
1. _____	0	1	2	3	4	5
2. _____	0	1	2	3	4	5
3. _____	0	1	2	3	4	5

D. Directions: Indicate the duration of the observation period in minutes. Then record each occurrence of the activities listed. Finally calculate the rate per hour of the occurrences by multiplying the total occurrences for each activity by 60 and dividing by the number of observation minutes, as in the formula:

rate = occurrence × 60/observation minutes

Teacher's Name: _____ Class: _____ Date: _____

Observation Period: From: _____ To: _____ Minutes: _____

Teacher Activity	Frequency	Rate
Asks question		
States information		
Uses blackboard or other AV		
Praises student(s)		
Criticizes student(s)		
Corrects mistakes		
Presents assignment(s)		
Responds to question		
Personal interaction		
Listening and observing		

E. Directions: Underline your rating of your teacher on each of the following qualities:

1. Knowledge of the subject:

Below Average Average Good Very Good Outstanding

2. Clarity of explanation:

Below Average Average Good Very Good Outstanding

F. Directions: Rate each of the following components of the program in regard to importance and emphasis. First, indicate its importance to your future success. Next, indicate how much emphasis was given to it in the program. (1 = none at all; 2 = a little; 3 = moderate; 4 = above average; 5 = very great)

Component	Importance	Emphasis
Pronunciation skills	1 2 3 4 5	1 2 3 4 5
Reading skills	1 2 3 4 5	1 2 3 4 5
Spelling skills	1 2 3 4 5	1 2 3 4 5

Linear Transformations

The simplest transformations are termed linear transformations, since they entail altering each score in the distribution by a constant. One example of a linear transformation is the conversion of raw scores to percentage scores. This procedure involves the formula,

$$Y = X \frac{100}{C} \tag{2.4}$$

where, Y = the percentage score
X = the raw score (i.e., number of correct items)
C = the total number of items (i.e., highest possible score)

Here C and 100 are constants applied to X in order to determine Y. A straight line could be drawn on a graph showing every possible Y score corresponding to every given X score.

Another common linear transformation formula, this one used in regression and prediction in chapter five, is,

$$Y = a + bX \tag{5.4}$$

Here a and b are constants used to determine a predicted value of Y for any value of X.

Normalization Transformations

Normalization transformations are used to standardize or normalize a distribution of scores. The most common of these is the z-score transformation discussed earlier in the chapter. By means of this transformation, ordinal data is changed to interval data, and scores are anchored to a norm or a group performance mean as a point of reference. Other interval scales discussed in the chapter also accomplish this end. Size and representativeness of the reference group must always be considered when one is making decisions based on normalized scores.

Fisher Z Transformation

One important transformation is named the Fisher Z transformation after its originator. This is quite distinct from z-score transformations discussed earlier. Fisher Z is used to transform correlation coefficients from a distorted ordinal to a normal interval scale so that arithmetic computations may be done using the coefficients. For example, if one wished to compare the ratio of two coefficients to that of two other coefficients as in the formula below, it would be appropriate to transform the coefficients first, then perform the arithmetic, and finally convert back to correlation coefficients.

$$X = \frac{r_1}{r_2} - \frac{r_3}{r_4}$$

The distortion that results from arithmetic operations involving nontransformed correlation coefficients is negligible when the coefficients are of mid-range magnitude, but becomes great as the coefficients approach unity.

The formula for the Fisher Z transformation is given below.

$$Z = \tfrac{1}{2} \ln (1+r) - \tfrac{1}{2} \ln (1-r) \qquad (2.5)$$

where, Z = the value of Fisher Z
r = the value of the correlation coefficient
ln = the natural or Napierian logarithm

One may transform the coefficient using this formula and making use of a table of natural logarithms, or one may consult Table C in Appendix A where the transformation has been calculated for every value of the correlation coefficient.

2.5 Summary

This chapter has presented important information about scaling in measurement. The four major scales (i.e., *nominal*, *ordinal*, *interval*, and *ratio*) were presented along with numerous examples of each and indications of purpose and methods of computation. Artificial scales used in the construction of questionnaires were presented with examples, including the Likert scale, the semantic differential, and other common rating schedules and scales. Finally, examples were given of scale transformations, including linear, normalization, and Fisher Z transformations.

Exercises

1. Name the four major measurement scales and their purposes, giving an example of each.
2. What is the main advantage of interval scales over ordinal scales in measurement?
3. Identify the scales represented by the following kinds of data:
   Raw scores from a listening comprehension test
   Adjectives on a word-association test
   Percentile scores from a spelling test
   Speed of note-taking in words per minute
   I.Q.-equivalent scores on a vocabulary test
   z-scores on the TOEFL
   The number of instrumentally motivated students in class
4. Compare and contrast the Likert scale and the semantic differential.
5. Name and give an example of two major scale transformations.
6. Given the following raw score distribution on a spelling test of 25 items, calculate (a) ordinal ratings, (b) percentage scores, (c) percentile ranks: 6, 9, 11, 12, 12, 13, 15, 17, 17, 21.
7. If you administered a multiple-choice test of phrasal verb usage to a class of ten students and got the following distribution of scores: 11, 13, 16, 16, 18, 20, 20, 25, 27, 29, what would be the corresponding z-scores, T-scores, and normal distribution area proportions if the standard deviation is 5.95?

8. If you administered a test of idioms to a class of 15 students and got the following raw score distribution: 10, 12, 15, 20, 21, 21, 21, 23, 26, 29, 32, 37, 38, 41, 42, what would be the corresponding stanines, and I.Q.(WISC)-equivalents, if the standard deviation is 10.29?
9. What score would a questionnaire respondent receive if she or he obtained a raw score of 36 on a ten-item Likert scale questionnaire assessing degree of motivation, and you performed a linear transformation to a percentage score?
10. If you performed a Need-Press Interaction analysis of an academic program, and students and faculty reported *mean importance* and *mean emphasis* of vocabulary expansion as 4.5 and 2.3 respectively, what would you probably conclude about the program curriculum?