PLES OF IONAL AND PSYCHOLOGICAL TESTING

THIRD EDITION

Frederick G. Brown

Principles of Educational and Psychological Testing

third edition

Frederick G. Brown
Iowa State University

Holt, Rinehart and Winston
New York Chicago San Francisco Philadelphia
Montreal Toronto London Sydney Tokyo
Mexico City Rio de Janeiro Madrid

Library of Congress Cataloging in Publication Data

Brown, Frederick Gramm.
Principles of educational and psychological testing.

Bibliography: p.
Includes index.
1. Educational tests and measurements. 2. Psychological tests. I. Title.
LB3051.B7636 1983 371.2'6 82-15642

ISBN 0-03-060103-7

Address correspondence to
383 Madison Avenue
New York, N.Y. 10017

Printed in the United States of America
3 4 5 6 038 9 8 7 6 5 4 3 2 1

CBS College Publishing
Holt, Rinehart and Winston
The Dryden Press
Saunders College Publishing

Preface

WHEN preparing the third edition of this book I asked myself several questions: What were my goals for the previous editions? How well were they accomplished? Have some of these goals changed or additional ones been added? In what ways could the book be improved, both in its presentation of current knowledge of educational and psychological testing and as a teaching tool?

As with the previous editions, my primary goal is to provide readers with the necessary background so that they can make informed and critical evaluations of tests (and other assessment methods) when the need arises. Because each testing situation is unique, one cannot say that a particular test is good or bad. Rather, test users must combine their knowledge of the principles of psychological measurement (which are presented in this book) with their knowledge of the test takers, the testing situation, and the proposed use of the test to make a reasoned evaluation of the usefulness of the test in a particular situation. Stated differently, my goal is for readers to learn what questions to ask when evaluating a test or a particular use of a test.

I believe that this goal can best be accomplished by stressing the logic of psychological measurement. By logic I mean the underlying principles that explain why psychological measurement proceeds as it does. Thus we must consider such questions as: Why do we measure psychological characteristics? What types of characteristics can be measured? Why should tests be reliable, valid, and standardized? What assumptions are made when measuring achievement, abilities, and personality characteristics? What do these assumptions imply regarding the nature of the measurement process and the interpretation and use of test scores? Consideration of these "why" questions leads to the "how" questions, those dealing with the methods for constructing and evaluating tests.

I also emphasize the applied aspects of testing, stressing how tests can be used to make educational and vocational decisions and to promote the development of individuals. Although this emphasis, in part, reflects my training as an applied psychologist, what is more important is that applied aspects are stressed because most readers will be test takers and test users rather than test developers or researchers. But test users, as well as test developers and researchers, must understand the nature of the measurement process and the reasoning and assumptions underlying it. As anyone who has followed the controversies over educational and psychological testing—both in the popular press and in professional journals—can attest, there is widespread misunderstanding and ignorance of the na-

ture and purposes of psychological measurement. Only by knowing how tests are developed, and why they are developed as they are, will you be able to make informed judgments about the reasonableness of the various arguments and points of view about the uses (and misuses) of educational and psychological tests.

Readers and instructors may ask how this book differs from others covering similar material. To me, the most distinctive feature is the emphasis on the logic of measurement—the emphasis on the basic principles and assumptions of psychological measurement. As a consequence, I have not attempted to survey or provide an extensive catalog of available instruments; instead, specific tests are introduced only to illustrate a particular principle, concept, or method. Another distinctive feature is the sequential development of topics from the nature of measurement, to consistency, to validity, to interpreting scores. This approach is used to emphasize that unless tests are carefully constructed, measure consistently, and are valid, interpretation of scores will be questionable. Only after the basic principles have been presented are particular types of tests described. Thus we are able to show how the basic principles of measurement are applied to the construction, interpretation, evaluation, and use of various types of tests and assessment methods. Third, I have emphasized certain topics that receive minor consideration in other books; for example, the decision-making approach to validity, base rates, incremental validity, and homogeneity.

A number of techniques have been incorporated to make the book an effective learning tool. All statistics used are described, both in the chapter on statistics (Chapter 3) and as they are applied to specific problems. Examples of statistical analyses are described in the text and illustrated. Statistical concepts are presented both verbally and by formulas. Each chapter has a summary reviewing the major points of the chapter and includes an annotated list of further readings for readers who want to pursue a topic in more depth. Numerous examples of test items, uses of tests, and types of analyses are described, as are guidelines for writing items and interpreting scores. A glossary of important terms is included at the end of the book and these terms are highlighted (by italics) in the text. Instructors may find the *Instructor's Manual* a useful source of test items, problems, and suggestions for class activities.

There are a number of changes from the previous editions. All chapters have been extensively rewritten, both to clarify the presentation and to include discussion of new and revised tests, new methods, and current issues. The two chapters on consistency have been combined into one chapter in order to integrate the material and give a more proper emphasis to this topic. On the other hand, the discussion of typical performance measures has been expanded to three chapters. The chapter on uses of tests has been deleted and the material integrated into other chapters. Information on published tests has been updated and topics of current concern have been discussed more extensively, including content-referenced (criterion-referenced) tests, test bias, competency tests, latent trait approaches, and coaching. And a glossary has been added.

Two comments on stylistic matters. Literature references have been cited in the text only when a point is controversial, to give credit to persons who have

developed a new technique or method, and to help readers locate sources that treat a topic in more depth. (The suggested readings at the end of each chapter also serve this third purpose.) And I have tried to balance male and female referents in examples.

Preparation of this edition was greatly helped by the thoughtful, detailed comments of Dr. John T. Cowles, Dr. Dale B. Harris, and Professor Peter Prunkl, who reviewed the manuscript in depth. I express my appreciation for their contributions.

Ames, Iowa
October 1982

F. G. Brown

Contents

• part one • measurement in psychology and education

MEASUREMENT pervades all aspects of our lives. Your birth certificate lists the date and time of your birth, and your birth length and weight—all measured characteristics. Every day you encounter measurements: when you receive a score on a test, when you price articles in a store, when you select clothes, and when you determine the distance you will travel on your vacation, to name but a few examples. Magazines and newspapers are filled with measurements—inflation and crime rates, athletic statistics, the cost of living index, and election returns. Science, of course, depends heavily on measurements. Perhaps the most famous equation of all time, Einstein's $E = mc^2$, shows the relationship among three measured quantities.

Educators, psychologists, and other behavioral scientists also make extensive use of measurements. Their focus, however, is not on economic or physical variables. Instead they measure abilities, achievements, aptitudes, interests, attitudes, values, and personality characteristics. These measurements are used for purposes such as planning and evaluating instruction, selecting workers and assigning them to jobs which match their abilities and interests, placing students in courses, counseling and guidance, and studying differences between groups and the nature and extent of individual differences. Underlying these uses of measurement is the belief that accurate information about characteristics of individuals is necessary for effective planning, decision making, and evaluation.

This book is concerned with the measurement of individual characteristics and focuses on one particular method, the psychological test. This first part is an introduction to and an overview of the book. Chapter 1 describes the nature of psychological measurement, contrasts psychological and physical measurement, and defines what is meant by a test. Chapter 2 describes the essential characteristics of a good test and the steps in constructing good tests. Inasmuch as measurement is

a quantitative endeavor and relies on statistical analyses, Chapter 3 reviews the descriptive statistics needed to understand the quantitative concepts used in testing. The topics introduced in these three chapters provide the foundation for the material discussed in the remainder of the book.

Chapter 1 • The Nature of Psychological Measurement

THAT people differ in numerous ways is apparent even to the most casual observer. Some people are short, others tall; some have blue eyes, others brown; some can run 100 yards in less than 10 seconds, others take 15 seconds or longer. People differ not only in physical appearance and skills, but in their abilities and personality characteristics as well. Some people find math easy, whereas others are unable to do simple calculations; some people are extroverted and outgoing, whereas others are quiet and retiring; some people emerge as leaders, whereas others are content to follow.

Because of these wide variations, the study of individual differences has been a continuing focus of interest in psychology. Part of this interest reflects our natural curiosity about other people—about what they are like and how they behave. Thus, in psychology as in everyday conversation, most terms used to describe people refer to characteristics that vary widely between people, such as intelligence, aggressiveness, mathematical ability, flexibility, and creativity, to name but a few. These differences also have practical consequences, perhaps the most obvious being that different skills and abilities are required in various occupations.

Psychologists interested in individual differences are concerned with a number of questions: In what ways do people differ? How large are these differences? How can these differences be most accurately measured? What are the practical implications of these differences? These are the types of questions addressed in this book. Although the focus will be on the identification and measurement of individual differences, the ultimate goal will be to apply this knowledge to help individuals live more satisfying and productive lives.

why measure psychological characteristics?

The goal when measuring psychological attributes—be they abilities, skills, interests, attitudes, or personality characteristics—is to obtain an accurate description of an individual or group of individuals. For some people the primary goal is to understand more about the nature and range of individual differences, as when a psychologist studies the range of intellectual abilities in a particular population. For others the goal is to understand and predict behavior more effectively. For example, a teacher might measure students' mathematical skills in order to determine what abilities are needed to solve problems and what teaching method will be most appropriate for each student. Others will be interested in making practi-

cal decisions, such as which applicants to hire for a position. In each situation, accurate information about abilities or personality characteristics is needed.

As you can see, there are various reasons for measuring the characteristics of individuals. To give a flavor of some of the common uses of psychological measurement, I will briefly describe several applications of psychological measurement. In these examples the characteristics of interest are measured by tests; keep in mind, however, that tests are only one of several possible methods for measuring psychological characteristics.

Descriptive Uses

In some situations tests are used to provide descriptions of an individual. For example, tests are frequently administered during counseling to provide clients with descriptions of their abilities, interests, or personality characteristics. Or an elementary school teacher may be interested in the abilities, knowledge, and study habits of his students. In other situations we may want to describe a particular population; for example, the academic ability of students attending Androscoggin College. (In turn, we might use these data to compare an individual student's ability either with that of other students attending Androscoggin or to compare the students at Androscoggin with students at other colleges.)

When using tests descriptively, we generally are interested in using the information at more than one time and for various purposes. Thus descriptive information is of most value when there is continuing contact between the test user and test taker over an extended period of time. For example, a classroom teacher can use students' scores on ability and achievement tests in a variety of interactions with students throughout the school year, not just at one particular time. Or a counselor may use test scores to help a client select courses, plan a program of study, improve study methods, or make a career choice. In other words, use of the test data is not restricted to one particular decision. Thus, descriptive uses generally involve measurement of broad abilities and characteristics, often ones that are relatively stable over time.

Decision-Making Uses

In contrast to the broad-based assessments just described, tests are also used as an aid in making specific, practical decisions. These decisions may involve individuals, such as who should be hired for a particular job or admitted to medical school, or groups, such as which of two methods of teaching physics results in higher achievement. In either case a decision must be made and the test scores are used as one basis for making the decision. Thus, in contrast to descriptive uses, the test content will usually be narrower and more focused on the specific situation.

The following paragraphs describe some common decision-making uses of tests.

Selection. Selection decisions are common in both academic and business settings. In selection there are more applicants than can be accepted or hired and the decision is which one (or ones) to accept. To illustrate, if a medical school has

places for only 100 students in its first-year class and 800 students apply, the admissions committee must decide which applicants to admit and which to reject. The role of a test in this situation is to help identify the most promising applicants—that is, those with the highest probability of success.

Placement. In placement, or classification, there are several individuals and several alternative courses of action; for example, there may be several academic tracks, training programs, or jobs, and each person is to be assigned to one alternative. The goal is to match individuals and alternatives in an optimal manner. Examples include using tests to assign recruits to occupational specialties in the armed services or college freshmen to various levels within a sequence of French courses.

Ranking. An illustration of the use of tests for ranking is grading. In most courses, grades must be assigned. Frequently all, or a large part, of the grade is determined by performance on examinations. The scores on the various tests and assignments are then combined, students are ranked in order of their achievement, and grades assigned on the basis of these ranks. In this situation, the decision involves which grade to assign to each level of performance.

Proficiency. Tests can also be used to establish proficiency. A familiar example is the examination required to obtain a driver's license. This examination usually consists of both a written test (on traffic laws and regulations) and a behind-the-wheel driving test. You must perform at a certain minimal level in order to obtain a license.

Other examples of proficiency tests include licensure exams in professions such as law and medicine, Red Cross swimming and lifesaving tests, and minimum competency tests for high school graduation. "Test-out" examinations in schools can also be considered as proficiency exams in that students who demonstrate the desired level of proficiency either receive credit for the course and/or are exempt from taking it.

The distinguishing feature of proficiency testing is that a minimal level of performance is specified in advance; the decision involves determining whether performance meets this standard. Usually there are only two possible outcomes: pass or fail.

Diagnosis. Diagnosis involves comparing an individual's performance in several areas in order to determine relative strengths and weaknesses. Generally, diagnostic procedures are instituted when an individual is having difficulty in some area. Once the areas of disability are identified, a program of remediation can be undertaken. Thus a diagnostic reading test might provide scores in phonetics, word meaning (vocabulary), sentence meaning, paragraph meaning, and reading rate. Here the goal would be to identify the student's particular weaknesses and strengths. Once the specific area of disability is identified, a program of remedial help can be arranged.

Evaluation. The previous examples all concerned decisions about individuals. But test scores can also be used to evaluate educational programs, treatments (such as alternative types of therapies), or procedures. Consider the example introduced at the beginning of this section, evaluating the effectiveness of two different methods of teaching physics. We could design an experiment where some students take a physics course taught by one method and others are taught by a second method. At the end of the term, both groups would be given the same final examination. By comparing the scores of the two groups of students on this exam, we could determine which course produced higher achievement.

To review, consider the decision made in each situation. In selection, the decision is whether to select or reject each applicant; in placement, which alternative course of action to instigate; in ranking, which grade or category to assign; in proficiency, whether or not the person attained the minimal level of performance; in diagnosis, which treatment to use; in evaluation, which method is more effective. Since no test (or other assessment method) is 100 percent accurate, the primary question is not whether the test provides accurate or inaccurate information but whether it increases the number the correct decisions made.

Note also that we referred to tests as decision-making "aids" because a test is usually only one of several elements in the decision-making process. Rarely are decisions made solely on the basis of test scores. Rather, all available relevant information should be used when making a decision. As stated in the *Standards for Educational and Psychological Tests:*

> For most purposes, . . . a single assessment or assessment procedure rarely provides all relevant facets of a description. . . . Decisions about individuals should ordinarily be based on assessment of more than one dimension; when feasible, all major dimensions related to the outcome of the decision should be assessed. . . (*Standards*, 1974, pp. 61–62)[1]

Because human behavior is complex, many factors enter into a decision; test scores are only one. Thus, when tests are used as decision-making aids, accurate test data alone will not guarantee accurate decision making; other factors and variables that contribute to the decision will also affect the quality of the decision.

Defining Constructs

In addition to descriptive and decision-making uses, tests can also aid in the development of educational and psychological theories. Theories in education and psychology, like those in other disciplines, involve both constructs and laws (Kerlinger, 1973). One way of operationally defining constructs is by tests. Suppose, for example, we are interested in anxiety. One way to measure anxiety would be to ask people to rate their degree of anxiety in various situations. The score on this test (the sum of these ratings) would be the operational definition of anxiety.

To validate a theory we conduct experiments. One possibility would be to

[1] Literature references can be found in the bibliography at the end of the book. For brevity, the *Standards for Educational and Psychological Tests* will be referred to as the *Standards*.

study the effects of anxiety on problem solving. We might hypothesize that the more anxious a person is, the poorer his or her performance on problem-solving tasks will be. To test this hypothesis we could administer an anxiety test to a group of subjects, administer the problem-solving tasks, and then see whether the predicted relation held. If it did, we would have more confidence in our theory. By conducting other studies we could increase our knowledge of the construct of anxiety and whether our test was a good measure of anxiety.[2]

The previous example illustrates the use of psychological measurement in *hypothesis testing*. That is, the anxiety measure was used to test a hypothesis about the relationship between anxiety and problem solving. In this example the anxiety measure was the independent variable, since we tested the effect of anxiety on problem solving. We could, by designing a different study, have used the anxiety measure as a dependent variable. For example, we could study whether anxiety test scores increase under stressful conditions.

Suggesting Hypotheses

Tests can also be used for *hypothesis building*. This use is illustrated by surveys and the use of tests in counseling. Suppose that a survey shows that persons living in a certain part of the United States score lower on achievement tests than persons of comparable age and education in other states. Why? Several hypotheses could be developed. One is that the people living in that area are less intelligent. Another is that the results reflect differences in the quality of education. A third hypothesis might attribute the effects to socioeconomic and cultural differences. These hypotheses could then be checked by further studies.

Counselors and therapists often use test results to build hypotheses about their clients. Suppose tests show that Clarence, whose father is an engineer, has excellent scholastic aptitude, is interested in literary and artistic activities, is submissive, and has conflicts with authority figures and a problem with family relationships. He is now enrolled in engineering and is failing. From the test data the counselor might hypothesize that Clarence is in engineering because of parental pressure. Unable to confront his parents directly with his dislike for engineering, he has chosen the indirect method of failing his courses. This hypothesis could then be checked through further interviews.

how are psychological characteristics measured?

A second question is: How can psychological characteristics be measured? For most characteristics, a wide variety of methods are available. Suppose, for example, we are interested in the mathematical ability of high school students. Several indices of ability come readily to mind: grades in math courses, scores on math tests, teachers' ratings, self-ratings, and performance in courses or situations requiring use of mathematics. (You can probably think of others.) Or, if we are interested in anxiety, we could have observers rate each person's level of anxiety, have people rate their own anxiety level or answer test items asking about their

[2] For further discussion see the section on construct validity in Chapter 7.