ESSENTIALS OF PERSONNEL ASSESSMENT & SELECTION

Scott Highhouse ▪ Dennis Doverspike ▪ Robert M. Guion

SECOND EDITION

ROUTLEDGE

ESSENTIALS OF PERSONNEL ASSESSMENT AND SELECTION

This second edition continues in the tradition of the first edition by giving managers and students the nuts and bolts of assessment processes and selection techniques. The book provides current and future managers with the knowledge and tools required to make informed personnel decisions based upon the results of tests and assessments. It emphasizes that good prediction requires well-formed hypotheses about personal characteristics that may be related to valued behavior at work and the need for developing a theory of the attribute one hypothesizes as a predictor—a thought process too often missing from work on selection procedures. In addition, it explores such topics as team-member selection, situational judgment tests, nontraditional tests, individual assessment, and testing for diversity. The book covers both basic and advanced concepts in personnel selection in a straightforward, readable style intended to be used in both undergraduate and graduate courses in Personnel Selection and Assessment.

Scott Highhouse is a Professor and Ohio Eminent Scholar in the Department of Psychology, Bowling Green State University, USA. Scott is Founding Editor of the journal *Personnel Assessment and Decisions* and serves on the editorial boards of *Journal of Applied Psychology* and *Journal of Behavioral Decision Making*.

Dennis Doverspike is a Full Professor of Psychology at The University of Akron, USA, Senior Fellow of the Institute for Life-Span Development and Gerontology, and Director of the Center for Organizational Research. He is certified as a Specialist in Industrial-Organizational Psychology and in Organizational and Business Consulting Psychology by the American Board of Professional Psychology (ABPP) and is a licensed psychologist in the State of Ohio.

Robert M. Guion (deceased) was Distinguished University Professor Emeritus at Bowling Green State University, where he was on the faculty from 1952 until his death in 2012. Honors include the Distinguished Scientific Contributions Award, Society for Industrial and Organizational Psychology; Award for Lifetime Contributions to Evaluation, Measurement, and Statistics, American Psychological Association (Div. 5); and the Stephen E. Bemis Memorial Award, International Personnel Management Association Assessment Council.

This page intentionally left blank

ESSENTIALS OF PERSONNEL ASSESSMENT AND SELECTION

Second Edition

Scott Highhouse, Dennis Doverspike, and Robert M. Guion

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Second edition published 2016
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

First edition published by Lawrence Erlbaum Associates, Inc., 2006

Library of Congress Cataloging-in-Publication Data
Names: Guion, Robert M., author. | Highhouse, Scott, author. | Doverspike, Dennis, author.
Title: Essentials of personnel assessment and selection.
Description: Second edition / by Scott Highhouse, Dennis Doverspike, and Robert M. Guion. | New York, NY : Routledge, 2016. | Earlier edition published in 2006 written by Robert M. Guion and Scott Highhouse. | Includes index.
Identifiers: LCCN 2015038976| ISBN 9781138914575 (hardback : alk. paper) | ISBN 9781138914599 (pbk. : alk. paper) | ISBN 9781315690667 (ebook)
Subjects: LCSH: Personnel management—Decision making. | Prediction of occupational success. Employees—Rating of. | Employment tests.
Classification: LCC HF5549 .G793 2016 | DDC 658.3/11—dc23
LC record available at http://lccn.loc.gov/2015038976

ISBN: 978-1-138-91457-5 (hbk)
ISBN: 978-1-138-91459-9 (pbk)
ISBN: 978-1-315-69066-7 (ebk)

Typeset in Bembo
by Apex CoVantage, LLC

It is a fine thing to have ability, but the ability to discover ability in others is the true test.

—Elbert Hubbard (1856–1915)

This page intentionally left blank

CONTENTS

PREFACE

Robert Guion wrote a book, *Personnel Testing,* published in 1965, which was used as a textbook in undergraduate and graduate courses in testing and selection. A second book was later undertaken to be a reflection of changes in assessment methods and in selection problems that occurred subsequent to that first book, and it was also intended to be a textbook. That book, published in 1998 (2nd ed., 2011) as *Assessment, Measurement, and Prediction for Personnel Decisions,* had a much longer title; moreover, in an effort to be comprehensive, its content was also longer and more complex. It turned out to be more appropriate for professionals in the field, and those industrial and organizational psychology students preparing to become professionals, than for undergraduate students or master's students preparing for broader HR roles.

The first edition of this book, *Essentials of Personnel Assessment and Selection,* distilled from the bigger book the essentials that managers and other well-educated people should know about the assessment processes so widely used in contemporary society—and so widely not understood. By most accounts, the book succeeded as a text for advanced undergraduates and master's level students interested in becoming users of research-based assessment and selection information and techniques.

It is now 10 years later and much has changed. Robert Guion is no longer with us. He passed away on October 23, 2012, at the age of 88. Bob was a model of integrity and deeply believed that the waste of human resources should pain the professional conscience of I-O psychologists. He worked tirelessly toward the development of a fundamental science that promotes human welfare at work. We are humbly moving forward with this *Essentials* text—which Bob made clear was his wish.

Like the earlier edition, this one emphasizes that good prediction requires well-formed hypotheses about personal characteristics that may be related to valued

behavior at work. We continue to emphasize the need for developing a theory of the attribute one hypothesizes as a predictor, a thought process too often missing from work on selection procedures. New to this book is increased attention to topics such as managerial and executive assessment, advances on the legal front, and global testing, as well as technology and testing. We also consider topics that were not of much concern in 2006, such as unproctored online assessment and "big" data. Considerable attention was also given to updating the book to incorporate recent research findings. Realizing that professors who use our book as a textbook prefer not to make major changes to their syllabi, we have made only one major revision, switching the order of Chapters 11 and 12, so as to discuss ratings after we complete our discussion of other types of assessments.

Although we have updated the book in some respects, we also have tried to stay true to the original vision of Robert Guion. In particular, in the first edition, Bob emphasized the philosophical and historical basis behind personnel selection. He included a good deal of research reflecting the origins of personnel selection. Therefore, the current edition continues to reflect the work of many of the early innovators in the field of personnel selection.

As in the first edition, our goal was to produce an accessible guide to assessment that covers basic and advanced concepts in a straightforward, readable style. Evaluating job candidates is an emotional topic, fraught with unsubstantiated claims from test publishers and baseless accusations from social critics. This book provides a review of the most relevant statistical concepts and modern selection practices that will equip readers with the tools needed to be competent consumers of assessment procedures and practices, and to be well-informed about the kinds of questions to be answered in evaluating them.

Finally, we would like to acknowledge the help of people who contributed their time and effort to make this book as good as we hope it is. A lot of people helped by critically reading parts of the earlier 2006 book. They include Neil Christiansen, Fritz Drasgow, Timothy Judge, Fred Oswald, and Charlie Reeve. A special thanks goes out to Catalina Flores, a graduate student at The University of Akron, who assisted with many of the administrative and editing tasks. On a personal note, Scott Highhouse would like to express his gratitude for distractions from his wife, Maggie, and their five kids: Carmen, Cole, Baye, Owen, and Willow. Dennis Doverspike would like to thank his wife, Ida, and sons, Dan and Tom, for keeping him centered and alive.

Thanks again to all of you.

—Scott Highhouse (Bowling Green State University) and
Dennis Doverspike (The University of Akron)

PART I

Deciding What to Assess

This page intentionally left blank

1

UNDERSTANDING PERSONNEL ASSESSMENT

Assumptions of This Book, Validation and Its Limits, and Theory and Practice

- In 1921, applicants who answered a job advertisement anonymously posted by the world-famous inventor Thomas A. Edison arrived at the Menlo Park facility only to find that they needed to answer a series of brainteasers such as "Is Australia larger than Greenland in area?" "If you were to inherit $1,000,000 within the next year, what would you do with it?" and "How is leather made?"
- Nearly 100 years later, applicants who made it through the initial screening process for a job with an Internet superstore were subjected to a grueling interview that included such oddball questions as "Why is a tennis ball fuzzy?" "Why are manhole covers round?" and "How many cows are in Canada?"

As these anecdotes show, employers are constantly inventing (or recycling) innovative methods for attempting to figure out if a job applicant has what it takes to succeed in their firm. What is vastly different between the two examples above is the public's reaction to such innovative methods. The public reaction to Edison's questions was almost uniformly negative (Dennis, 1984). The *New York Times* published 23 articles about the Edison questions in one month alone. Most of these articles ridiculed Edison for attempting to assess the fitness of job candidates with outrageous questions ("More Slams at Edison," May 22, 1921). Today, companies such as Microsoft, Zappos, and Xerox are praised for using brainteaser interview questions, presumably because they enable candidates to provide atypical responses and demonstrate their creativity (e.g., Fuscaldo, 2014; Poundstone, 2012). Despite this, there is no evidence that such methods have any utility for predicting future job performance. For instance, the senior vice president of "people operations" at Google commented, "On the hiring side, we found that brainteasers are a complete waste of time . . . They don't predict anything. They serve primarily to make the interviewer feel smart" (Bryant, 2013).

Brainteaser questions are just one example of how employers often become enamored by their personal theories of what good applicants should be like in order to be successful at work. We believe that personnel assessment in practice will not be taken seriously by upper management until the people who use it become serious advocates for tests, acknowledge and master the complexities of selection, and thoroughly and persistently communicate the utility of using sound methods to reach decisions to key stakeholders.

Human Resource (HR) managers need to make a case to upper management for giving employee selection as much research and development (R&D) attention as is given to patent development. Staffing courses need to give the science of employee selection as much attention as they give to designing performance management systems or strategizing about human capital. Getting a "seat at the table" is about proving to management that you can find diamonds in the rough, using state-of-the-art techniques in performance prediction. It is not about talking the right business lingo or rejecting proven methods as old-fashioned.

Wise Decisions

An organization functions through its members. New members are chosen in the belief that they will benefit the organization. Employees benefit the organization by accepting fairly specific organizational roles—fairly specific sets of functions, duties, and responsibilities. When existing members of an organization seek a new hire for a designated role, the dominant consideration is the suitability of the candidate for that role. Once in the organization, a person may keep the original role, be transferred or promoted, be trained for a somewhat changed role, or be terminated. All are personnel decisions. All are based, if the organizational leaders are not too whimsical and impulsive, on some sort of *assessment* of the person. Organizational decision makers hope to make wise decisions and competent assessments help.

Results of wise decisions can range from the mere absence of problem hires to the acquisition of genuine superstars, or top talent, who promote organizational purposes. Good hiring decisions can result in substantial increases in performance levels and productivity. Consequences of unwise decisions can range from inconvenience to disaster. An examination of past U.S. presidential elections or NFL draft choices can provide ready examples of good and bad hiring decisions.

Wisdom in selection decisions depends greatly on knowing the characteristics that are truly important in an anticipated role and on not being distracted by irrelevant characteristics. Assessing relevant characteristics may be as easy as looking at a driver's license and noting whether it is current, but most are more abstract and harder to assess. If it is inferred from job analysis that qualifications include skill in getting along with others, that skill might be assessed in an interview, or from personal history information, but special efforts are needed to be sure that these assessments provide valid information related to future behavior on the job. Many qualifications are best assessed by tests or specially developed work samples.

This book emphasizes work organizations and how they may improve the chances that their personnel decisions will be wise ones. Wisdom in decision making is elusive; there are opposing points of view about what is wise, desirable, and valued. In this book, we want to state our view explicitly and assist managers in refining and analyzing their own philosophy toward decisions concerning human resources.

Organizations exist when people join forces voluntarily to reach a common goal; they earn their existence by producing goods or services valued in at least a segment of the larger society. An organization, therefore, prospers according to its contribution to society (Eels & Walton, 1961), and individual members contribute by functioning well in their assigned roles. The interests of the consumers of the goods or services are compromised, no less than the personal interests of those in the organization, when a person who can function very well is denied a position given to one less qualified. Enough multiplication of such selection errors, and the organization fails—with resulting human and economic waste. If there are more applicants than openings, choices must be made. Choices could be random, or quasi-random, like "first come, first chosen." Choices might be based on social values, giving preference to veterans, women, or minorities. The choices might be based on nepotism, prejudice, or a similar-to-me bias. Or they can be based on the science of selection and result in the proven prediction of future performance.

We believe the principal basis for personnel decisions should be *merit*. Some people reject merit as elitist. Some consider profit-oriented concepts of merit inimical to the interests of a broader society. Some dismiss the idea of merit in the belief that situational factors (e.g., having a good boss) influence work performance more than the personal characteristics people bring to the job. If the merit principle is accepted, however, methods for establishing relative merit are needed. We prefer psychometric methods that give standardized, even-handed assessments of all candidates, similar results from one time or situation to another, and demonstrable relevance to performance.

The term ***psychometric*** results from the combination of two Greek words and, literally translated, means "measurement of the mind." The psychometric approach involves developing imperfect indicators of some underlying concept. They are imperfect because they are subject to measurement error.

It is wasteful to deny qualified people employment for invalid reasons, including whims known only as "company policy." Wasting human resources is as inexcusable as wasting physical resources. An organization has a responsibility to itself, to

the society that supports it, and to the people who seek membership in it, to be sure that it conserves and optimizes human talent.

The Role of Research in Staffing Decisions

The history of assessment for personnel selection is old. The ancient Chinese developed civil service examinations (Bowman, 1989; DuBois, 1970). Plato devised procedures for selecting the Guardians in his Republic. Another example is Biblical. Gideon had too many candidates for his army. On God's advice, he used a two-stage personnel testing procedure. The first was a single-item preliminary screening test ("Do you want to go home?"); on the basis of the answers, he cut 22,000 candidates down to 10,000. A behavioral exercise—to observe candidates drinking from a stream—was used for those remaining; 300 were chosen. No one questioned the validities of these procedures for they were given by God. Unfortunately, many contemporary testers behave as if they believe that they, too, have God-given tests and do not need to worry about research evidence. Selection researchers, however, recognize that tests and interpretations of results are fallible and that the validity of any given procedure for assessing candidate characteristics needs to be questioned. Such questioning has led to fairly standard procedures for evaluating (validating) selection procedures.

Fundamental Assumptions

Freyd (1923) identified five assumptions that were fundamental to the research process. With some updating, they are also fundamental to this book:

1. People have abilities and other traits: mental abilities, psychomotor abilities, knowledge, specifically learned skills (including social skills), and habitual ways of dealing with things and events (including personality or temperament). We do *not* assume that traits are permanently fixed, either by heredity or early life experiences. We do assume, however, that some of them, especially abilities, are reasonably stable for most adults, stable enough that the level of ability observed in a candidate will stay pretty much the same for some time. Thus, even if traits or characteristics cannot be directly observed, they can be inferred on the basis of their effects and are, thus, real. Psychometricians often refer to the existence of underlying *latent traits*.
2. People differ in any given trait. Those with higher levels of abilities relevant to the performance of a job are expected to perform better, other things being equal, than those with lower levels. Thus, individual differences exist on traits and characteristics.
3. Relative differences in ability remain pretty much the same even after training or experience. People with higher levels of a required ability before being hired will be the better performers on that job after training or after a period of time has passed.

4. Different jobs require different traits. For example, one job may require specialized mathematical skills; another may require conscientious attention to procedural detail.
5. Required abilities can be measured. Cognitive abilities, for example, can be measured with many different kinds of tests. Not only can traits or abilities be measured, but the resulting scores or numbers have some real mathematical meaning.

Cognitive tests have been used successfully for employee selection and for many other purposes. The measurement of *motivational* requisites of successful performance has a less impressive record of success in employee selection. The record may be more impressive when the research effort expended on the definition and measurement of such traits approaches that expended on cognitive abilities.

Steps in Traditional Validation

Personnel research has traditionally focused on jobs that employ large numbers of people. For such jobs, traditional employment test validation follows steps like these:

Analyze Jobs and Organizational Needs. These procedures are sometimes casual, sometimes very systematic (see Chapter 2). Both job and organizational need analysis inform judgments of whether the need is for improved selection or some other sort of organizational intervention, such as redesigning the job or training current employees. Clearly, no new selection procedure can solve a problem that springs primarily from inadequate equipment or inept management.

Job analysis asks what a worker does, how it is done, and the resources (personal and organizational) used in doing it. Jobs are analyzed to get enough understanding of the job to know what applicant characteristics are needed to perform it effectively.

Choose a Criterion. The criterion in personnel research is that which is to be predicted: a measure of performance, of a limited aspect of performance, or of some valued behavior associated with the assigned job role. It might be a measure of trainability, production quality and quantity, attendance, or something else. Criterion choice is a matter of organizational values and organizational needs.

The ***predictor*** is what we use to assess the job candidate's (future) suitability for the job. The ***criterion*** is the thing we use to assess the employee's (current) performance on the job. If we used a test of personality to predict number of sales made by sales associates, the predictor would be the test of personality, and the criterion would be number of sales. ***Validation*** is the process of estimating the relationship between the predictor and the criterion.

Form Predictive Hypotheses. More than one kind of ability or trait likely must be measured if the criterion is to be predicted in all of its complexity. Each predictor–criterion pair is a hypothesis open to research (see Chapter 3). For example, an analysis of the job of potato chip sorter may have revealed that chip quality is an important work outcome to be predicted. One predictive hypothesis might be that individual differences in attention to detail should be related to better performance in monitoring chip quality. A predictive hypothesis may be rather casual and still prove to be a good one. More systematically developed, well-reasoned hypotheses ordinarily will be more likely to be supported by research.

Select Methods of Measurement. We tend to have more research on tests and questionnaires than on other methods—for good reasons. Practical research follows success, and the predictive value of tests has been demonstrated more persuasively and more frequently than for competing approaches to assessment. Further, testing is easily standardized, enabling a fairer assessment than is possible when the method of assessment varies from one person to another (as with an unstructured job interview). Test use is not, however, free from problems. One serious problem is the tendency to assess candidates only on traits for which tests are available, rather than to assess characteristics (such as interpersonal skills) not easily assessed by available testing procedures.

Design the Research. Good research tries to ensure that findings from the research sample can generalize to the population of interest, which is job applicants. One aspect of research design is the choice of research participants. Inappropriate participants may spoil the generalizability of results. In particular, incumbents and applicants may differ in motivation to do well on a test, in means and variances on the measured predictors, or in demographics. Demographic *diversity* has become a watchword in organizational staffing. The research implications of tapping currently underused sources of job candidates in the search for diversity must be monitored carefully.

When the complexity of criterion performance calls for multiple predictors, some means of considering the predictors in combination is needed. Considering them in combination requires a choice of methods for forming a composite, and it is that composite of predictors that is to be evaluated. Sequential approaches to selection call for some rules for advancing from one step to the next. Any composite or sequence anticipated in operational use should be the composite or sequence used in research.

Collect Data. Predictors must be administered with both standardization and tact. The first of these is technical; the second is both technical and civil. Standardization of assessment procedure has long been accepted as a sine qua non of good practice; it has been virtually unquestioned throughout most of the history of

personnel selection research. Everyone who is tested is given the same set of items, identically worded; any established time limits are rigidly followed whenever the test is given, and instructions are the same for everyone. With that said, appreciating the apprehension of people being assessed is important. Standardization does not mean treating people in a way that is not courteous and respectful.

Evaluate Results. Freyd (1923) referred to evaluating measurement; the idea subsequently became known as *validating* the predictor as measured. Whether called evaluation or validation, the traditional procedure has been to correlate scores or ratings on predictor variables with numerical values on criterion measures. If the correlation is high, the predictor is said to be a good one (i.e., a *valid* one), and if the correlation is low, the predictor is said to be poor. High and low are relative terms, evaluated more against experience than against specified numbers. In employment testing, empirical evaluation of predictions traditionally has been deemed essential.

The tradition of empirical validation needs to be qualified in light of views developed later in Chapter 5. An even older psychometric tradition defines *validity* as how well the predictor (usually a test) "measures what it purports to measure" (Drever, 1952, p. 304). These views of validation are not the same. A test that purports to measure spelling ability may do so very well, but it is not likely to be very good at predicting how well mechanics repair faulty brakes. For this reason, we distinguish between the validity with which a trait or attribute is measured and the validity with which the measured trait predicts something else—between validity of measurement (psychometric validity) and validity as the job-relatedness of a predictor. Evidence for either concept of validity may be collected by any of several forms of empirical investigation.

Validation Designs

From the early days of employment testing, validation has followed one of two basic design methods: the *present employee* method, studying people already on the job, or the *follow-up* method, testing job applicants and getting criterion data later for those hired. The follow-up method is widely (but not universally) considered the better design because it tests actual applicants.

In an idealized follow-up design, sometimes called the *Cadillac* version, the tests are given to all applicants but not scored until criterion data are available for those who are hired. (This is to ensure that neither employment decisions nor subsequent criteria are affected by knowledge of the test scores.) Decisions are made as if the tests were not available at all, using existing methods—application forms, interviews, references, tests, hunches, or whatever—whether previously validated or not. After a time, criterion data are collected for those hired; the tests are then scored, and the scores are compared to criterion data.

In the early days of employment testing, such ideal data collection procedures were rare; now they are virtually nonexistent. Nevertheless, the ideal provides a standard against which other designs can be discussed. Traditionally, the only other option was the present employee method where employees are taken off the job, tested, and the test scores are correlated with existing or concurrently obtained criterion measures. It is a faster method, and practical considerations often seem to favor it.

The two different approaches are referred to as "predictive" and "concurrent" research designs. These terms distinguish time spans for data collection, not the employment status of the research subjects. ***Predictive designs*** include a substantial time interval between the availability of predictor data and collection of subsequent criterion data; in ***concurrent designs,*** both are collected at about the same time. Thus, a predictive design may use present employees if the data to be evaluated can be collected from them at one time and criterion data collected some weeks or months later.

Does it matter whether the research design is concurrent or predictive? Opinions differ. Barrett, Phillips, and Alexander (1981) argued that the importance of the issue has been exaggerated. Acknowledging that the design differences are potentially important, they presented arguments to show that the differences do not, in fact, have much impact on the results of studies. If anything, concurrent studies generally have given somewhat larger correlations (e.g., Gupta, Ganster, & Kepes, 2013). Moreover, abilities are enhanced through job training and experience; people who do well on the job develop their abilities more than do those who do less well.

Concurrent and predictive designs are all variations on a single theme: the correlation between a predictor and a criterion. Validation research is not limited to that theme. This book considers other designs and considerations for assessing not only job-relatedness as an aspect of validity but also for assessing the meaning of scores on an assessment procedure. Because a predictor–criterion correlation is the traditional meaning of a "validity coefficient," it serves as a way to introduce the problems and complexities of validation, but it is only an introduction.

Problems With Traditional Research

This recital of traditional personnel research is quite conventional, but it describes a paradigm that needs to be reexamined. It is subject to several potentially serious problems.

Numbers of Cases. Conventional research needs large numbers. "Large" once meant 30 or more; considerations of power in evaluating statistical significance have shown that "large enough" may require hundreds of research subjects. The power of statistical tests depends on the statistic. Generally, the more complicated the statistical analysis, the larger the sample needed. Major changes in the U.S.

workforce have occurred and seem likely to continue. Most people do not work in large corporations on jobs performed by hundreds of coworkers. Technological growth has produced a wider variety of jobs. Many employment decisions must now be made where only a few people are to be hired (perhaps only one) from a relatively small group of candidates. Further, more hiring is being done in professional, semi-professional, and managerial occupations, where one person must be chosen from perhaps as few as a half-dozen candidates. In short, the numbers for many decisions are too small for reliable correlation coefficients (i.e., less than 100). The traditional paradigm makes no provision for the small business, for choosing the replacement for a retiring manager, or for hiring a one-of-a-kind specialist.

Consideration of Prior Research. Traditional validation ignores prior research. Earlier, it was thought that validities were unique, specific to a situation at hand. Now it is known that validities often generalize well across different situations (see Chapter 7).

Need for Judgment. The traditional approach to selection is purely statistical; it leaves no room for judgment. In one sense, that is good. The idea that human judgment yields better predictions than statistical equations do is a myth (or a superstition based on hope) persisting in spite of overwhelming evidence to the contrary. Nevertheless, statistical prediction is often impossible, infeasible, or insufficient; judgment is necessary (see Chapter 8). Even with research, the circumstances for a candidate at hand may differ enough from the research circumstances that use of the research is questionable. The most obvious example lies in testing the skills of people with disabilities. One cannot intelligently (or legally in the United States) refuse to consider a blind applicant for a job in which visual acuity is not a genuine requirement just because the applicant does not match the research sample of people with sight. One can, of course, make some modification of the selection procedure (such as reading items orally), but the research does not apply to these nonstandard modifications (see Chapter 4). The decision maker must, therefore, make a judgment based on the applicant's performance on a procedure of unknown validity, on interviewer judgments of unknown validity, prior work experience of unknown validity, or on a random basis known not to have any validity. To disqualify an applicant because the possible assessment procedures have not been validated is not very wise.

Global or Specific Assessments. A guiding theme of this book is that a predictive hypothesis can specify that people strong in a certain trait, or collection of traits, are likely to do well on the criterion. An alternative point of view is the *whole person* view—the idea that people are more than bundles of independent traits, that assessments should be holistic, looking globally at "the whole person."

Dachler (1989) suggested that selection be considered a part of personnel development, considering patterns of behavior rather than scorable dimensions,

focusing more on probability of future growth and adaptability than on fitness for a particular job. There is much to recommend his position.

Accepting one of these views may not wholly exclude the other. Two major differences between them are not insurmountable. First, traditional correlation uses measures of dimensions, not patterns. This does not, however, preclude correlating *X* and *Y* where *X* is the degree to which people fit a designated pattern of behaviors. Second, at least in the United States; the *Uniform Guidelines* (Equal Employment Opportunity Commission [EEOC], Civil Service Commission, Department of Labor, & Department of Justice, 1978, Section 5I, p. 38298; see Chapter 4) follow traditional methods. Although holistic evaluation of people and their future growth are nowhere mentioned in the guidelines, we suspect that a well-reasoned, well-developed selection procedure with evidence that it improves productivity, without violating the values of the larger society, will be permitted by the courts. Traditional research may seem to preclude more holistic approaches because not enough traditional researchers have thought about holistic approaches often enough or deeply enough to develop a solid paradigm for its use.

Ethical Testing

- The person conducting the assessments must have knowledge and understanding of the psychometric instruments being used.
- The assessment process should be standardized and each candidate being assessed should be treated the same.
- Applicants should be informed of the purpose of the assessments and how the results will be used.
- Who will see the results of the assessments should be clearly explained to the candidate.
- The testing professional must take reasonable steps to ensure that the results are not misused by others in any way.
- Where feasible, the testing professional should respect the applicant's desire for feedback.

Two important resources on ethical testing are the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014), *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, and the Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author. For further discussion of ethical issues in employee selection, see Lefkowitz, J., & Lowman, R.L. (2010). Ethics of employee selection. In J.L. Farr and N.T. Tippins (Eds.), *Handbook of employee selection* (pp. 571–591). New York, NY: Routledge.

Theory and Practice

Good practice requires understanding of what one is doing. An existing, relevant theory can promote understanding, but its existence does not ensure it. We call for more attention to theory to promote understanding of what is done in practice. Too much of what we know about personnel assessment and decision making, and, therefore, too much of this book, is limited to techniques. Better theories of work and work effectiveness can sharpen, prune, and expand those techniques and improve decisions. If there is a theme to this book, it is that we need to develop much greater knowledge of how managers use assessment results to make selection decisions and that we need to provide managers with sufficient knowledge concerning assessment methods, so that they have a strong basis for making more informed, rational, and accurate selection decisions.

An unfortunate but growing gap seems to separate academic science from organizational practice. Academics often seem interested only in building theories. Practitioners tend to decry the triviality and impracticality they perceive in academic theories, yet some of the theories they decry could inform many practical decisions in their organizations. There is, or should be, a symbiotic relationship between theory and practice and between basic and applied research. To be practical, a theory has to be a good one, internally consistent, supported by solid data, and tested in practice to find out how well it works beyond the boundaries of an experimental situation.

A third member of this mutual relationship is society at large. Both science and practice must heed the social issues and problems they solve or exacerbate. Many scientific questions, especially in the behavioral sciences, stem from the concerns of that larger society. Practice within an organization is also practiced within that larger society; for many practical decisions, both the relevant scientific foundations and their social effects must be considered.

Research should not be limited to just one chosen criterion; decision outcomes are likely to be plural. They need to be understood. Understanding requires HR research and development programs at least on par with product and market research, and these programs work best if informed by competent theory. Outcomes and reasons for unexpected ones can be clarified through research, providing further practical guidance for decision making. All of this occurs within a community (including the larger society) that experiences the effects of outcomes and seeks to influence them. With a well-funded R&D program, unspecified and unintended outcomes, whether relevant to community concerns or to organizational needs, could be investigated much as medical research looks for side effects of medical interventions.

We must not, however, be so wrapped up in psychometric research, statistical analyses, and the contextual influences of the community that we forget that the purpose of all this is to optimize the process by which some people get rewards and opportunities and others do not. The central focus of this process—the one intended

to reach the best possible outcomes—is a decision. Decisions are based on assessments; they also imply judgment, preferably informed judgment. Some of the information comes from research and theory, some of it comes from knowing the organization's needs, and some of it comes from community influences. We do, in fact, need more theory; and more theory needs to be informed by practice.

Discussion Topics

1. In the chapter, the authors argue for hiring based on merit. However, the definition of "merit" is open to interpretation; how would you define "merit"? Is it ever appropriate to hire on the basis of some other standard?
2. How do you think companies most commonly deviate from using psychometrically sound selection procedures? What are the consequences of this?
3. How does the selection approach of choosing the person who will be the best or highest performer in the job differ from choosing the person who has the best fit to the job, or is least likely to leave the job within a short time period? What are the implications of each?
4. What are some unique questions you have been asked when applying for jobs? If you have ever served as an interviewer, what are some of the more creative questions you have asked a job candidate?

References

(Note: In addition to citations contained in the text in Chapter 1, we have provided references that we believe are helpful to anyone involved in the practice of assessment.)

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1073.

American Psychological Association (2010). *Publication manual for the American Psychological Association* (6th ed.). Washington, DC: Author.

Arthur, W., Jr., Doverspike, D., Barrett, G. V., & Miguel, R. (2013). Chasing the Title VII Holy Grail: The pitfalls of guaranteeing adverse impact elimination. *Journal of Business and Psychology, 28,* 473–485.

Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology, 66,* 1–6.

Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist, 44,* 576–578.

Bryant, A. (2013, June 20). In head-hunting, big data may not be such a big deal. *The New York Times*. Retrieved from http://www.nytimes.com/2013/06/20/business/in-head-hunting-big-data-may-not-be-such-a-big-deal.html

Civil Rights Act of 1964 § 7, 42 U.S.C. § 2000e *et seq* (1964).

Civil Rights Act of 1991 § 109, 42 U.S.C. § 2000e *et seq* (1991).

Cohen, D. B., Aamodt, M. G., & Dunleavy, E. M. (2010). *Technical advisory committee report on best practices in adverse impact analyses*. Washington, DC: Center for Corporate Equality.

Dachler, H. P. (1989). Selection and the organizational context. In P. Herriot (Ed.), *Assessment and selection in organizations: Methods and practice for recruitment and appraisal* (pp. 45–69). Chichester, England: Wiley.

Dennis, P. M. (1984). The Edison questionnaire. *Journal of the History of the Behavioral Sciences, 20*(1), 23–37.

Drever, J. (1952). *A dictionary of psychology*. Baltimore, MD: Penguin.

DuBois, P. H. (1970). *A history of psychological testing*. Boston, MA: Allyn & Bacon.

Eels, R., & Walton, C. (1961). *Conceptual foundations of business*. Homewood, IL: Irwin.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290–38315.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, Department of Justice (1979). Interpretation and clarification of the Uniform Employee Selection Guidelines. *Federal Register, 44,* 11996–12009.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, Department of Justice (1980). Adoption of additional questions and answers to clarify and provide a common interpretation of the Uniform Guidelines on Employee Selection Procedures. *Federal Register, 45,* 29529–29531.

Freyd, M. (1923). Measurement in vocational selection: An outline of research procedure. *Journal of Personnel Research, 2,* 215–249, 268–284, 377–385.

Fuscaldo, D. (2014, January 11). Why HR should consider asking oddball interview questions. *Glassdoor*. Retrieved from http://employers.glassdoor.com/blog/why-hr-should-consider-asking-oddball-interview-questions/

Gupta, N., Ganster, D. C., & Kepes, S. (2013). Assessing the validity of sales self-efficacy: A cautionary tale. *Journal of Applied Psychology*, *98*, 690–700.

Lefkowitz, J., & Lowman, R. L. (2010). Ethics of employee selection. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 571–591). New York, NY: Routledge.

More slams at Edison; Experts pronounce his questions only one-tenth effective in gaining their purpose. (1921, May 22). *The New York Times*. Retrieved from http://www.nytimes.com

Poundstone, W. (2012). *Are you smart enough to work at Google?: Trick questions, Zen-like riddles, insanely difficult puzzles, and other devious interviewing techniques you need to know to get a job anywhere in the new economy*. Oxford, England: Hatchet.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

2

ANALYZING ORGANIZATIONS AND JOBS

Organizational Analysis, Job Analysis, and Competency Modeling

Organizations face many HR challenges, only some of which require solutions involving improved employee selection. Before deciding how to assess, what to assess, or even whether to assess, the source of HR problems must be identified. The three general approaches to the analysis and identification of HR issues are as follows:

1. Organizational Analysis
2. Job Analysis
3. Person Analysis

Most of this book is devoted to person analysis, whether accomplished through tests, inventories, or ratings. In this chapter, we briefly cover the topics of organizational analysis and job analysis.

Organizational Analysis

The scope of *organizational analysis* is broader than can be addressed fully or appropriately in this book. Considering organizational analysis as providing the context in which selection and other interventions can be compared as possible solutions, however, is important. For example, organizational analysis will tell us whether there is enough money available to pay for job analysis or test validation studies. Organizational analysis may also identify where turnover is extremely high or where there are deficiencies in an organization's succession plan.

Organizational analysis typically is precipitated by a problem (e.g., poor productivity, high turnover) or by changes in organizational goals. Effective organizational analysis identifies gaps between current states and future needs and generates

hypotheses about future courses of action, which might be used to close the gap between current and desired states.

> For a classic guide to organizational analysis, see Levinson, H. (2002). *Organizational assessment: A step-by-step guide to effective consulting*. Washington, DC: American Psychological Association.

In any given situation, some actions may be more effective than others. For example, a specific problem may be addressed by improved selection, improved training, job redesign, or changes in organizational structure or policy. Informed judgments of relative effectiveness (and relative costs) of the options determine the focus of further study and action.

Organizations function as systems, and the needs and actions that appear to focus on only one aspect of the organization will also have implications for others. If organizational analysis suggests improved selection as potentially useful, it also suggests criteria important enough to measure and predict. It may also identify needed personal characteristics. In these ways, organizational analysis can identify a selection problem that cuts across organizational units.

Organizational-Level Outcomes

Economic outcomes at the overall organizational level include profit or loss, stability or fluctuation of stock value, market share, and so on. Reasonable profit, stable or rising stock value, and growing market share please organizational leaders. Losses, fluctuations, and low market share are not pleasing; they are problems to be avoided or overcome. Overcoming them may call for new strategy, capital investment, high-level personnel changes, changes in manufacturing processes or inventory controls, or human capital investments such as supervisory training or hiring more highly skilled employees.

People-oriented outcomes include performance (quality and quantity), workforce dependability (low turnover, absenteeism, or tardiness), workforce health and well-being (low stress and stress-related illness, few accidents), employee attitudes and motivation (job satisfaction, organizational commitment), or responsible versus counterproductive or antisocial behavior (no use of alcohol or drugs, theft, or sabotage). Some of these are organization-wide; others may be concentrated in specific units. An enormous number of variables may contribute to performance problems, including declining morale, worn equipment, inadequate training, poor quality of tools and work aids, poor supervision, or failure to select people on the basis of assessment of genuinely important qualifications. Each suggests its own corrective action, and several of them may be needed. New selection procedures may be one of many changes that are needed.

Approaches to Organizational Analysis

Organizational analysis must attempt to solicit different points of view and, to the extent possible, reconcile them. The methods used for organizational analysis will depend upon the background and skills possessed by the analyst. Three approaches used frequently by HR professionals are conference methods, SWOT analysis, and organizational assessment surveys.

Conference Methods. Dialogue among people with different views is an essential condition for effective communication, which in turn is essential for clear identification and definition of problems and their solutions. One approach to organizational analysis, then, gets knowledgeable people together to talk about an issue or problem. Talking may sometimes become argumentative or excessively formal, so it can be useful to bring in an outside consultant who can facilitate the process and its focus.

The dialogue often begins with questions such as the following:[1]

1. What is the nature of the issue or problem at hand?
2. What is its history?
3. What are the perceived outcomes or consequences of the problem? What observations have led to these perceptions? How consistent or how variable are people's perceptions?
4. What is occurring system-wide that is or might be related, or that might have an impact on the investigation of the issue or problem?

The facilitator probes beyond these basic questions, with the probes varying depending on the answers. Questions of structure, processes, key systems or subsystems, policies, and external forces might be explored. The approach is planned in advance only in a general way; it is not a standardized approach to be used consistently for all organizations, all issues, or all problems.

SWOT Analysis. A classic approach to analyzing and organizing information from an organizational analysis is referred to as SWOT. The SWOT acronym corresponds to the following:

- Strengths—What does the organization do well? Corresponds to positive, internal factors.
- Weaknesses—What does the organization do poorly? Corresponds to negative, internal factors.
- Opportunities—What is going on in the organization's environment? What are some current trends that offer the organization opportunities? Corresponds to external, positive factors.
- Threats—How will changes in the environment or technology have a negative future impact on the company? What are future trends that may

represent major problems for the organization? Corresponds to external, negative factors.

In conducting a SWOT analysis, a consultant may use a mix of methods including focus groups, conferences, and surveys. The SWOT framework is used to guide discussion, organize the information obtained, and generate feedback. An effective SWOT process should lead to the identification of future goals, gaps, and action steps. An alternative to SWOT, which focuses on the positive side of organizational culture and life, is called *appreciative inquiry* (see Cooperrider & Srivastva, 1987).

Organizational Assessment Surveys. On the other hand, survey methods are often proposed precisely because they are standardized. Questionnaires can be developed after interviews and conferences to be sure that major questions are asked and to obtain quantitative results. Any organizational analysis is useful when it helps people in organizations overcome force of habit in studying organizational problems. Raising questions about problems should be *systematic* and ongoing. To some organizational experts, *systematic* requires careful measurement, and *ongoing* means periodic. Regularly scheduled surveys can meet both requirements.

Van de Ven and Ferry (1980) developed the Organizational Assessment Instruments for survey research, one of many possible structured approaches to developing an organizational analysis survey. However, today, many organizations develop their own surveys; for a short sample of typical questions see Table 2.1. There are a number of software companies that provide easily customizable online software for designing and delivering surveys. As a student or HR professional, you may already have taken multiple surveys constructed using SurveyMonkey. However, the analysis and interpretation of the data from an organizational survey still required the trained eye of an HR professional.

A General Approach to Organizational Analysis

Organizational analysis is a managerial function, not a research function. Its immediate purpose is to generate hypotheses, not to test them. It must be done systematically, recognizing that the outcome of organizational analysis is a judgment (or a set of judgments) that can be framed in the language of hypotheses, and the quality of the judgment depends on the experience, knowledge, and wisdom of those who reach it. The best advice we can offer, whether the focus is on dialogue or on questionnaires, is to consider seven general questions as carefully as possible:

1. What work outcomes are most in need of improvement? That is, what outcomes are most highly valued and not satisfactorily attained?
2. How widespread is the problem? Is it pervasive throughout an organization or organizational unit, or is it found in specific instances (i.e., specific people or specific units)?

TABLE 2.1 Sample Questions From an Organizational Survey

Overall Attitude Toward Company

- Overall, this is a good company.
- Considering everything, I am satisfied working here.
- I would recommend the company to others as a place of employment.
- I enjoy working for the company.
- I feel a sense of security working for the company.
- I feel the future is bright for the company.

Communication

- The company does a good job of communicating information to employees.
- The company openly shares information with employees.
- I am satisfied with the information I receive from Management on what's going on in the company.
- This company is willing to listen to alternative points of view.
- Employee input is used when making decisions.

Human Resources

- The Human Resource department contributes in a positive manner to the functioning of the company.
- I agree with most of the decisions made by Human Resources.
- The Human Resource policies of this company make sense.

Satisfaction With Job

- Overall, I am satisfied with my job.
- I feel good about my job.

Satisfaction With Pay and Benefits

- I feel that my current level of pay is fair.
- I am satisfied with the way decisions are made regarding pay levels.
- I am paid fairly for the work I do.
- I have a clear understanding of the pay policies.
- I am happy with my benefits.
- The benefits offered by this company are excellent.
- The benefits offered by this company meet my needs and the needs of my family.

Note: All questions are responded to on a scale of (1) *strongly disagree* to (5) *strongly agree*.

3. At what level of analysis (individual or organizational) is the problem most accurately defined and approached? Consider, for example, a serious absenteeism problem. Should it be approached at the individual employee level or a broader organizational level?
4. What kinds of corrective actions are plausible? That is, what might reasonably be expected to help? Discussions with different people in the organization, and perhaps with outside consultants, can provide an initial list of plausible actions.
5. How effective have the various options been in prior use, in this organization or elsewhere? This question is probably the one that gives some edge to attempts to improve employee selection decisions when the problem is one of

improving performance levels. Most other activities lack the strong research base, with the substantial levels of predictive power and utility that characterizes the testing literature.

6. To what extent is money and personnel available to respond to the identified gaps? How much time does the organization have to respond to the problem before it becomes a crisis? What is the risk of potential litigation?
7. Is organizational leadership committed to working to solve the problem? All too often, an organizational analysis leads to the identification of severe problems. Solutions are then offered, but nothing is really done to respond to the issues. A year later, another organizational analysis is conducted and identifies the same areas of concern. This goes on year after year; meanwhile, the employees of the organization become more frustrated with consultants and HR, and more cynical considering the possibility of true change. If you are going to identify problems in your organization, be prepared to act on the proposed solutions.

Organizational analysis may lead directly to a hypothesis about appropriate selection procedures. Usually, however, hypothesis development requires analysis of individual jobs (or job families). Traditionally, job analysis has been considered necessary for building a selection system. Its use is strongly encouraged in equal employment opportunity (EEO) case law, it contributes to criterion development by identifying the most important aspects of performance, and it is a basis for choosing potential predictors. In the next section, we cover job analysis.

Job Analysis

When organizational needs require improved personnel decisions for people in specific positions, jobs, or groups of jobs, job analysis (or position analysis) is necessary. The purpose is to understand the needs clearly enough to know what aspect of job behavior should be predicted and to identify variables or constructs that might be effective predictors—that is, to develop predictive hypotheses.

> ***Job analysis*** is a study of what a jobholder does on the job, what must be known in order to do it, what resources are used in doing it, and perhaps the conditions under which it is done. One of the most common errors made by job analysis is to be overly influenced by qualities of the person, rather than the real requirements of the job.

What the jobholder does may be defined in several ways: as tasks, classes of duties or responsibilities, broad activities, or general patterns of behavior. What must be

known includes job knowledge and job skills. Based on the EEOC's *Uniform Guidelines,* we would argue that any job analysis must describe the following:

- The major work behaviors, including the tasks that make up the major work behaviors. For each task, information should be provided on frequency and importance.
- The knowledge, skills, and abilities (KSAs) required for the job. For each KSA, information should be provided on whether the KSA is learned on the job or required at entry, and the importance or criticality of the KSA.

Resources used may include those the person brings to the job (relevant experiences, general abilities, or other personal characteristics), tools and materials used (e.g., supplies or equipment) or the work products of other jobs or work units. Following McCormick (1979), with some additions and liberties of our own, Table 2.2 provides some standard definitions relevant to job analysis.

Fundamentally, all job analysis consists of observing what can be seen and asking questions about what cannot. A job analyst watches, questions, understands, and summarizes the information received to form a job description. Some jobs can be adequately analyzed just by watching workers work; others require extensive questioning by interview or survey. Job incumbents can be observed or questioned in several ways; a job analysis that provides the best job understanding usually uses several methods.

> A common question is whether there is any preferred or required method of job analysis. Based on court cases and the *Uniform Guidelines,* the answer would be that there is no required method of job analysis. However, in recent years, the regulatory agencies in the United States have shown a preference for job analyses that result in verifiable, quantitative results. For this reason, many validation studies now make use of questionnaires, which result in objective data.

Starting Points

Very rarely must the job analyst start from scratch. Unless a job is brand new or does not exist yet, there are usually two sources that the job analyst can turn to as a starting point. The first is previous job descriptions and the second is the O*NET.

Previous Job Descriptions. In most cases, there will already be an existing job description. It may be rather incomplete or quite dated, but it probably exists. Even if there is no formal write-up, other resources are usually available such as

TABLE 2.2 Common Terms and Standard Definitions Related to Job Analysis

Term	*Definition*
Position	The duties and tasks carried out by one person. A position may exist even where no incumbent fills it; it may be an open position. There are at least as many positions in an organization as there are people.
Job	A group of positions with the same major duties or tasks; if the positions are not identical, the similarity is great enough to justify grouping them. A job is a set of tasks within a single organization or organizational unit.
Occupation	An occupation is a class of roughly similar jobs, found in many organizations and even in different industries. Examples include attorney, computer programmer, mechanic, and gardener.
Job Family	A group of jobs similar in specifiable ways, such as patterns of purposes, behaviors, or worker attributes. An example of a job family might be "clerical and technical," which could include receptionists, accounting clerks, secretaries, and data entry specialists.
Element	The smallest feasible part of an activity or broader category of behavior or work done. It might be an elemental motion, a part of a task, or a broader behavioral category; there is little consistency in meanings of this term.
Task	A step or component in the performance of a duty. A task has a clear beginning and ending; it can usually be described with a brief statement consisting of an action verb and a further phrase.
Duty	A relatively large part of the work done in a position or job. It consists of several tasks related in time, sequence, outcome, or objective. A clerical duty might be "sorting correspondence." One task in correspondence sorting might be "identify letters requiring immediate response."
Job Description	A written report of the results of job analysis. A job description is usually narrative, sometimes given in a brief summarizing paragraph. It may be more detailed. Where job analysis was done by survey methods, the description may include listings of task statements found to define or characterize the job being studied, along with statistical data.

training materials or performance checklists. Existing job information can be used to create an initial job description, which can then be distributed to incumbents and supervisors for review, comment, and confirmation.

Occupational Information Network (O*NET or ONET). The most extensive job analysis program was that of the United States Employment Service (USES) in developing the *Dictionary of Occupational Titles* (DOT; United States Department of Labor [USDL], 1977). The brief DOT descriptions were backed by extensive descriptions based on combined methods of observation and interviewing. The

procedures were described in the *Handbook for Analyzing Jobs* (USDL, 1972). Many experienced job analysts may still talk favorably about the DOT and may still make use of the DOT in searching for job information.

Today, most analysts rely upon a new method of large-scale occupational analysis that has been developed by the USDL. The O*NET database is accessible online at www.onetonline.org, and provides a set of ready-to-use instruments for describing jobs. O*NET OnLine has detailed descriptions of jobs or occupations for use by HR professionals. The O*NET uses general worker-specific descriptors that can be used to describe multiple jobs, as well as lists of major duties and tasks statements. A number of potential applications of O*NET to job analysis and predictor selection have been presented (e.g., Jeanneret & Strong, 2003; Peterson, Borman, Hanson, & Kubisiak, 1999).

As a note of caution, because the O*NET is so convenient, there is a tendency among job analysts to overuse the descriptions and material provided online. The O*NET provides very detailed descriptions, but also attempts to describe occupations throughout the United States economy. Thus, some of the material provided will describe a particular job in a specific organization, but other information will be inaccurate. The O*NET should be used as a starting point, as an inexpensive source of occupational information. However, the responsible job analyst will want to confirm the information provided with incumbents and supervisors.

Try the O*NET for yourself. First, go to www.onetonline.org. Then, in the upper right corner you should find a search box labeled *occupation quick search*. Type in the name of a job and then click on the search arrow. You should be taken to a page that lists a series of occupations corresponding to your search term. Click on the first occupation listed and you will be taken to a wealth of information on that particular occupation.

Observation

Direct observation consists of watching—and taking appropriate notes. It is the most obvious way to learn about a job, but it poses problems. The incumbent may work differently in the presence of an observer, perhaps going more by the book than is necessary, or perhaps inflating the job by adding things not ordinarily done, or perhaps failing to do some things because of nervousness about being watched. Observation is time-consuming and expensive. To observe a sample of several workers requires extensive observer time and skill. The biggest problem, however, is that much work is simply not observable. Consider the job of computer technician; much of the content of this job goes on in the head of the incumbent. Questions must be asked and answered to augment and interpret what can be directly

observed. Another problem is that some jobs may involve dangerous conditions or the presence of a job analyst may be inappropriate.

Instead of actually observing the job, the analyst may rely upon introspective reports of incumbents, which may be useful, especially in the development of *experience samples*. In this procedure, a job incumbent identifies with a brief note in a smartphone or notebook what is being done at specific times during the day. The record allows the analyst to identify not only tasks and activities but sequences of these. Inferences from the experience samples should be verified in a follow-up interview of the incumbents who filled them out.

Interviews

Interviews are one of the most frequently used methods of collecting job analysis information. Individual or group interviews may be conducted. Interviews may be conducted with managers, supervisors, or incumbents. The interview is a remarkably flexible approach and is useful in many ways.

An initial interview with an incumbent before observing work can clarify the nature and purpose of the observation, provide a broad view of the job being observed, and reassure the person being watched. Will the work be done all in one place, or will it be necessary to move around to see everything? Are certain crucial aspects of the work likely to be done so quickly that only an alert observer will see them? Such questions, answered during an initial interview, can guide the analysis. Verification interviews after observations can verify (or modify) other information, and the other information may stimulate incumbents to mention things otherwise overlooked.

The advantage of interviewing as an adjunct to observation seems obvious. What may be less obvious is that questioning, as in an interview, may need to be augmented with observations. Landy (1989) pointed out several problems with interviewing alone. One is that interview results may describe what should be done ideally or in theory rather than the way the job is actually done. Another is that experienced incumbents often have forgotten just how they do a job, as much of it has become automatic.

Most employees seem to enjoy talking about their job. However, a skilled job analyst knows how to guide the interview and maximize its effectiveness. For some jobs, you can simply ask the employee to talk you through his or her day, starting with what they do when they come into work to set up for the day, what tasks they engage in during the day, and how they complete or finish up their day. For other jobs, it may be more effective to simply say "Most jobs involve 5–7 duties or major work behaviors. Can you list the 5–7 major work behaviors involved in your job?" In either case, the job analyst digs deeper by asking about specific inputs, tasks, and outputs, as well as the associated KSAs. In conducting the job analysis, the expert analyst also keeps in mind the purpose, which in this case would be developing and choosing effective selection instruments.

Sentence Element	Task Statement
Subject: Who?	(Always the worker; unstated)
Action verb: Perform what action?	Schedules
Object of verb: To whom or what?	Appointments, meetings, events
Phrase: Upon what instruction? Source? Specificity?	Supervisor, caller, or memo; usually a vague "set it up," perhaps with deadline
Phrase: Using what tools, equipment, work aids?	Calendar, appointment pad, telephone, or conference room schedule book as needed
In order to … : To produce or achieve what? (Expected outcome)	To ensure presence of those expected to be present

Task Statement: Schedules appointments, meetings, or events according to instructions from supervisor or memo, or requests received from callers, using as needed appointment pads, calendars, telephone, or conference room schedule book in order to ensure that all those expected to be present at the meeting or event will be able to do so.

FIGURE 2.1 An example of a task statement developed by task sentence structure in functional job analysis.

Functional Job Analysis (FJA). A job analyst must distinguish what people do on the job from what gets done as a result; FJA provides a grammar for doing so. Within the FJA system, things people do are called ***worker functions*** (Fine & Cronshaw, 1999); they are action verbs in task statements in which the subject is always understood to be the worker, statements fleshed out in subsequent phrases more fully describing the task, as in Figure 2.1. Instructions for using the FJA approach to develop an inventory of task statements are available in Fine and Cronshaw (1999). The system assumes that everything workers relate to in the course of their work can be subsumed under one of three headings: people, data, and things. These headings are even broader than they seem; interactions with *people* may include analogous interactions with animals; *data* includes a full spectrum of information, ideas, statistics, and so on; and *things* include virtually any tangible object touched or handled, such as complex machinery, books, or the top of the desk.

Critical Incidents. In the critical incident method, developed by Flanagan (1954), the job analyst meets with a group of incumbents (or others with expert knowledge of the job) and draws from them their recollections of things people have done that resulted in noteworthy consequences, good or bad. For each incident, the description is limited to facts, not inferences, but it is detailed, reporting environmental contributing factors (e.g., equipment problems) or antecedents that may have contributed to the incident. Extraordinary events are more likely than ordinary ones to have memorable consequences and to be recalled, so this is not a

good technique for getting complete job descriptions. It is, however, an excellent way to gain insights into crucial aspects of performance, which makes it an ideal technique for developing rating scales for use in performance appraisal. The critical incident method is also very useful in identifying content for employment interview questions and situational judgment tests.

A typical critical incident interview might involve asking a supervisor to describe a recent action taken by an employee that was *unusually effective* in his or her job role. What were the circumstances leading up to the incident? What exactly did the employee do? Why was this so helpful in getting the job done?

Job Analysis Surveys

Observational and interview methods are useful when studying a single job where adequate information can be obtained from a few experts. When simultaneously studying sets of related jobs, especially in multiple locations, survey research may be the preferred method of collecting observations. Surveys are useful where many people have jobs with the same or similar titles but do different things. Even where jobs seem standard, such as police patrol officers, work performed may vary widely. A survey of job incumbents permits study of virtually all positions to determine whether there is enough uniformity among them to treat them as one job, or whether the positions should be grouped into different jobs with distinguishable patterns of duties.

Survey methods of job analysis are amenable to statistical analyses. They call for the development of items that can be combined into internally consistent job dimensions. Dimension scores can identify differences or similarities of positions within jobs, or they can be used to infer predictor constructs of greatest potential job relevance. Quantification helps even for analysis of a single job or position in a single location. Scores on major job dimensions clarify the degree to which one feature of a job is more important than another, and they help in choosing criterion or predictor constructs. As previously mentioned, surveys also appear to be more objective than interviews and result in numbers that can be subject to statistical analysis.

Task Inventory Development. McCormick (1959) distinguished two types of inventories for surveys: job-oriented and worker-oriented. A ***job-oriented inventory*** is a set of brief task or activity statements (usually much briefer than the example in Figure 2.1), each of which may describe what is done and what gets done as a result, for example: (a) "translates correspondence from French

to English," (b) "coordinates departing, en route, arriving, and holding aircraft by monitoring radar and communicating with aircraft and other air traffic control personnel," or (c) "writes special reports." Each example includes an action verb saying what is done and a further phrase describing what is accomplished.

In contrast, ***worker-oriented inventories*** describe work activities in terms that describe behavior, not accomplishments. McCormick's example, instead of describing a baker's job with the statement "bakes bread" (a job-oriented statement), described the activity with statements such as "manually pours ingredients into containers" and "observes condition of product in process" (McCormick, 1959, p. 411). These are worker oriented in that they describe behaviors that might be required in a variety of jobs, for example, chemists, some quality-control inspectors, or candy makers. The use of worker-oriented questions should be distinguished from asking questions about knowledge, skills, abilities, and others (KSAOs), even though there will be a great deal of overlap.

The core of a job-oriented or task inventory is a set of task statements, or items, each with its action verb, direct object, and necessary delimiting phrases. Developing the set of statements is usually an iterative process in which the preliminary statements are edited, perhaps several times, during the various phases of inventory development. A first step in item writing is to consult available information such as training manuals, earlier job descriptions, organization charts, reference materials or manuals used in doing the work, or procedural guides and work aids. Such documents provide initial understanding of the job and may, perhaps, suggest some preliminary task statements. Information gleaned from documents can be augmented (or corrected) through on-the-job observations and interviews with knowledgeable people; such information can add to the pool of preliminary task statements.

Using Job Experts. Job experts, meeting in groups, can add, delete, or edit statements. Job experts may be incumbents, supervisors, engineers, quality-control staff, trainers, occupational safety officers, job evaluation staff, or others who have relevant knowledge about the targeted jobs.

> Job experts are often referred to as subject matter experts or simply as SMEs.

Writing Items. Inventory items must be descriptive but brief. They may be written by staff members or consultants, or in conference by groups of experts. They may be written at various levels of specificity. General statements are usually preferable to highly specific ones, but not always; what seems important is to keep the

level at least somewhat similar across statements. Unfortunately, there seems to be no standard method for doing so. We offer a suggested, but untested, method:

1. Write preliminary items without concern for level of generality.
2. Have job experts sort them, independently, into sets with fairly consistent content.
3. Conduct a consensus meeting to reconcile differences in items placed in the different content sets.
4. Within each set, have each expert arrange the items in a hierarchy from most specific to most general, placing the most general at the top of the stack with more specific statements below to illustrate it.
5. Conduct a consensus meeting for a final arranging of statements and to judge the similarity of level among statements topping the sets.

If the experts think those statements are comparable in generality, item development may be complete, but items judged too general or too specific probably need editing.

The type of items written will depend upon the purpose, intended use, and the authors' own approach. However, most job analysis inventories used today include a mix of task-oriented items, items assessing KSAOs, and more general ability attributes, which could be seen as similar to the older worker-oriented approach. Some of the common item types and associated example items appear in Table 2.3; this inventory might be used with someone in a general manufacturing job.

Response Scales. Ratings of importance or criticality are almost always requested. Ratings of complexity or task difficulty can help target levels of ability to be assessed. *Frequency* of task occurrence, or *time spent* on it, are commonly used scales. It is often important to ask whether the task can be performed as soon as one gets on the job or only after extended training or experience. Multiple scales are used because analysts want a variety of task information, but they may also increase reliability of ratings by forcing greater attention to them. A problem with multiple response scales, however, can be that the distinction seen by the investigator may not make much difference to respondents. Task inventories often call for ratings of both importance and frequency of performance; correlations between these scales often approach 1.0, suggesting that they mean the same thing to respondents. Nevertheless, both scales may be necessary (e.g., A lifeguard may hardly ever save a life, but it is the most important part of the job!).

The distinctive features of a job might define it and its critical components or its essential functions. A distinctive feature might be one that (a) takes up the bulk of the respondent's work time, (b) is crucial to some important work outcome (something would not result, or would not turn out well, if the task were not done effectively), or (c) is undertaken by the respondent only. Such tasks may define the

TABLE 2.3 Examples of Different Types of Job Analysis Survey or Inventory Questions

Type of Question	*Example of Typical Item Stem (Task, KSAO, Worker Characteristic, or General Ability)*	*Typical Types of Rating Scales Used*
Task-Oriented	Monitors and troubleshoots packaging machines	• Frequency (How Often) • Importance
Task-Oriented	Records production data such as number of pieces produced and quality of pieces	• Frequency (How Often) • Importance
KSAOs	Knowledge and skill in troubleshooting packaging machines	• Criticality • When Learned
KSAOs	Knowledge of and skill in recording production data such as number of pieces produced and quality of pieces	• Criticality • When Learned
Worker-Oriented	Monitors processes	• Frequency • Importance • Scale ordered from simple to complex monitoring
Worker-Oriented	Performs record-keeping duties	• Frequency • Importance • Scale ordered from simple to complex types of record keeping
General Ability	Knowledge of mechanical principles	• Criticality • Scale ordered from simple mechanical principles (operation of simple tools) to complex (principles of programming of automatic controls)
General Ability	Knowledge of math	• Criticality • Scale ordered from simple math (copy numbers) to complex (apply calculus)

job; to differentiate among somewhat similar jobs, the following response scale combines all three. On a 4-point response scale, the responses may be as follows:

0. I do not do the work described in this statement.
1. This statement describes something I may *occasionally* do, but it is neither an important nor a frequent part of my work.

2. This statement clearly *describes* my work; I do it, but it is not very time-consuming, nor as important as other things I do, nor unique to my job.
3. This statement *defines* my job either because it is one of the most important things I do, or because it describes my work a great part of my time, or because no one else in my work unit is responsible for doing it.

It is best if local job experts choose wording they think communicates best for a given survey. Whatever the precise words, the scale is a composite of three scales and may, therefore, seem ambiguous. However, it has a unifying theme of job definition, and job experts seem not to be bothered by the ambiguity.

Pilot Studies. Inventories should be pretested for clarity and content. One kind of pretest asks a few people to read instructions and complete the inventory, "thinking aloud" throughout. As they verbalize their thoughts, ambiguities, unintended meanings, and other problems come to light. At some point, the draft inventory should be given to a sample of job incumbents. If possible, it should be completed in the presence of investigators so that problems with individual task statements or response scales can be observed and recorded. The task list for this preliminary study should include places for incumbents to identify tasks they perform that did not appear in the list.

Task Inventory Administration. It is often desirable to gather data independently from incumbents and their supervisors. People who have held a job a long time, at least one permitting some autonomy, may come to do some things routinely without the supervisor's awareness; they may do some things so automatically that they are not aware of it themselves. Supervisors may expect some things to be done without checking to be sure they are clearly communicating the expectation to the worker. An incumbent may inflate the nature of the job (Morgeson, Delaney-Klinger, Mayfield, Ferrara, & Campion, 2004); a supervisor may disparage it. If the incumbent and supervisor both complete the survey questionnaire independently, the two versions of the job can be reconciled in meetings with job analysts. The resulting description can be more readily accepted as correct.

Grouping Task Statements. Task statements may be grouped to assist in distinguishing jobs within a group of jobs, to define criterion measures for evaluating performance, and to infer predictors of those criteria. Two ways are used to group task statements to facilitate meeting these objectives. They can be grouped rationally by job experts who, independently, identify broad categories and assign the statements to them. If statements are grouped as part of the inventory construction, grouping is by definition rational. The alternative is a statistical analysis of responses in a pilot study or in the actual job analysis. Cluster or factor analysis is often used, but these approaches have produced

erroneous results (Cranny & Doherty, 1988). We suggest the following steps (or some variant):

1. When the list of task statements is available, and the responses from the survey are at hand, compute mean responses for each statement and select the most critical, important, or defining (or whatever response scale is used) tasks within each group; an arbitrarily chosen point on the scale will separate those selected from others.
2. For each group of task statements, have a panel of judges, meeting in concert, select a pair of statements describing the most clearly different tasks.
3. Have the panel sort the remaining statements (from Step 1) into one of three stacks: like one, like the other, or like neither (presumably the "like neither" stack will be largest).
4. Repeat the process with the "like neither" stack until all of the most important or most critical task statements have been allocated.
5. Check for consistencies of responses across task statements within each group. Where inconsistencies are noted, either reassign the statement, or remove it from consideration.

Linkage of Required Worker Characteristics to Activities. It is important to distinguish between describing the job activities and inferring the personal characteristics required for completing the activities. These personal characteristics have come to be known widely as KSAs. Sometimes the term is expanded to KSAPs or KSAOs, where the letter *O* or *P* stands for "other personal characteristics." We are not fond of these terms, in part because of problems distinguishing skills from abilities, or knowledge from skills, with reasonable satisfaction. Nevertheless, reference to KSAs or KSAPs has become so widespread, and is such a convenient shorthand, that we will use it rather than fight it. However, *when* we use it, readers should see it as shorthand for a more inclusive term, *job requirements*.

Deriving the worker characteristics needed to perform the work activities is a critical component of job analysis. Indeed, it is this component that allows for the leap from describing the job to hypothesizing about what kinds of workers will do the job well. Sanchez and Levine (2001) noted that the translation of job activities into worker characteristics is what makes job analysis a *psychological* endeavor. Many job analysis inventories include items listing possible job requirements, and many of them ask respondents to link job requirements to tasks or, more often, activities. One way to request linkage judgments is illustrated in Figure 2.2. In the inventory from which it is drawn, general duties were listed and KSAs were restricted to ability factors. Each cell represents a potential predictive hypothesis. Job experts entered a 0, 1, 2, or 3 in each of the cells, and mean judgments of experts was computed. Arbitrarily, a predetermined value (perhaps a mean of 1.5, 1.7, or 2.0) may be interpreted as suggesting a useful hypothesis, and that value might change from one predictor construct to another.

Linkage of KSA Categories to Major Job Duties

In the table below, the major job duties have been listed down the left hand side. The KSA categories agreed upon have been listed across the top. Each job duty is a row in the table; each KSA is a column. The place where a row and a column intersect is a cell. The definitions for the brief phrases here are given in the help sheets; please keep those definitions before you all the time you're going through this exercise.
Each cell calls for your judgment about the relevance of the ability listed in the column to performance of the duty listed in the row. You should record your judgment as a 0, 1, 2, or 3 according to this scale:

0 – not at all relevant to the performance of this duty
1 – relevant, but only slightly, to performance of this duty
2 – relevant to an important degree to performance of this duty
3 – of the highest relevance to the performance of this duty

Job Duty	KSA				
	Verbal Comp	Clerical Sp & Acc	Interview Skill	Number Facility	General Reasoning
1. Questions clients					
2. Evaluates documentation					
3. Explains, answers questions					
4. Refers clients to resources					
5. Codes information					
6. Develops budget worksheet					
7. Calculates needs, allowances					

FIGURE 2.2 Linkage of KSA categories to major job duties.

A linkage matrix like Figure 2.2 can be used to generate predictive hypotheses in several ways. If an overall performance criterion is to be used, a summary statistic for each ability column may be used; it might be a simple average or an average weighted by ratings of the importance of the various duties. It might be simply the number of cells in the column with cell means exceeding the predetermined value. The same options exist if duties are grouped for independent performance evaluations. If job relatedness is to be determined by criterion-related validation, errors in the linkages will be corrected by failure to find satisfactory correlations. If job relatedness is to be determined by expert judgments, however, the duty and KSA definitions must be tested in pilot studies to assure

common interpretations, and the rules for inferring job relatedness must be carefully considered in advance; any subsequent deviations must be justified, if, indeed, they can be, with very great care.

Ready-to-Use Surveys

Although in many cases, organizations develop their own tailored surveys, there are also a number of standardized surveys available. One of the older standardized measures, the *Position Analysis Questionnaire,* or PAQ, is one of the best known job analysis survey methods. Today, many testing and consulting firms offer online job analysis surveys. For example, Harvey (1993) developed and offered an online questionnaire known as the Common Metric Questionnaire. In this section, we briefly discuss the PAQ as well as a newer survey that can be used for the assessment of personality dimensions related to job performance.

Position Analysis Questionnaire (PAQ). A job-oriented inventory must be explicitly developed for each job or occupational group studied. This fact reduces the generality of job dimensions identified and renders comparisons across jobs and occupations difficult. By definition, a worker-oriented approach is applicable across widely differing occupations (For examples of worker-oriented items see Table 2.3). For this reason, some people prefer to use a worker-oriented approach to position, job, or occupational surveys.

The most widely known worker-oriented approach is the PAQ by McCormick, Jeanneret, and Mecham (1969).[2] The items are organized into six general areas: (1) information input, (2) mental processes, (3) characteristics of work output, (4) relationships with others, (5) physical and social environment, and (6) others.

The recommended procedure for completing the PAQ involves the use of job analysts as data collectors. The job analysts may be people in the organizations (or outside consultants) whose major job is analyzing the jobs of others, or they may be other employees specifically chosen for this ad hoc assignment. They interview job incumbents, their supervisors, or both; after the interviews, the analysts themselves complete the PAQ forms. Alternatively, an analyst may meet with small groups of incumbents and supervisors who fill out the form in the presence of the analyst who can answer questions about the statements or the rating scales.

Personality-Based Job Analysis Surveys. In his presidential address to the Division on Measurement and Evaluation, American Psychological Association, Douglas Jackson said that job analysis techniques have largely overlooked personality predictors; a fact which he said accounted for their poor history. So challenged, a Bowling Green research group developed an inventory specifically intended to generate hypotheses about potential predictors among personality traits (Raymark, Schmit, & Guion, 1997). It is based on a list of 12 personality dimensions, shown, with definitions and contrasts, in Table 2.4. A sample page is shown in Figure 2.3.

TABLE 2.4 Work-Relevant Personality Dimensions

Broad Trait	*Dimension*	*Brief Description*
I. Surgency		
	I-A: General Leadership	Tendency to take charge of situations and motivate behavior of others.
	I-B: Interest in Negotiation	Ability to see and understand differing points of view and achieve harmony.
	I-C: Achievement Striving	A desire to advance and to excel relative to others or a personal standard.
II. Agreeableness		
	II-A: Friendly Disposition	Tendency to be likable, warm, and approachable.
	II-B: Sensitivity to Others	Tendency to be considerate, understanding, and have a concern for others.
	II-C: Collaborative Work Tendency	Desire to be part of a group and work well with clients, customers, and coworkers.
III. Conscientiousness		
	III-A: General Trustworthiness	A reputation for following through on promises, commitments, or agreements.
	III-B: Adherence to Work Ethic	A tendency to work hard and to be loyal.
	III-C: Attention to Details	A meticulous approach to one's own work or the work of others.
IV. Emotional Stability		
		A calm, relaxed approach to situations, events, or people.
		Adaptability to changes in the work environment.
V. Intellectance		
	V-A: Desire to Generate Ideas	A preference for original or unique ways of thinking about things.
	V-B: Tendency to Think Things Through	Considering the consequences or effects of alternative courses of action.

Other efforts in this direction include a job analysis inventory geared to personality traits that have been used for occupations such as police, managers, and bus drivers (Inwald, 1992). There are also two inventories linked to specific instruments or theories. Costa, McCrae, and Kay (1995) described the NEO Job Profiler, designed for use with the Revised NEO Personality Inventory, and Gottfredson and Holland (1994) developed the Position Classification Inventory to match the Holland RIASEC personality theory (for a summary, see Rounds, 1995).

EFFECTIVE PERFORMANCE IN THIS POSITION REQUIRES THE PERSON TO:		Not Required	Helpful	Essential
	Set 1			
1.	lead group activities through exercise of power or authority.	☐	☐	☐
2.	take control in group situations.	☐	☐	☐
3.	initiate change within the person's work group or area to enhance productivity or performance.	☐	☐	☐
4.	motivate people to accept change.	☐	☐	☐
5.	motivate others to perform effectively.	☐	☐	☐
6.	persuade coworkers or subordinates to take actions (that at first they may not want to take) to maintain work effectiveness.	☐	☐	☐
7.	take charge in unusual or emergency situations.	☐	☐	☐
8.	delegate to others the authority to get something done.	☐	☐	☐
9.	make decisions when needed.	☐	☐	☐
	Set 2			
10.	negotiate on behalf of the work unit for a fair share of organizational resources.	☐	☐	☐
11.	work with dissatisfied customers or clients to achieve a mutually agreeable solution.	☐	☐	☐
12.	help people in work groups settle interpersonal conflicts that interfere with group functioning.	☐	☐	☐
13.	help settle work-related problems, complaints, or disputes among employees or organizational units.	☐	☐	☐
14.	negotiate with people outside the organization to gain something of value to the organization.	☐	☐	☐
15.	mediate and solve disputes at individual, group, or organizational levels.	☐	☐	☐
16.	negotiate with people within the organization to achieve a consensus on a proposed action.	☐	☐	☐
17.	mediate conflict situations without taking sides.	☐	☐	☐

(Go on to next page)

FIGURE 2.3 A sample page from a personality-based inventory of general position requirements.

Putting It All Together—The Job Narrative

After the job analysis information has been collected and collated, it must be put together into a usable and useful format. The write-up of the job analysis information is referred to as the *job analysis narrative* or the *job description*; although to make matters more confusing, the term *job description* is often used to refer to the whole write-up as well as the section of the narrative dealing with the major work behaviors and tasks. The major sections of the job analysis narrative include the following:

- Identifying Information—Basic information such as job titles, job analysts, SMEs, and dates of study.
- Major Responsibilities—A paragraph describing the major responsibilities of the job.
- Job Description—A listing of the major work behaviors and the associated tasks; for selection purposes, it is also useful to list major textbooks or courses that cover the knowledge required for successful performance of the job.
- KSAs—A list of the KSAs corresponding to the major work behaviors and tasks.
- Job Specifications—Usually based on factors similar to those found in a job evaluation system, such as supervisory responsibility, supervision received, financial responsibility, effort required, and the working conditions.
- Tools and Equipment Used—A list of the common tools and equipment used on the job. May include information on weights lifted, carried, or dragged.
- Minimum Qualifications—The minimum education and experience required for the job along with the rationale for the minimum qualifications.
- Americans with Disabilities Act Information—A description of why certain major work behaviors are considered essential functions of the job.

Some Warnings About Job Analysis

- *Different sources of information may yield different information, at least some of it wrong.* Different sources yield different information. Observing one incumbent rather than another may get biased information. An unusually effective worker may do different things with different resources. People with strong verbal skills can describe tasks and resources more clearly than others—and, perhaps, say more to embellish their jobs. As noted earlier, be especially wary of developing an overreliance on the O*NET.
- *Using all of the complex information a job analysis provides is not necessary for accurate prediction.* Overall performance in any job, or any aspect of job behavior, can be optimally predicted by only a few predictors. After

one or two variables, further variables rarely make more than trivial contributions to predictive accuracy.

- *Job analysis tends to yield static descriptions of "the way we've always done it."* Job analysis typically describes the job as it is, not how it might be, ought to be, or will be in the future. Job analysis should, but rarely does, include planning for future contingencies and alternatives.
- *Job analyses rarely recognize alternative ways to do the job or to qualify for it*. Most jobs can be done in more than one way. More attention should be given to "if–then" hypotheses: If an applicant can be expected to do the job one way, then one set of attributes will provide the best predictors, but if the applicant is likely to do it differently, then a different set of attributes may be better.
- *Job analysis is typically descriptive, not prescriptive.* It might often be useful to describe *effective* ways to do a job. Differences in information from high performers and low performers can highlight the actions and personal resources that lead to effectiveness.
- *No one method of job analysis is clearly superior to another.* For personnel research, the purpose of job analysis is to understand the job well enough to form sensible, rationally defensible hypotheses about the characteristics of people that predict criterion variables of interest. That purpose is not likely to be optimally met by any one method, nor is it likely to be met if one uses any method or set of methods uncritically.
- *Describe the job, not the person.* It is very easy to fall into the trap of describing the person holding the job rather than the job itself. The purpose of the job analysis is to describe the job requirements, not the characteristics of the individual currently holding the job. This is not always easy, especially when you are dealing with single incumbent jobs. As an example, when analyzing jobs at Universities, it is not uncommon to find clerical jobs held by individuals with PhDs, even if the job requires only a high school degree. It is the responsibility of the job analyst to report that the minimum requirement for the job is a high school degree, regardless of the educational accomplishments of the incumbent.

A Recommended Procedure

Although we have indicated that there is no required approach to job analysis, on the basis of current developments in the field, we can recommend a multimethod approach to job analysis. In order to conduct a comprehensive job analysis for most jobs, we recommend the following eight steps:

1. Review existing information and O*NET.
2. Create and distribute a first draft of the job description to SMEs.

3. Conduct the first set of confirmatory job analysis interviews with SMEs.
4. Create and distribute the second, revised draft of the job description.
5. Create a job analysis survey or questionnaire based on the draft job description.
6. Distribute questionnaire to a sample of incumbents and supervisors.
7. Analyze the data from the questionnaire using Excel or statistical software.
8. Create a final version of the job description based on all the information and data.

General Caveats

Even the most careful job analysis is subjective. Job analysis is not science, even when it is used in scientific research or guided by scientific thought. It is an information-gathering tool to help managers or researchers decide what to do next. If developed well and used systematically, it yields reliable information that leads to defensible predictive hypotheses with strong likelihood of being supported empirically. The insight needed to choose predictors that improve organizational functioning is more likely if one acquires correct information through well-considered job analysis. Not every job analysis must be comprehensive or even thorough, but they must be well considered.

This chapter has barely scratched the surface of the topic of job analysis. Perhaps the most important caveat of all is to point out how much information is available that goes well beyond this chapter.

Competency Modeling

The term "competency modeling" has become widespread in the popular HR literature. The notion behind the competency modeling movement is that traditional or "old school" job analysis cannot meet the demands of the changing workplace. Although the meaning of the term *competency* remains unclear (Schippmann et al., 2000), the notion behind it is that one should identify the characteristics or attributes that are related to exceptional performance on the job.

Competency modeling, or a *competency analysis,* can be defined as *a technique for identifying competencies*. Just as there are many methods for job analysis, there is a range of methods that can be used to complete a competency analysis. For example, both interviews and surveys can provide useful information in competency modeling.

One of the questions that comes up in competency analysis is whether to work from the top down in an organization or to work from the bottom up. This leads to the following four (at least) different approaches that differ in their basic philosophies and can also lead to quite different results:

1. Start at the top of the organization by identifying general organizational goals and strategies, which leads to the identification of desired,

organization-wide competencies, and then work down through the organization for the purpose of performing a microanalysis of KSAs and behavioral exemplars associated with individual jobs. This would be the top-down approach.
2. Start at a micro or job level in order to identify competencies and then through aggregation and an analysis of common trends, work toward more general organization-wide competencies. This would be the bottom-up approach.
3. Adopt a standard set of competencies either from the literature or based on a model offered by a consulting firm. For example, you will see many references to Bartram's (2005) Great Eight competencies. Campbell, McCloy, Oppler, and Sager (1992) also offered a frequently cited list of basic job competencies. Many consulting firms offer their own set of competencies.
4. Have SMEs choose from a long list of competencies. Many consulting firms offer such lists, which are referred to by a number of names including banks or dictionaries. Even in this era of high technology, one can still find competency card decks, which are intended for easy sorting by SMEs; of course, there are also computerized versions of such card decks.

Products and Tools

Once the competency model is identified and information collected, the information must be translated into useable tools. In terms of our focus on selection, three important products from the competency analysis are as follows:

1. The Competency-Task-KSA Linkage Chart—This chart identifies the relations among various competencies and the tasks and KSAs from a traditional job analysis.
2. The Competency Measurement Matrix—This matrix suggests or recommends methods for measuring competencies. This matrix is particularly important in selection, as it suggests a number of methods for assessing competencies for either initial selection or promotion.
3. A Competency Rating Guide—This guide provides rating scales by linking various competency levels to specific behavioral exemplars. For example, for a competency such as Leadership, it is common to distinguish between levels such as entry-level, advanced, and expert. Another common series of levels is Below Expectations, Meets Expectations, and Exceeds Expectations. Such competency rating guides play an important role in selection as they can serve as the basis for interview questions as well as the development of interview rating scales. A simple, sample rating guide for Leadership appears in Table 2.5.

TABLE 2.5 A Competency Rating Scale for Leadership

Definition of Leadership: A leader motivates and drives others. A leader is willing to make decisions and resolve conflicts. A leader engages in ethical behavior.

Level of Competence	*Prototypical Behaviors*
Easily Exceeds Expectations and Performs at Expert Level	Seen by others as a leader. Is confident. Drives change and inspires others to take ownership of change. Motivates others to high levels of performance. Is willing to make decisions even when criticized by others or in times of crisis. Maintains high levels of ethical behavior and inspires others to do the same. Identifies conflicts before they occur.
Meets Expectations	Is seen by others as a leader in some situations. Makes a case that change is necessary and acts in a manner to try to get others to accept changes. Motivates others to meet set goals. Is willing to make decisions, especially when it is an area in which they have experience. Adheres to ethical standards in own behavior. Listens to others and resolves disagreements when they occur.
Below Expectations	Lacks confidence in self. Others lack confidence in the ability to lead. Fails to make a strong case for change. Actions fail to motivate others. Prefers to leave decisions to others. Does not always consider ethical issues. Avoids dealing with conflicts.

What is a ***competency***? Unfortunately, there is no simple answer. As an attempt at definition, we would argue that a competency is *a worker-oriented higher level characteristic, which can be linked to associated KSA clusters and defined in terms of specific behavioral exemplars, which thus allow for operationalization and measurement*.

Interested in learning more about competencies? The United States Government offers an excellent site that describes in detail the process of developing competency models. You can find this information at www.careeronestop.org/CompetencyModel. Go to this site and explore; the great aspect of working online is that you cannot break anything.

A Quick Caveat

Sackett and Laczo (2003) observed that competency modeling's entire *raison d'être* is the misguided assumption that job analysis focuses only on tasks done on the job, and not on the attributes required for success on the job. Indeed, there are many examples of worker-oriented methods presented in this chapter. Unfortunately, existing methods of competency modeling often fall short in terms of methodological rigor (see e.g., Lievens, Sanchez, & De Corte, 2004). In addition, many organizations find that general competency models do not provide sufficient detail for selection purposes and, in response, move to methods that look a great deal like traditional job analysis.

Discussion Topics

1. What is the difference between an organizational analysis and a job analysis?
2. Given the numerous methods available for job analysis, what are some considerations in deciding which method to use? What method might you use with a plant manager, where there are only two incumbents? What method, or methods, would you use to analyze the job of firefighter in a large city with 5,000 incumbents and a history of litigation over hiring and promotion?
3. What is the role of the job expert, or SME, in the job analysis process? How would you go about selecting SMEs?
4. In what ways do you believe competency analysis differs from job analysis? What are the advantages of competency analysis as compared with job analysis?

Notes

1 Vicki V. Vandaveer suggested these illustrative questions.
2 The Position Analysis Questionnaire is available from PAQ Services Inc.

References

Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90,* 1185–1203.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1992). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco, CA: Jossey-Bass.

Cooperrider, D. L., & Srivastva, S. (1987). Appreciative inquiry in organizational life. In W. Pasmore & R. Woodman (Eds.) *Research in organizational change and development* (Vol 1, pp. 129–169). Greenwich, CT: JAI Press.

Costa, P. T, McCrae, R. R., & Kay, G. G. (1995). Persons, places and personality: Career assessment using the revised NEO Personality Inventory. *Journal of Career Assessment, 3,* 123–139.

Cranny, C. J, & Doherty, M. E. (1988). Importance ratings in job analysis: Note on the misinterpretation of factor analysis in industry and the public sector. In S. Gael (Ed.), *The job analysis handbook for business, industry and government* (Vol. 2, pp. 1051–1071). New York, NY: Wiley.

Fine, S. A., & Cronshaw, S. F. (1999). *Functional job analysis: A foundation for human resource management*. Mahwah, NJ: Lawrence Erlbaum Associates.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327–358.

Gottfredson, G. D., & Holland, J. L. (1994). *Position Classification Inventory*. Odessa, FL: Psychological Assessment Resources.

Harvey, R. J. (1993). *Research monograph: The development of the CMQ*. San Antonio, TX: The Psychological Corporation.

Inwald, R. (1992). *Hilson Job Analysis Questionnaire*. Kew Gardens, NY: Hilson Research.

Jeanneret, P. R., & Strong, M. H. (2003). Linking O*NET job analysis information to job requirement predictors. An O*NET application. *Personnel Psychology, 56,* 465–492.

Landy, F. J. (1989). *Psychology of work behavior*. Pacific Grove, CA: Brooks/Cole.

Levinson, H. (2002). *Organizational assessment: A step-by-step guide to effective consulting*. Washington, DC: American Psychological Association.

Lievens, F., Sanchez, J. I., & De Corte, W. (2004). Easing the inferential leap in competency modeling: The effects of task-related information and subject matter expertise. *Personnel Psychology, 57,* 881–904.

McCormick, E. J. (1959). Applications of job analysis to indirect validity. *Personnel Psychology, 12,* 402–413.

McCormick, E. J. (1979). *Job analysis*. New York, NY: AMACOM.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1969). *The development and background of the Position Analysis Questionnaire* (Contract NONR-1100(28), Report No. 5). West Lafayette, IN: Purdue University, Occupational Research Center.

Morgeson, F. P., Delaney-Klinger, K., Mayfield, M. S., Ferrara, P., & Campion, M. A. (2004). Self-presentation processes in job analysis: A field experiment investigating inflation in abilities, tasks, and competencies. *Journal of Applied Psychology*, *89*, 674–686.

Peterson, N. G., Borman, W. C., Hanson, M. A., & Kubisiak, U. C. (1999). Summary of results, implications for O*NET applications, and future directions. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), *An occupational information system for the 21*st *century: The development of O*NET* (pp. 289–295). Washington, DC: American Psychological Association.

Raymark, P. H., Schmit, M. J., & Guion, R. M. (1997). Identifying potentially useful personality constructs for employee selection. *Personnel Psychology, 50,* 723–736.

Rounds, J. (1995). Vocational interests: Evaluating structural hypotheses. In D. Lubinski & R. V. Dawis (Eds.), *Assessing individual differences in human behavior* (pp. 177–232). Palo Alto, CA: Davies-Black.

Sackett, P. R., & Laczo, R. M. (2003). Job and work analysis. *Handbook of Psychology: Industrial and Organizational, 12,* 21–37.

Sanchez, J. I, & Levine, E. L. (2001). The analysis of work in the 20th and 21st centuries. In N. Anderson & D. S. Ones (Eds.), *Handbook of industrial, work and organizational psychology: Volume 1. Personnel Psychology* (pp. 71–89). London, England: Sage.

Schippmann, J. S., Ash, R. A., Battista, M., Carr, L., Eye, L. D., Hesketh, B., . . . & Sanchez, J. I. (2000). The practice of competency modeling. *Personnel Psychology, 53,* 703–740.

United States Department of Labor. (1972). *Handbook for analyzing jobs*. Washington, DC: U.S. Government Printing Office.

United States Department of Labor. (1977). *Dictionary of occupational titles: Definitions of titles* (4th ed.). Washington, DC: U.S. Government Printing Office.

Van de Ven, A. H., & Ferry, D. L. (1980). *Measuring and assessing organizations*. New York, NY: Wiley.

3

DEVELOPING PREDICTIVE HYPOTHESES

Conceptual and Operational Definitions, Criteria, and Predictors

A predictive hypothesis concerns the relation of one variable, a *criterion,* to at least one other, a *predictor.* In selection, the predictor is usually, but not always, some type of test or assessment device. The criterion is usually job performance, although other possibilities such as turnover or salary exist. A predictive hypothesis is not a universal truth; it may be expected to hold only within certain boundary conditions. Specifying the operational hypothesis and its boundaries requires job and organizational knowledge and, beyond that, knowledge of psychology, psychological research, and psychometric tools.

Conceptual and Operational Definitions

Predictors and criteria can be defined at two levels, *conceptual* and *operational.* Hypothesis formation should ordinarily begin with simple statements about how predictor and criterion *constructs* relate. For example, one expects, for an electronics assembly worker, "Quality of performance is a function of ability to make fine, precise manipulations of small objects." Performance quality is a criterion construct; the predictor construct is the ability to make fine, precise manipulations of small objects. This theoretical statement is not testable because no measurement operations are specified. If the predictor construct is conceptually defined as finger dexterity, its operational definition may be a score on a standardized test of finger dexterity.

A ***construct*** is simply a general idea about something that is unobservable. Each predictor construct can be defined at a *conceptual* level (e.g., strong work ethic) and at an *operational* level (e.g., scores on a self-report measure of work ethic).

Some predictors are used even when the construct underlying them is unknown. Predictors defined only at the operational level can still be effective (one company reportedly found that high-performing truck drivers had more tattoos than their lower performing colleagues). We do not disparage predictors that work (although, we may draw the line at counting tattoos). We believe, however, that finding out what they mean can promote understanding and, eventually, better measures.

Synergy of Theory and Practice

Professional practice is the hallmark of applied psychology, but continued application without understanding never progresses. Theory is understanding—or the attempt to understand. It is more than a hunch. Binning and Barrett (1989) pointed out that developing a predictive hypothesis requires both theory building and theory testing. With some modifications in terms and numbering, we follow their presentation in Figure 3.1.

1. A predictor construct—an idea of a way applicants differ from one another—is related to a criterion construct, usually the ultimate criterion of true job performance or total worth to the organization. This is the basic theoretical statement at a conceptual level.
2. Predictor measure *X* is related to criterion measure *Y,* a relation expressible mathematically, usually by a correlation coefficient. This is a predictive hypothesis and, unlike #1, is empirically testable.
3. Predictor measure *X* is a valid measure of, or reflection of, the predictor construct. For example, if our predictor construct is cognitive ability, we could use a test such as the *Wonderlic Personnel Test* as an indicator of the underlying construct.

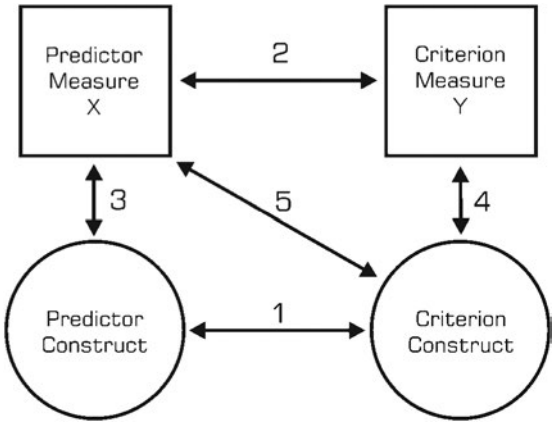

FIGURE 3.1 Basic linkages in the development of a predictive hypothesis.

From Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478–494. Adapted with permission.

4. Criterion measure *Y* is a valid measure of, or reflection of, the job performance behavior construct. For example, the job performance construct could be assessed, and frequently is, using supervisor ratings of observed employee behaviors. Please note that evidence with regard to tests of Inferences 3 and 4 is provided through both reliability studies and construct validation. In Chapter 5 we discuss the topic of reliability and also introduce the concept of *psychometric validity,* which attempts to operationalize the idea of construct validity.
5. The predictor measure *X* is related to the criterion construct in a manner consistent with its presumed relation to the criterion measure *Y*. The truth of this inference depends on the validities of Inferences 1 and 3.

Much can be added to this framework. Binning and Barrett (1989) have, in fact, done so, but we will stop here; Inference 5 is the basic operational hypothesis. It rests on the reasonableness of Inference 1, the inferred relationship between two hypothetical constructs not directly measurable, and of Inference 3, the construct validity of the predictor as used. A framework like that proposed by Binning and Barrett (1989) is theoretical in calling for understanding, beginning with the constructs. When one can define the constructs with some clarity, one has an idea *why* assessment of a certain trait is likely to predict subsequent employee performance. If, in fact, it does not predict as logically expected, one must change one's understanding of the criterion, the predictor, or both; progress ensues.

Specification of Population

To whom does the predictive hypothesis apply? Anyone? Only experienced applicants? New entrants into the labor pool? People with required credentials (e.g., degrees or licenses)? In short, who is an applicant? The question has both legal and technical implications. Definitions of applicant populations are elusive, but the basic idea is a population to which research results should generalize. Population boundaries may be defined by prior conditions such as required credentials or passing a screening test.

This is where practice and theory clash. Most theories assume a highly artificial world in which there is random sampling from a population. However, in real life, people are not randomly sampled. People make informed choices to apply for jobs based on available information.

> In order to illustrate the nonrandomness of applications, consider applications of undergraduate psychology majors to graduate program in Industrial/Organizational (IO) Psychology. The typical student spends a great deal of time studying details of various programs—what are the minimum and average Graduate Record Exam (GRE) scores of applicants and of students accepted, what are the minimum and average Grade Point

Averages (GPA), how many people apply, how many people are accepted, and what the costs are in time and money to apply. Based on careful consideration of the available information, the prospective student decides what program to apply to and what programs are not worth spending time in order to submit an application. As a result, the applicants to programs such as Bowling Green State University and The University of Akron are anything but a random sample of all undergraduate psychology majors. They are a very select, nonrandom sample. (However, by a twist of logic that is too involved to explain here, we can still treat that select sample as a random sample from some population, even if we have no idea how to describe the underlying population. At this point, we start to enter into ground covered in philosophy of science or research methods classes, involving topics such as internal validity, external validity, the appropriate use of statistics, and the generalizability of results.)

Specification of Time Intervals

Usually (not always), criteria collected early—after a few months or perhaps a couple of years—are better predicted than those collected after longer intervals. The idea that the validity of predictors may change as a function of the amount of time someone spends on the job is at the heart of the dynamic criterion debate (Barrett, Alexander, & Doverspike, 1992; Barrett, Caldwell, & Alexander, 1985; Hulin, Henry, & Noon, 1990). Murphy (1989) suggested that validities and most valid predictors change with shifts in career stage from a *transitional stage* of new learning to a *maintenance stage* of doing more or less routinely what had been learned. Cognitive variables, for example, may be better predictors of performance in transitional stages and motivational predictors better for maintenance stages. Helmreich, Sawin, and Carsrud (1986) found that achievement orientation did not predict performance well until after a "honeymoon" period on the job. The idea is that employees, like new spouses, are usually on their best behavior for the first few months. In time, however, employees (and spouses) revert to their characteristic ways.

A study of pharmaceutical sales representatives found that agreeableness and openness (aka intellectance) predicted performance during the transitional period, but only conscientiousness predicted sales during the maintenance period (Thoresen, Bradley, Bliese, & Thoresen, 2004). For some jobs, the learning period may go on and on; Ghiselli (1956) identified an investment broker job in which performance improved linearly for six years. Cognitive ability may continue to predict performance for jobs with ever-changing tasks (Farrell & McDaniel, 2001). In generating hypotheses, one must decide whether they refer to predictions of performance during an early learning period, a later maintenance stage, or long-term career growth and development. The time interval can be approximate, but it should make sense.

Specifying Functional Relations

The term "functional relation" implies that the level of one variable (usually *Y*) varies "as a function of" variation in another. The nature of the relation (i.e., the function) may ordinarily be expressed as a mathematical equation. Functions are discussed in more detail in Chapter 6, but discussion of two issues should not wait.

One is that predictive hypotheses usually assume (deliberately or by default) a linear function (one graphed as a straight line throughout the predictor range) as the relation between predictor and criterion. There are good reasons for the assumption, but there are also reasons for considering alternative functions. Consider the following example: We may hypothesize that more educated bank tellers will stay in the job longer than less educated ones. The reasoning might be that less-educated tellers will make more mistakes and become more frustrated. It might be that more educated tellers had enough persistence to complete further schooling, and that persistence may also keep them on the job.

Figure 3.2 shows three examples of simple functional relations plausible for various kinds of predictions. Panel *a* describes the common linear function in which any difference in *X* always has a corresponding difference in *Y*; that is, adding a point to a

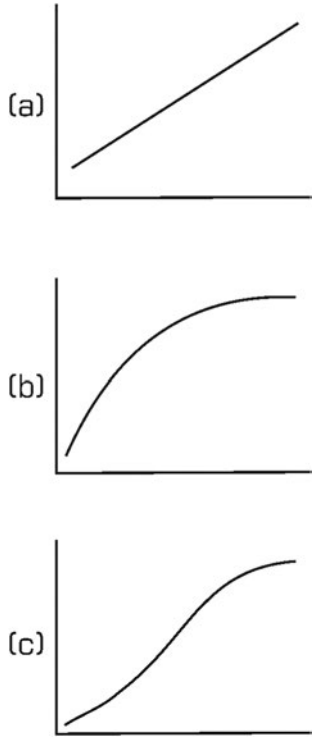

FIGURE 3.2 Three examples of functional relationships.

score implies the same added level of criterion performance whether the point is added to a low score, a moderate one, or a high one. In panel *b*, this is not true in the higher predictor levels; adding a point to the lower predictor scores is associated with a bigger criterion difference than at the higher levels where the curve may be asymptotic to some criterion level. Panel *c* shows a similar loss of advantage at both lower and higher score levels; that is, differences in actual predictor levels in either a low-scoring or a high-scoring range have only trivial counterparts in criterion performance, whereas predictor differences in the middle range are associated with substantial criterion differences. These are by no means the only functional relations that may be plausible; others may also deserve consideration. Consider the bank teller example earlier; it may be that educated tellers have better job opportunities elsewhere, or may quit because the job routine is boring. Whereas our first hypothesis was based on a linear positive relation between education and tenure on the job, further consideration led to a revised, curvilinear hypothesized relation. Failure even to think of alternatives to the linear function means failure to test them. Although linear functions will ordinarily be specified, they should be specified intentionally, not by default.

Criteria

The word *criterion,* for a predictive hypothesis, is simply a "dependent variable, the variable to be predicted" (English & English, 1958, p. 130). There is little virtue in trying to predict a criterion nearly everyone (or no one) does well; it is not useful, and it will not work. Prediction of individual levels of criterion performance requires individual differences—variance. For instance, if everyone who is hired for a specific job leaves within one year on the job, then it makes no sense to identify individual differences related to turnover—everybody turns over. If nearly everyone is at the low end of the scale, some other intervention is needed, such as better pay, job redesign, or recruiting from a different applicant pool.

Criterion Constructs

Criteria are measures of behaviors or events that are important to the organization. Too often, they are simply accepted as givens, without much concern for their meaning. Clarity of the constructs they represent provides clarity for the meaning of predictions.

Inferring Constructs From Measures. Events worth counting, recording, and predicting may include accidents, quitting, completion of training, or receipt of letters of commendation (e.g., letters from the public praising something done by a police officer or by a truck driver). The meanings of such measures are often unclear.

Absence (or absenteeism) provides a useful example. What does it mean to count the number of days at (attendance) or away from work (absences) over a given period? Psychologists once interpreted absence only as withdrawal from an aversive situation.

In this view, being late is a mild form of withdrawal, being absent is a stronger form, and quitting is the ultimate withdrawal from the job or organization.

Reasons for absence may not be readily apparent. Even classifying absences as necessary or avoidable is difficult. Smith (1977) computed average scores on six attitude scales for organizational units in Chicago and in New York. Scores were correlated with percentages of people attending work on a certain day, which happened to be the day after a severe snowstorm in Chicago; weather was no problem that day in New York. Mean attitude scores did not correlate well with attendance in New York, but did in Chicago. Is attendance a matter of attitude? Only, it seems, if it requires more effort than mere habit.

This example shows that the meaning of attendance or absence is unclear. Why would an organization want to predict either? One reason is economic: Absence is expensive. But before the cost, absence is psychological. Withdrawal or escape from work is psychologically interesting, but is it clearly indicated by absence? Probably not. Perhaps a trait construct such as *acceptance of responsibility* is one reason organizations worry about absenteeism, but absenteeism may not measure it very well.

> Most organizations are not concerned about truly uncontrollable reasons for absence as much as they are concerned about employees skipping work for no good reason. Thus, the *conceptual* definition may be voluntary absenteeism, but the *operational* definition may be number of days of work missed.

Starting with the measure (e.g., counting absences) and then trying to determine what it means is the wrong way to go. It makes more sense to decide first on the criterion concept; only when the concept is reasonably clear can a measure of it be tried and evaluated. With such a complex construct as responsible work behavior, a composite of several measures (maybe including attendance) may be more valid. Predicting one component, in short, may be less useful, and less well done, than predicting a *pattern* of behaviors tapping a common and clearly defined construct.

A Theory of Performance. Performance is a construct, measurable in many ways. Campbell, McCloy, Oppler, and Sager (1993) defined performance as cognitive, motor, psychomotor, or interpersonal behavior controllable by the individual, relevant to organizational goals, and scalable in terms of proficiency. Work outcomes (such as production), effectiveness (evaluation of outcomes), and productivity (an aggregate, not an individual, measure) are not part of their definition. Performance is work-related activity—behavior. Note that Campbell et al.'s (1993) definition differs from that given by the Society for Industrial and Organizational Psychology (SIOP) "the effectiveness and value of work behavior and its outcomes" (SIOP, 1987, p. 39).

Regardless of the definition, performance is not unidimensional; it may have many components. Ranking employees by level of proficiency in one component may not match their rank order on another. Supervisors, however, may have

difficulty differentiating between different performance dimensions or factors. As a result, even when we believe that performance dimensions are unique, the correlations between different dimensions when rated by supervisors may approach one. There is a strong counterargument to the multidimensional view of performance that posits that the ultimate criterion for employee performance is the worth of that performance to the organization, which can be expressed in or converted into money or dollars.

Performance Components and Determinants. Campbell et al. (1993) postulated three determinants to account for proficiency in any performance component: ***declarative knowledge***, factual knowledge and understanding of things one must do; ***procedural knowledge***, skill in knowing how to do them; and ***motivation***, the direction, degree, and persistence of effort in doing them (see Kanfer & Ackerman, 1989, for background). In the workplace, both declarative and procedural knowledge may be combined as job knowledge; those with a wealth of knowledge that can be tapped by others are valued highly, but those for whom that knowledge is not accompanied by actual skill in applying it are often dismissed, somewhat contemptuously, as merely "talking a good job." The theory also suggests eight general factors of performance, although in a 2012 handbook chapter, John Campbell offered an expanded view of the possible number of factors underlying job performance (Campbell, 2012). The number eight is interesting because it seems to underlie a number of models of job performance. For example, Bartram offered a model of performance based on eight great competencies (Bartram, 2005).

In Table 3.1, we offer our own interpretation of eight critical factors that serve as the basis of job performance, As with other models, not all of the factors are relevant to every job (e.g., many jobs have no supervisory component, although it could be argued that leadership is relevant for all positions in an organization).

Contextual Behavior. Valued behavior at work includes more than doing assigned job tasks. Regularly coming to work on time, staying with the organization rather than leaving, staying overtime on short notice when unexpected problems arise, helping others when needed, minimizing or solving conflicts within the work group, training or mentoring newcomers, justifying trust, or simply providing a good model for others—all form part of the context in which work is done.

> ***Contextual performance*** refers to aspects of performance unrelated to specific tasks. These include activities directed at enhancing the interpersonal and psychological environment that facilitates task completion. Other similar concepts include organizational citizenship behaviors (Organ, 1988), prosocial behaviors (Brief & Motowidlo, 1986) and extra-role behavior (Organ, Podsakoff, & MacKenzie, 2006) as well as the converse, counterproductive work behavior (Dalal, 2005; Sackett, Berry, Wiemann, & Laczo, 2006).

TABLE 3.1 An 8-Factor Model of Job Performance and Competence

Factor	*Definition*
1. Task Performance and Job Knowledge	Understands factual and procedural information required for the job and uses it to accomplish goals
2. Motivation and Effort	Shows direction and persistence in working toward goals as well as a true desire to accomplish them
3. Quality of Work	Completes tasks efficiently according to protocol and with minimal errors
4. Communication Ability	Conveys thoughts and ideas clearly and concisely to others
Oral	Conveys ideas clearly through the spoken word (e.g., one-on-one conversations, presentations, discussion facilitations)
Written	Conveys ideas clearly in a written format (e.g., e-mails, reports, presentation materials)
5. Leadership	Assumes control, takes initiative, sets an example
Supervision	Delegates tasks, provides direction, coaches, and gives feedback
6. Administrative Competence	Follows protocols and procedures, meets deadlines, and prioritizes tasks
7. Teamwork	Cooperates, collaborates, and shares responsibilities equally with others
8. Emotional Intelligence	Perceives, understands, manages, and uses emotions in interactions with others

Borman and Motowidlo (1993), calling such things *contextual* activities, differentiated them from task or job performance in four ways: (1) task activities contribute directly to the technical core of an organization's production of goods and services, contextual activities contribute to the organizational or social environment in which that technical core functions; (2) task activities differ across different jobs, contextual activities are common to many if not all jobs; (3) task activities are associated with skills or abilities, contextual activities are more associated with motivational or personality variables; and (4) task activities are things people are hired to do, contextual activities are desirable but less likely to be demanded.

A first question in choosing a contextual criterion is to ask whether it represents organizationally required behavior for everyone, or at least everyone within a specified group of jobs, or whether it is merely desirable. A second question is whether the desired behaviors are more likely and more safely elicited from day-to-day managerial influence than from antecedent traits.

Trainability. How quickly tasks are learned may be an important construct, especially where people must frequently adapt to changing technology or assignments. Even on static jobs, the idea that anyone can become expert given enough time is a myth; those who need long learning time for complex tasks generally do not reach the level of proficiency after training reached by those who learned more quickly. Selecting or promoting people who will learn their duties quickly or adapt quickly to job changes is organizationally useful.

Status Quo, Change, and Criterion Choice

Organizations need to grow and adapt through change. Criteria should promote effective change, maintain useful stability in the face of change, and help develop an organization that continues to function effectively in a changing world. William Whyte (1957), in his classic book *The Organization Man,* criticized U.S. organizations that, through employment testing, choose conformers who resist change. Many of his criticisms were off target but, in fact, many criteria, if predicted well, tend dangerously to maintain the status quo. For example, the personality factor "conscientiousness" is widely advocated for selecting cooperative employees, but we have seen no discussion of how selection on this trait may create a workforce of followers. Avoiding criteria that merely reinforce the status quo requires intelligent recognition of the inevitability and usefulness of change.

Predictors and Predictive Hypotheses

What variables are likely to predict the criterion? How should they be measured? Forming a predictive hypothesis is a two-part logical argument: First, that the criterion is related to certain traits and, second, that the chosen predictors are valid measures of those traits.

> The ***predictive hypothesis*** involves (1) identifying what traits relate to the criterion, and (2) reasoning that certain measures are valid measures of these traits.

Constructs are inferred from knowledge of the job or organization. People with different backgrounds may infer different constructs. Psychologists may choose constructs from factor analysis or general theories. Managers and job incumbents may rely on their experience, using what Borman (1987) in another context called *folk theories*; if psychologists ignore ideas based on such experience, they risk ignoring some very good bets for predictors. Good hypotheses depend on prior knowledge and logic. One needs to know what has worked before, and what has failed to work in similar situations. Some things are well established; for example, job performance is predicted better by abilities than by other traits, and cognitive abilities predict better than noncognitive abilities, for most jobs. Although many people find it hard to believe, the same factors that lead to success in school also seem to lead to success in one's career (Kuncel, Hezlett, & Ones, 2004).

Cognitive Factors

Cognitive abilities are abilities to perceive, process, evaluate, compare, create, understand, manipulate, or generally think about information and ideas. Common work-relevant cognitive activities include reading verbal or graphic materials,

understanding the principles that make things work, planning events or procedures, solving problems, or perceiving signs of trouble in equipment or in human interactions or in contradictions in plans. Mental abilities are diverse and somewhat overlapping. More than 75 years of factor analytic research, however, has clarified and defined many components of mental abilities.

Factor analysis identifies latent or underlying factors in the responses to measures. Verbal ability and quantitative ability are two of the (latent) factors underlying responses to items on tests such as the GRE.

Factor analysis examines intercorrelations among measures to identify or infer underlying (latent) traits accounting for the correlations. Several lists of mental abilities have been based on factor analyses, beginning with the Thurstone (1938) list of seven primary mental abilities: verbal comprehension, word fluency, spatial ability, perceptual speed, numerical facility, memory, and inductive reasoning. Subsequent research has made finer distinctions. Differences in the measures analyzed, in methods of analysis, and in the focus of researchers will result in slightly different factors and differences in their specificity. The possibilities for differences make it impressive that factor analytic results have been as much alike as they are. If every nuance is treated as a difference, the lists of factors across studies can be very large; unfortunately, in many areas of selection we lack agreed-upon taxonomies, with the one notable exception being the Big 5 in personality measurement. We will not try to list and define them all, but offer an illustrative (*not* definitive) list of frequently recurring cognitive factors in Table 3.2.

General Mental Ability

Even the lowest correlations between measures of mental ability are positive, suggesting a general mental ability. Traditionally, the general trait is called *intelligence.* Different authors, with different perspectives, emphasize different features of behavior called intelligent. One of the better definitions was given by Humphreys (1979): *"Intelligence is the resultant of the processes of acquiring, storing in memory, retrieving, combining, comparing, and using in new contexts information and conceptual skills; it is an abstraction"* (p. 115).

Spearman (1927) emphasized a general intellectual ability he abbreviated *g,* a symbol once again in frequent use. Cattell (1963) argued that the general factor had two components: *fluid intelligence,* involving basic reasoning, and *crystallized intelligence,* measured by tests such as vocabulary. Carroll (1993), in an encyclopedic reanalysis of mountains of factor analytic studies, proposed a three-stratum model.

TABLE 3.2 An Illustrative List of Cognitive Factors Identified via Factor Analysis

Factor Label	*Description*
Verbal Comprehension	The ability to understand words and their meanings, and to apply such understanding in verbal communications. Nearly any mental ability test will include a verbal comprehension score in some form.
Fluency	The ability to produce quickly a lot of ideas or associations. Different jobs may require different kinds of fluency, such as verbal fluency, ideational fluency, or number fluency.
Perceptual Speed	The ability to identify figures, make comparisons, or match visually perceived figures quickly and accurately. Perceptual speed is a generally useful predictor of clerical performance.
Spatial Orientation and Visualization	The ability to perceive spatial patterns, to orient oneself or an object relative to other objects in space. Engineers, mechanics, and others who must work from drawings need such abilities; drivers, pilots, or others who plan trips probably need it as well.
Number Facility	The ability to do elementary arithmetic operations quickly and accurately. This is an obvious requirement for jobs requiring arithmetic computation.
General Reasoning	The ability to understand relational principles among elements of a problem and to structure the problem in preparation for solving it.
Problem Recognition	The ability to tell from early and perhaps subtle warnings that something is wrong or likely to go wrong; the problems may develop in equipment, people, social systems, or data. Sensitivity to potential or existing problems seems useful in jobs such as physician, air traffic controller, or machinery operation or monitoring.
Associative Memory	The ability to recall bits of information previously associated with unrelated information; for example, to remember numerical information associated with names.
Span Memory	The ability to recall in proper sequence a series of items (numbers, words, symbols) after a single presentation of the series; for example, looking up a telephone number and remembering it.

The first stratum consisted of the first order factors, somewhat like (but more narrowly and precisely defined) the list offered in Table 3.2. The second stratum included more general factors such as fluid and crystallized intelligence and others, and the third corresponded pretty well to Spearman's concept of *g*.

Job-Specific Knowledge and Skill

"Know-how" is a folk construct. People who have it—who know and understand thoroughly a job's requirements—are better workers than those who do not have it. To be useful, the term needs cleaner definition. Job knowledge may be general or limited to specific kinds of information or skill. The O*NET system of occupational information lists skills in three categories: *basic, cross-functional,* and

occupation-specific. Basic skills are capacities developed over a relatively long period of time that promote or provide a foundation for learning other types of material. Cross-functional skills are those useful in a wide range of occupations. Occupation-specific skills, of course, focus on tasks specifically required in occupations or jobs. The three categories are segments of a continuum from general to specific. The nearer a particular skill or knowledge is to the basic or general end of the continuum, the more likely it is to be expected of all qualified candidates; the nearer to the job-specific end, the more likely it is to be the content of in-house training programs. For organizational entry, hypotheses usually emphasize more general skills and knowledge; for promotions, they may emphasize skills and knowledge specific to the work to be done.

Personality Constructs

Personality is a mixture of values, temperament, coping strategies, character, and motivation, among other things. Compared with cognitive traits, conceptual definitions of personality traits can be developed more easily for particular jobs or purposes, but finding operational definitions to fit them is more difficult. From the 1960s to the 1980s, research on personality predictors was sparse. Some people attributed the demise of such research to a critical survey of the published research on personality testing (Guion & Gottier, 1965). Two other influences were probably greater. First, the Civil Rights Act of 1964 explicitly permitted the use of "professionally developed ability tests," but it included no such enabling statement for personality inventories. Such inventories were targeted for severe social criticism, so many employers quietly stopped using them in fear of litigation. Second, the views of Mischel (1968), insisting that behavior is determined more by situations than by traits, were widely accepted. The idea of personality traits was widely abandoned by psychologists, but trait psychology, which never fully disappeared, reappeared in the 1980s.

A ***personality trait*** is a habitual way of thinking or doing in a variety of situations. It may be a general value, goal, or behavioral tendency to seek or avoid certain kinds of situations. It might be a need, even a metaphorical need, for a goal (e.g., need for interpersonal affiliation). It may be a role that one habitually plays—the role of leader, clown, scholar, or teacher. It may be a constellation or combination of traits, a syndrome or type. The O*NET taxonomy does not refer to personality traits but "occupational values" and "work styles."

Most personality inventories measure several traits; if the list of traits named in them were placed end to end, it would stretch far! Consider the variety of constructs implied in this partial list of names of scales in existing measures of personality: alienation, anxiety, coping styles, emotional empathy, hopelessness, level of aspiration, perceptions of daily hassles and uplifts, response style, rigid type, risk-taking orientation, self-confidence, self-esteem, stress tolerance, team builder, Type A, and vigor. So many possible constructs must overlap; they require some means of reduction, commonly factor analysis.

The Five-Factor Model. Languages have thousands of words describing people's traits. Many words have overlapping meanings; for example, *timid, shy, nervous,* and *irresolute* all describe people who tend to falter in social situations. Meaningful distinctions can be made among these terms, but the more general idea of social faltering can be inferred from the similarities. The example is an "arm chair" or intuitive factor analysis. Actual (statistical) factor analysis applied to such descriptive words has often resulted in five factors (e.g., McCrae, 1992). The five-factor solution, sometimes called the Big Five, has been found in languages other than English and in using different measurement techniques, including adjective checklists, phrases, and even a nonverbal approach. It has for some time dominated personality research; Goldberg (1993) described the domination as an "emerging consensus."

Labeling the Big Five

Names given to the factors by various researchers have differed. Some differences in preference can be attributed to bipolarity, with some names describing the positive and others describing the negative end of a bipolar scale. Generally, however, name differences seem to reflect the different nuances different researchers think most worthy of emphasis.

Our preference for Factor I is ***Surgency***. It suggests the interpersonal aspect associated with extraversion, the common alternative, but it also includes the dominance and in-your-face visibility implied by wave-like "surging"; it is partly defined by adjectives such as *aggressive, assertive, unrestrained, daring,* and even *flamboyant*.

For Factor II, we prefer ***Agreeableness***. It encompasses terms like *likeability* or *friendliness* without putting much emphasis on conformity or compliance or implying emotional attachment to others.

For Factor III, we prefer ***Conscientiousness***; it seems the most relevant to the work context. One set of key terms identified by Hofstee, de Raad, and Goldberg (1992) includes *organized, neat, precise, exacting*; another includes terms like *conscientious, responsible,* and *dependable* clustering together.

For Factor IV, we prefer ***Emotional Stability***. It is a familiar term, measured well by many inventories, positive rather than negative; it seems to generate no controversy, and it frequently has been a valid predictor.

Naming Factor V is not merely a matter of preference; substantive differences exist in the factors identified. "Openness to experience" is substantively different from "intellect," and neither reflects the central traits very well. We believe the most useful term is ***Intellectance***; it is a liking for thinking about things, whether within the culture or in personal experience: problems to be solved, or things to be created.

Conscientiousness and Integrity. Factor III merits special attention. Employee theft of cash or merchandise is common enough that it has led to the use of tests to screen job applicants for honesty or integrity. These are not easy constructs to define. At first they seem to mean "theft potential," but that is too narrow. A person of integrity is not simply not a thief but a person whose word can be trusted, whose work is reliably or dependably performed even without monitoring, who, in short, can be counted on to do the right or good thing. A closer look at honesty testing and related validation research suggests the broader construct. Some test publishers have called their instruments predictors of "counterproductive behavior," but perhaps the more common, and more positive, construct would emphasize the *dependability* or *trustworthiness* aspects of conscientiousness (Goldberg, Grenier, Guion, Sechrest, & Wing, 1991).

Predictors under *conscientiousness* have been found to have the highest, albeit modest, mean correlations with job performance (Barrick, Mount, & Judge, 2001). Other factors have stable modest correlations within specific occupational groups. For example, agreeableness and emotional stability have been found to predict performance reliably in customer service jobs (Hurtz & Donovan, 2000). Thus, the five factors appear to have some utility for predicting performance, particularly when they are not correlated with other predictors in a selection battery.

Sackett and Walmsley (2014) set about identifying the personality traits that are most important in the workplace, in general. To do so, the authors examined (a) predictive validity of traits in the five-factor model, (b) the attributes that employers found most important to evaluate applicants in the job interview, and (c) attributes rated as important (i.e., using O*NET) for performance in a broad sample of occupations across the U.S. economy. The authors concluded that attributes related to Conscientiousness and Agreeableness were nearly universal in importance for workforce readiness in a variety of occupations. The importance of the other factors in the five-factor model appeared to differ upon specific circumstances.

Some critics say important personality constructs are not included in the Big Five. A popular alternative framework is the HEXACO model (Ashton, Lee, & de Vries, 2014). This model considers Honesty–Humility as a sixth personality factor.

Commentary. The five factors may be too broad for personnel assessment. Funder (1991) asserted that global traits (like the five) are best for explanations and theory development, but that in prediction, narrower trait constructs are better. In choosing cognitive and psychomotor constructs, the trend favors more general ones, but for personality, the trend seems to be toward greater specificity, favoring constructs more explicitly related to specific aspects of work. Examples include specific orientations such as *work orientation* (Gough, 1985) or *service orientation* (J. Hogan, Hogan, & Busch, 1984), or the breakdown of the Type A personality construct into two narrower constructs, achievement striving and impatience–irritability, each

predicting different sorts of outcomes (Spence, Helmreich, & Pred, 1987). Even the *Hogan Personality Inventory* (R. Hogan & Hogan, 1992), heavily influenced by the five-factor model, includes seven scales, not five. It divides surgency into two components, ambition (the surgency emphasis) and sociability (the extraversion emphasis); intellectance is divided into one called intellectance and another emphasizing school success, liking academic pursuits and achievements. There is good reason to question the adequacy for applied purposes of the five broad factors.

Identifying ***job-relevant personality traits*** involves specifying required behaviors of the job and choosing traits that reflect consistencies in those behaviors.

Physical and Sensory Competencies

The Americans with Disabilities Act (ADA) and the Civil Rights Act of 1964 (as amended) have dampened what little enthusiasm existed for physical and sensory competencies for personnel decisions. For many kinds of work, however, they are potential predictors of performance and may be genuine prerequisites for some jobs and, therefore, defensible in litigation.

Physical Characteristics. Physical traits *can* be relevant to work outcomes; accommodation for physical differences may not be as simple as it might seem. Remodeling or computerizing a work area might be prohibitively expensive; providing work aids for some people might create hazards for others. Job analysis should show just how important apparent physical requirements really are and how the job might be done differently, and it should form a foundation for imaginative thinking about potential methods of accommodation. Hogan and Quigley (1986) reported that height and weight requirements had been approved in litigation only where there was no adverse impact or where job relatedness was clearly demonstrated.

Physical Abilities. Many jobs, not merely laboring jobs, require physical skills. Package deliverers, firefighters, power line repairers, tree trimmers, construction workers, and paramedics are among those for whom strength, endurance, and balance are relevant. Nevertheless, few psychologists have studied physical abilities and their relevance in employment practices, although recently Deborah Gebhardt has written extensively on physical ability assessment (Gebhardt & Baker, 2010a, 2010b).

Most of what we know has come from the work of Edwin A. Fleishman and his associates. Joyce Hogan (1991a, 1991b) considered seven of these sufficient in personnel selection, arguing that static and dynamic distinctions rarely made sense in job descriptions. Condensing further, she identified three general fitness factors: (1) muscular strength, the ability to apply or resist force by contracting muscles; (2) cardiovascular endurance, or aerobic capacity, and (3) coordination, or quality of movement. A combination of the ideas of Hogan and Fleishman's contributes to an overall model of physical abilities as offered in Figure 3.3.

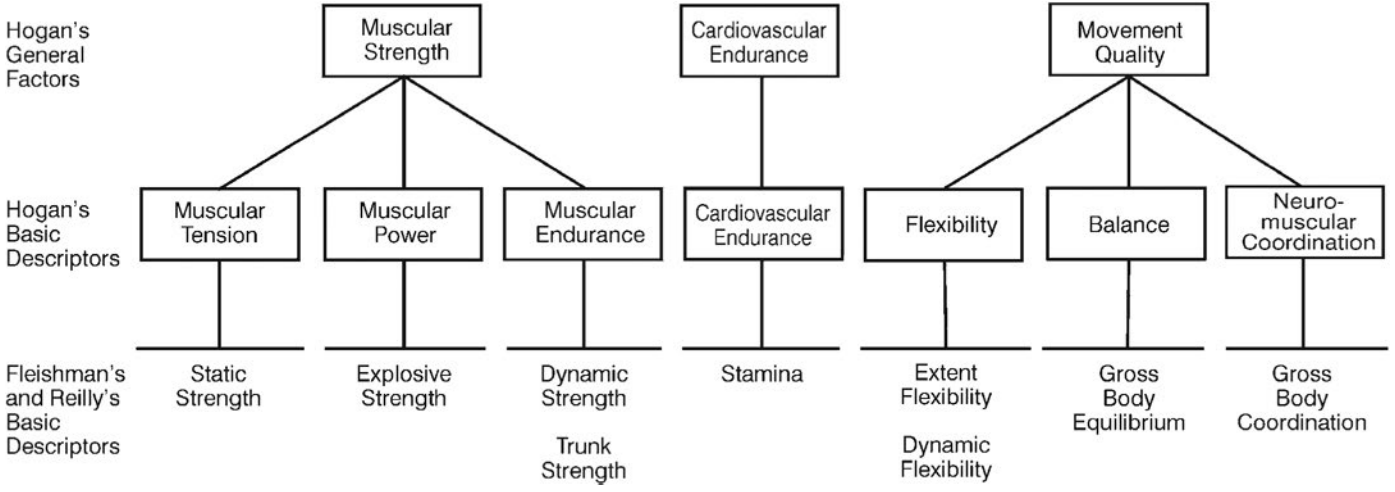

FIGURE 3.3 A model of physical abilities, using the J. Hogan (1991a, 1991b) and Fleishman and Reilly (1992) descriptors.

How general should definitions of physical abilities be? Again, it depends on the generality or narrowness of the criterion to be predicted.

Sensory Abilities. Vision and hearing ability are not unitary; "good vision" or "good hearing" means quite different visual or auditory skills for different jobs. Fleishman and Reilly (1992) listed 12 different visual and auditory abilities, including with near and far acuity such specialized abilities as night vision, color vision, depth perception, and a corresponding variety of sounds related to hearing. Postwar vision research at Purdue University, and others, was reviewed in Guion (1965). No such extensive research has been done on hearing. It is likely that, in both cases, strong cognitive components are involved as well as the sensory ones. A certain pitch with low volume might be emitted from a piece of machinery; two people may have the acuity to hear it, but the better worker is the one who understands its implication.

Emotional Intelligence

The concept of *emotional intelligence* has received considerable attention in the popular press (Goleman, 1995) and is receiving serious attention as a psychological construct (Matthews, Zeidner, & Roberts, 2002). Although there appears to be no universally agreed-upon definition of emotional intelligence, it is generally described as the ability to accurately perceive, appraise, express, and regulate emotions (Mayer & Salovey, 1997). Certainly, such an ability would seem to be useful for a wide range of occupational arenas, including customer service, sales, politics, and many others. In a comprehensive and critical review of the literature, however, Matthews, Roberts, and Zeidner (2004) concluded that the construct lacked conceptual coherence, and that attempts to measure emotional intelligence had fallen woefully short of standard psychometric criteria. At this point, we simply do not know if emotional intelligence is distinct from personality, or that it correlates with job performance when cognitive ability and personality are controlled. For example, a meta-analysis by Joseph, Jin, Newman, and O'Boyle (2015) concluded that, after controlling for personality and ability, the correlation between emotional intelligence and job performance is close to zero. Thus, emotional intelligence measures work because they assess personality and cognitive ability, as well as emotional intelligence. However, from the point of view of the practitioner, one could argue that if the label "emotional intelligence" is attractive to consumers and if the measures do predict, then measures of emotional intelligence are useful for selection.

Experience, Education, and Training

Some predictors are hypothesized without clearly articulated constructs; specified training or experience requirements are among them. Credential requirements are rarely useful; too often people with fine credentials do not have the competencies to match (Ash, Johnson, Levine, & McDaniel, 1989). They *can* be useful, if

systematically evaluated and based on job analysis, (Howard, 1986; McDaniel, Schmidt, & Hunter, 1988). Ash et al. (1989) suggested that education requirements, including a specific college major, might be justified if (a) the job requires extended knowledge comparable to that of recognized professions, (b) the knowledge and ability requirements are hard to evaluate by other methods, (c) the consequences of *not* requiring the degree and major are likely to be severe, and (d) the degree program is the only way to acquire the knowledge demanded by the job. Even in these cases, however, it may be better to identify the competencies sought and to distinguish *preferred* from *necessary* qualifications.

In addition to being used as predictors, education and experience often serve as minimum qualifications. That is, in order to be a minimally qualified applicant for a job, I must be able to document that I have a certain level of education and experience. For example, an HR manager job may require a four-year college business degree plus three years of progressive HR experience. When used as minimum qualifications, education and experience still constitute selection tests and should be validated. Buster, Roth, and Bobko (2005) described a process for the content validation of education- and experience-based minimum qualifications that won approval from a federal court.

Predictors for Team Selection

Individual candidates for assignment to a team must be assessed, and we stress here the assessment of *individual* candidates for team assignments. Technical competence is surely among them, but the skills, knowledge, and motivation needed to function well in a team go beyond the core technical skills. Stevens and Campion (1994) identified a generic list of teamwork KSA requirements. Their framework was based on two primary categories of KSAs that they termed *interpersonal* (i.e., conflict resolution, collaborative problem solving, communication) and *self-management* (i.e., goal setting and performance management, planning and task coordination). Stevens and Campion (1999) used this framework to develop a selection test for teamwork settings, appropriately dubbed the "Teamwork Test." Although preliminary research on the Teamwork Test has been promising (McClough & Rogelberg, 2003; Stevens & Campion, 1999), its success may be based on its large general mental ability component. Personality approaches to team member selection remain largely a matter of conjecture (cf., Kichuk & Wiesner, 1998).

An alternative approach to team member selection is based on the idea of achieving optimal "fit" between the applicant's preferred teamwork style and that of the employing team. Anderson and Burch's (2003; Burch & Anderson, 2004) Team Selection Inventory extends the notion of person–job fit to the domain of teamwork. Thus, person–team fit is assessed by first assessing the work team's emphasis on safety, innovation, goals, and quality, and then assessing the relative importance the applicant places on these dimensions. The usefulness of this approach for predicting team member performance remains to be seen.

Discussion Topics

1. Develop a predictive hypothesis, on both a conceptual and an operational level, using one of the five factors of personality.
2. How does contextual performance differ from traditional task performance? Would you like to be evaluated for your contextual performance in the classroom?
3. What kinds of factors would predict who would perform well as a team member in a group project? Consider group projects that have been assigned in your classes.

References

Anderson, N., & Burch, G. St. J. (2003). *The Team Selection Inventory.* Windsor, England: ASE/NFER-Nelson.

Ash, R. A., Johnson, J. C., Levine, E. L., & McDaniel, M. A. (1989). Job applicant training and work experience evaluation in personnel selection. *Research in Personnel and Human Resource Management, 7,* 183–226.

Ashton, M. C., Lee, K., & & de Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality Factors: A Review of Research and Theory. *Personality and Social Psychology Review, 18,* 139–152.

Barrett, G. V., Alexander, R. A., & Doverspike, D. (1992). The implications for personnel selection of apparent declines in predictive validities over time: A critique of Hulin, Henry, and Noon. *Personnel Psychology, 45*(3), 601–617.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1985). The concept of dynamic criteria: A critical reanalysis. *Personnel Psychology, 38,* 41–56.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9,* 9–31.

Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90,* 1185–1203.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478–494.

Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an Army officer sample. *Organizational Behavior and Human Decision Processes, 40,* 307–322.

Borman, W. C., & Motowidlo, S. M. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel Selection in Organizations* (pp. 71–98). San Francisco, CA: Jossey-Bass.

Brief, A. P., & Motowidlo, S. J. (1986). Prosocial organizational behaviors. *Academy of Management Review, 11,* 710–725.

Burch, G. St. J., & Anderson, N. (2004). Measuring person-team fit: Development and validation of the team selection inventory. *Journal of Managerial Psychology, 21,* 406–426.

Buster, M. A., Roth, P. L., & Bobko, P. (2005). A process for content validation of education and experience-based minimum qualifications: An approach resulting in federal court approval. *Personnel Psychology, 58,* 771–799.

Campbell, J. P. (2012). Behavior, performance, and effectiveness—in the 21st century. In S. W. J. Kozlowski (Ed.), *The Oxford Handbook of Organizational Psychology* (pp. 159–194). Oxford, England: Oxford University Press.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel Selection in Organizations* (pp. 35–70). San Francisco, CA: Jossey-Bass.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, England: Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1–22.

Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90,* 1241–1255.

English, H. B., & English, A. C. (1958). *A comprehensive dictionary of psychological and psychoanalytic terms.* New York, NY: Longmans, Green, and Co.

Farrell, J. N., & McDaniel, M. A. (2001). The stability of validity coefficients over time: Ackerman's (1988) model and the General Aptitude Test Battery. *Journal of Applied Psychology, 86,* 60–79.

Fleishman, E. A., & Reilly, M. E. (1992). *Handbook of human abilities: Definitions, measurements and job task requirements.* Palo Alto, CA: Consulting Psychologists Press.

Funder, D. C. (1991). Global traits: A neo-Allportian approach to personality. *Psychological Science, 2,* 31–39.

Gebhardt, D. L., & Baker, T. A. (2010a). Physical performance. In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment: Evidence-based practices for selecting and developing organizational talent* (pp. 165–196). San Francisco, CA: Jossey-Bass.

Gebhardt, D. L. & Baker, T. A. (2010b). Physical performance tests. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 277–298). New York, NY: Routledge.

Ghiselli, E. E. (1956). Dimensional problems of criteria. *Journal of Applied Psychology, 40,* 1–4.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34.

Goldberg, L. R., Grenier, J. R., Guion, R. M., Sechrest, L. B., & Wing, H. (1991). *Questionnaires used in the prediction of trustworthiness in pre-employment selection decisions: An APA task force report.* Washington, DC: American Psychological Association.

Goleman, D. (1995). *Emotional intelligence.* New York, NY: Bantam Books.

Gough, H. G. (1985). A work orientation scale for the California Psychological Inventory. *Journal of Applied Psychology, 70,* 505–513.

Guion, R. M. (1965). *Personnel testing.* New York, NY: McGraw-Hill.

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18,* 135–164.

Helmreich, R. L., Sawin, L. L., & Carsrud, A. L. (1986). The honeymoon effect in job performance: Temporal increases in the predictive power of achievement motivation. *Journal of Applied Psychology, 71,* 185–188.

Hofstee, W. K. B., de Raad, B., & Goldberg, L. R. (1992). Integration of the big five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology, 63,* 146–163.

Hogan, J. (1991a). Physical abilities. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 753–831). Palo Alto, CA: Consulting Psychologists Press.

Hogan, J. (1991b). Structure of physical performance in occupational tasks. *Journal of Applied Psychology, 76,* 495–507.

Hogan, J., Hogan, R., & Busch, C. M. (1984). How to measure service orientation. *Journal of Applied Psychology, 69,* 167–173.

Hogan, J., & Quigley, A. M. (1986). Physical standards for employment and the courts. *American Psychologist, 41,* 1193–1217.

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory: Manual.* Tulsa, OK: Hogan Assessment Systems.

Howard, A. (1986). College experiences and managerial performance. *Journal of Applied Psychology, 71,* 530–555.

Hulin, C. L., Henry, R. A., & Noon, S. L. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relationships. *Psychological Bulletin, 107,* 328–340.

Humphreys, L. G. (1979). The construct of general intelligence. *Intelligence, 3,* 105–120.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85,* 869–879.

Joseph, D., Jin, J., Newman, D., & O'Boyle, E. H. (2015). Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. *Journal of Applied Psychology, 100,* 298–342.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74,* 657–690.

Kichuk, S. L., & Wiesner, W. H. (1998). Work teams: Selecting members for optimal performance. *Canadian Psychology, 39,* 23–32.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86,* 148–161.

Matthews, G., Roberts, R. D., & Zeidner, M. (2004). Seven myths about emotional intelligence. *Psychological Inquiry, 15*(3), 179–196.

Matthews, G., Zeidner, M., & Roberts, R. D. (2002). *Emotional intelligence: Science and myth.* Cambridge, MA: MIT Press.

Mayer, J. D., & Salovey, P. (1997). *What is emotional intelligence?* New York, NY: Basic Books.

McClough, A. C., & Rogelberg, S. G. (2003). Selection in teams: An exploration of the teamwork knowledge, skills, and ability test. *International Journal of Selection & Assessment, 11,* 56–66.

McCrae, R. R. (1992). The five-factor model: Issues and applications [Special issue]. *Journal of Personality, 60,* 175–532.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of the validity of methods for rating training and experience in personnel selection. *Personnel Psychology, 41,* 283–314.

Mischel, W. (1968). *Personality and assessment.* New York, NY: Wiley.

Murphy, K. R. (1989). Is the relationship between cognitive ability and job performance stable over time? *Human Performance, 2,* 183–200.

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA, England: Lexington Books/D.C. Heath and Com.

Organ, D. W., Podsakoff, P. M., & MacKenzie, S. B. (2006). *Organizational citizenship behavior: Its nature, antecedents, and consequences.* Thousand Oaks, CA: SAGE Publications.

Sackett, P. R., Berry, C. M., Wiemann, S. A., & Laczo, R. M. (2006). Citizenship and counterproductive behavior: Clarifying relations between the two domains. *Human Performance, 19,* 441–464.

Sackett, P. R., & Walmsley, P. T. (2014). Which personality attributes are most important in the workplace? *Perspectives on Psychological Science, 9,* 538–551.

Smith, F. J. (1977). Work attitudes as predictors of attendance on a specific day. *Journal of Applied Psychology, 62,* 16–19.

Society for Industrial and Organizational Psychology. (1987). *Principles for the validation and use of personnel selection procedures.* (3rd ed.). College Park, MD: Author.

Spearman, C. (1927). *The abilities of man.* New York, NY: Macmillan.

Spence, J. T., Helmreich, R. L., & Pred, R. S. (1987). Impatience versus achievement strivings in the Type A pattern: Differential effects on students' health and academic achievement. *Journal of Applied Psychology, 72,* 522–528.

Stevens, M. J., & Campion, M. A. (1994). The knowledge, skill, and ability requirements for teamwork: Implications for human resource management. *Journal of Management, 20,* 503–530.

Stevens, M. J., & Campion, M. A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management, 25,* 207–228.

Thoresen, C. J., Bradley, J. C., Bliese, P. D., & Thoresen, J. D. (2004). The big five personality traits and individual job performance grow trajectories in maintenance and transitional job stages. *Journal of Applied Psychology, 89,* 835–853.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs,* (1).

Whyte, W. H., Jr. (1957). *The organization man.* New York, NY: Doubleday.

4

KNOWING WHAT'S LEGAL (AND WHAT'S NOT)[1]

Title VII, *Uniform Guidelines*, Affirmative Action, and Negligent Hiring

In this age of litigation, personnel decisions based on whim, stereotypes, prejudices, or expediency are just plain foolish. This chapter emphasizes U.S. laws promoting equal employment opportunity (EEO), particularly the Civil Rights Act of 1964. That legislation has dominated employment practices in the United States for over 50 years and is the foundation for other antidiscrimination laws. Its importance is widespread; it has influenced legislation in other countries, and U.S. EEO laws apply anywhere in the world where United States citizens are employed by an U.S.-controlled company. Businesses incorporated in other countries are subject to these laws for their operations in the United States.

The Civil Rights Act of 1964

The Civil Rights Act (1964) was to social policy in the United States what the continental divide is to the flow of rivers. It put the full power of the federal government to work on behalf of African American citizens having equal access to schools and public accommodations as well as employment opportunities. During Congressional debate, a Civil Rights *opponent* offered an amendment to include sex as a proscribed basis for decision. Some say this was to derail the bill, others suggest that it was done to protect the rights of White women (see Highhouse & Gutman, 2011). Nevertheless, the Act has also provided women with a significant source of relief from unequal opportunity. The Act's importance as a signal of a shifting concept of government cannot be overemphasized. Previously, the federal government had regulated things and standards (e.g., food and drugs, weights and measures). This Act regulated behavior.

For those too young to remember, which now is almost all readers of this book, common practices with regard to employment were very different prior to the

Civil Rights Act of 1964. If you picked up the classified section of the newspaper back in the early half of the 1960s, you would find jobs listed by race and sex; there would be a column of jobs for Black males, for Black females, for White males, and for White females. The segregation of jobs by race and sex was common and was accompanied by separate compensation systems and promotional ladders. The Civil Rights Act of 1964 was an attempt to remedy the effects of this segregation, specifically the resulting income inequality.

Although the Civil Rights Act was very broad, dealing with a variety of discrimination issues, *Title VII* of the Act was directed explicitly to employment issues.

Unlawful Employment Practices

Title VII of the Civil Rights Act specified several unlawful employment practices:

1. Employers may not fail or refuse to hire, or discharge, anyone on the basis of race, color, religion, sex, or national origin.
2. They may not segregate or classify employees or applicants so as to deprive anyone of employment opportunities on the basis of race, color, religion, sex, or national origin.
3. Employment agencies may not fail or refuse to refer candidates on the basis of any of these characteristics. This holds as well for labor unions with regard to membership or influencing employers to discriminate.
4. All provisions apply equally to employers, labor organizations, or joint labor–management committees controlling training programs.
5. Advertising employment or training opportunities may not indicate preferences for any group under any of these designated characteristics. Separate classified columns for "Help Wanted—Men" and "Help Wanted—Women" were discontinued as were statements of preferences for characteristics that only men, or Whites, or speakers of English are likely to have.
6. It is unlawful to retaliate against people who have opposed unlawful employment practices under the Act.

Exemptions. The Act did not "apply to an employer with respect to the employment of aliens outside any State," nor did it prevent religious organizations from hiring their own adherents to carry out religious work, although the issue of religious discrimination has always been a complicated one. Some preferential hiring

was endorsed explicitly such as preferential hiring of American Indians on or near reservations, or veterans' preference. Bona fide seniority systems were also protected. The Act does *not* prohibit or discourage discrimination on the basis of actual qualifications to do a job.

Students and readers always seem most interested in two types of cases that are infrequent but nevertheless seem to capture people's attention. The first involves Bona Fide Occupational Qualifications (BFOQs). A BFOQ might be a male attendant for a men's restroom or advertising for a female actress to play Madonna in a movie. In truth, the BFOQ defense and associated court cases are rare, as the number of situations where a BFOQ might apply is extremely limited. It should be noted that the BFOQ defense is not applicable where the rationale is that a customer preference exists. The BFOQ defense has been used by Hooters of America in limiting its hiring of Hooters servers to female applicants (see, for example, *Latuga v. Hooters Inc.*, 1996). Another famous court case involving restaurants and servers is *EEOC v. Joe's Stone Crab* (2001).

The second involves religious discrimination. As an example of the complexity of this issue, in *Hosanna-Tabor Evangelical Lutheran Church and School v. EEOC* (2012), the Supreme Court applied the principle of the ministerial exemption to the firing of a teacher with narcolepsy from a school where her religious duties were limited to approximately 45 minutes a day and most of her responsibilities involved teaching nonreligious subjects. The ministerial exemption states that ministers cannot sue their churches claiming they had been fired in violation of employment discrimination laws.

The Civil Rights Act of 1991

Differences in opinions about fairness in employment were neither resolved nor clarified by 25 years of EEO enforcement and litigation. If anything, they froze as polar opposites, held not as reasoned policy but as deeply held emotional commitments. For some, Supreme Court decisions seemed overdue statements of sanity in the EEO arena. To many others, they seemed to signal a weakening of basic EEO principles, including the Court's standards in *Griggs v. Duke Power Co.* (1971).

The Civil Rights Act of 1991 clarified the results of a number of court decisions and their impact on Title VII of the Civil Rights Act of 1964. In particular, many legal authorities and testing experts point to the Supreme Courts' decision in *Wards Cove Packing Co. v. Atonio* (1989) as a major case leading to the passage of

the Civil Rights Act of 1991. In *Wards Cove,* a five-justice majority affirmed most of the plurality decision in *Watson v. Fort Worth Bank & Trust* (1988). It affirmed the extension of disparate impact analysis to subjective procedures (e.g., the job interview), the need to specify the practice being challenged, and maintaining the burden of persuasion on the plaintiff. Prior to *Watson,* the burden of persuasion shifted to the defendant, after the plaintiff had met the burden of production, although this notion of a shifting burden has always been the subject of debate in both the legal and selection literature.

The most controversial aspects of the *Wards Cove* decision involved the further requirement that evidence of disparate impact compare the demographic data on a specific job to the available supply of people for that job; that is, disparate impact statistics must be based on the relevant applicant pool. In addition, the Court also reduced the "business necessity" language to "business justification," saying that a practice need not be essential to survival of the business or in some other sense indispensable.

Leading up to the Civil Rights Act of 1991, there was also growing concern over the use of quotas or "race norming." (Race norming is a way to "get the numbers right" by using percentiles or standard scores in different score distributions for different subgroups and using top-down selection based on the percentiles; we discuss this issue in more detail in Chapter 9). Shortly before the Congressional debates, controversy erupted over the practice in state Employment Services referrals that used the United States Department of Labor's *General Aptitude Test Battery* (GATB; see Hartigan & Wigdor, 1989 and Chapter 9). Race norming does not seriously affect mean job performance, but making it illegal quieted the charges that the Civil Rights Act of 1991 was a quota bill. Section 106 of the Civil Rights Act of 1991 places severe limits on the use of such adjustments, as it states that

> It shall be an unlawful employment practice for a respondent, in connection with the selection or referral of applicants or candidates for employment or promotion, to adjust the scores of, use different cutoff scores for, or otherwise alter the results of, employment related tests on the basis of race, color, religion, sex, or national origin.

Of the Supreme Court decisions opposed in the 1990 bill, only the *Watson* view that the burden of persuasion remained with the plaintiff was changed by the 1991 Act. Definitions of business necessity, job relatedness, and even the concept of disparate impact appeared to have been codified by the 1991 Act. Nevertheless, they remain as ambiguous (some say "flexible") as before, and common sense definitions may yet prevail. Another provision addressed intentional discrimination, providing for jury trials and for compensatory and punitive damages. Good sense, if not morality, requires organizations to make sure that intentional discrimination on irrelevant grounds, or even the appearance of it, did not occur.

The Equal Employment Opportunity Commission

The Civil Rights Act of 1964 established the Equal Employment Opportunity Commission (EEOC), empowered to investigate charges of prohibited employment practices; to dismiss charges deemed unfounded; to use conference, conciliation, and persuasion to eliminate practices where charges were found to be true; and to work with authorities in states or other jurisdictions where the practices are prohibited by local law. Where there is a finding of "reasonable cause" to believe the charge is true, the EEOC can file suit in the federal courts. Early in EEO history, working with employers through "gentle persuasion" lost out procedurally to the adversarial posturing of litigants.

The EEOC is one of three federal agencies that play a prominent role in the administration of employment discrimination laws. The second agency is the Office of Federal Contract Compliance Programs (OFCCP), which is responsible for the enforcement of legislation and Executive Orders with organizations receiving federal money through government contracts. The third agency is the Department of Justice, which often intervenes in cases involving public jurisdictions, most commonly cases involving hiring and promotion in public safety positions such as fire and police. In addition, discrimination cases may be brought by state fair employment commissions or by private individuals.

Uniform Guidelines

The ***Uniform Guidelines on Employee Selection Procedures,*** identified hereinafter simply as *Uniform Guidelines* (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, 1978, 1979, 1980) help guide employee selection aspects of the Act and Executive Orders. The course of the *Uniform Guidelines* development was not smooth (see Guion, 1998, for a history) but they, along with case law and the provisions of the Civil Rights Act of 1991, now define "the legal context" for personnel decisions.

Disparate Impact and Disparate Treatment

Discrimination may be charged and litigated under two distinct legal theories. One is ***disparate impact*** (sometimes called adverse impact; in Chapter 9 we discuss the definition of disparate impact, various approaches to quantifying disparate impact, and possible approaches to reducing disparate impact) in which an action is said to affect different groups differentially. Although the purpose of the law and its enforcement is to protect individual citizens from discrimination based on group identity, disparate impact refers to a differential and unintended effect on protected groups. Evidence that a group as a whole is less likely to be hired is preliminary (prima facie) evidence of discrimination against protected group members (but no more than that).

The second theory is ***disparate treatment***, where there is evidence that a candidate from a protected group intentionally was treated differently from other candidates in the employment process. In principle, all applicants should receive the same treatment—the same kinds of interviews, tests, application forms, and hiring standards.

The *Uniform Guidelines* note that disparate impact requires justification in terms of business necessity. That term does not imply something necessary for the survival of the business; rather, ***business necessity*** means that a selection procedure must be related to job behavior or performance—usually that it is a valid predictor of an important criterion—and, therefore, serves a useful business purpose not as well served by a known alternative with less disparate impact.

The 80% (Four-Fifths) Rule

The 80% (four-fifths) rule is used to determine the existence of disparate impact (in Chapter 9 we discuss in detail the identification or analysis of disparate impact and the definition of protected groups). Disparate impact exists if the selection ratio in one group (presumably the minority group) is less than 80% of the selection ratio in the other. Consider a situation in which a company has 80 White applicants for a job and 20 Black applicants. The 80% rule says that if the company hires 25% of the White applicants (in this case, 20), it is reasonable to expect that 20% (i.e., four-fifths of 25%) of the Black applicants would be hired (in this case, 4). Originally, the 80% rule was intended to be an enforcement trigger used by federal agencies to move to an investigation. It was viewed as lacking the force of law and requiring interpretation in light of other information. One employer might have a disparate impact ratio well under 80%, and, therefore, be suspected of potential discrimination, only because vigorous affirmative action had increased the number of nonqualified minority applicants. Another employer may have a disparate impact ratio above 80% because of the chilling effect of a reputation suggesting that application to that employer would be futile for members of certain demographic groups. Although a "chilling effect" argument requires substantial proof to succeed in court (e.g., *International Brotherhood of Teamsters v. United States,* 1977), its inclusion in the *Uniform Guidelines* emphasized that the four-fifths rule was subject to interpretation in specific contexts. However, both case law and the Civil Rights Act of 1991 have placed the analysis of disparate impact and the four-fifths rule in a much more central and critical role (see Chapter 9). The 80% rule is often interpreted as a measure of practical significance rather than as an enforcement trigger.

Options Under Disparate Impact. A selection procedure having disparate impact on any protected group may be modified, eliminated, or justified by validation. Modification is an option to be undertaken only with carefully designed research. Elimination is not an acceptable option for procedures with useful levels of validity. To abandon the use of a valid selection procedure because of fear of litigation is to return to essentially random selection—not a wise way to run a business. Of the three, validation in support of a business necessity defense is the only organizationally sound option. Organizational leaders should require it without waiting for federal agencies to do so. A selection procedure should be replaced if validation fails to show that it serves an important business purpose. Statistical validation is not the only way to show job relatedness. Case law gives several examples where validity was accepted on the basis of content strategies and there is no preference for any specific approach to validation. In addition, in some cases, the business necessity defense has been made based on rational, "common sense" grounds, as opposed to through validation or job relatedness.

A further option is to substitute an alternative selection procedure with less disparate impact (*Johnson et al. v. City of Memphis, 2014*). Actually, alternatives should be considered from the outset. In planning the research, a predictive hypothesis should be formed with certain applicant traits hypothesized as predictor constructs (see Chapter 3). Then one must choose one or more operational definitions of the hypothesized constructs to use in the study. In making that choice, prior literature should be considered, including literature describing evidence of validity and evidence of disparate impact, and evidence of other challenges to the validity of the methods (See Chapter 9 for a discussion of disparate impact reduction techniques).

Requirements for Validation

New research evidence, and the thinking stimulated largely by EEO regulations, has placed the *Uniform Guidelines* somewhat at odds with contemporary professional views. The *Uniform Guidelines* have treated criterion-related, content, and construct validity as representing three different roads to validity. Actually, the idea of three distinctly different *kinds* of validity has been inconsistent with professional views at least since the terms were introduced in the *1954 Technical Recommendations for Psychological Tests and Diagnostic Techniques* (American Psychological Association [APA], American Educational Research Association [AERA], & National Council on Measurement in Education[NCME], 1954) and explicitly denied in the *1999 Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, see also the *2014 Standards*). The more integrated view, where these terms refer to different *aspects* or varieties of evidence of validity, is discussed more thoroughly in Chapter 5.

Criterion-Related Validation. Obtaining a statistical relation between a predictor and criterion may not always be technically feasible. The *Uniform Guidelines*

explicitly recognize three conditions in the definition of technical feasibility: (1) adequate samples, (2) adequate score ranges on both predictor and criterion, and (3) an acceptable criterion (i.e., unbiased, reliable, and relevant). If criterion-related validation is determined to be feasible, the *Uniform Guidelines* specify the following:

1. Job information must be reviewed, or the job must be analyzed, to determine criteria; a criterion must "represent critical or important job duties, work behaviors or work outcomes. . . ." (EEOC et al., 1978, p. 38300).
2. Samples used should be like the relevant labor market for the job in question.
3. Relations between predictors and criteria should be statistically expressed and should be statistically significant, typically at the 5% level of confidence.
4. If, in general, the results show that a selection procedure is a valid predictor of an important criterion, studies of fairness should be conducted (where technically feasible). The *Uniform Guidelines* are ambiguous about the definition of fairness to be used.

Content Validation. To establish content validity, the *Uniform Guidelines* require job analysis to identify work behaviors required for effective performance, their relative importance, and the work products expected to result. The analysis should focus on observable task behavior, although questions about what is observable are sometimes raised (e.g., is *planning or leadership* an observable behavior?). The apparent narrowness is relieved somewhat by the inclusion of tests developed previously by others, in other circumstances, if the test content matches in some convincing way the content of the job as revealed by the job analysis; although this is probably better handled using the provisions of the *Uniform Guidelines* related to the transportability or transfer of validity.

> Relying on ***content validity*** evidence requires the use of a task-focused job analysis, rather than one that (only) focuses on the identification of KSAs.

Our interpretation of the *Uniform Guidelines* suggests the following:

1. A content domain must be defined on the basis of a thorough job analysis, one that not only identifies tasks and resources used but also determines their relative importance to the job overall. Implicit in this statement is the assumption that acceptable content validity arguments are job specific.
2. If the defined job content is but a portion of the job, it must be critical to overall job performance.
3. The content of selection procedures defended on the basis of content validity must match the content defined by the job analysis.

4. Required prior training or experience may be justified as valid content if its content closely resembles the content of the job, as identified by job analysis.

Construct Validation. The *Uniform Guidelines* also require job analysis as the first step in a defense invoking construct validity. It should identify behavior required for effective performance and constructs believed to underlie effective behavior. Such constructs should be clearly named, defined, and distinguished from other constructs, and selection procedures chosen should be supported with empirical evidence that they are related to the intended constructs.

Unfortunately, the *Uniform Guidelines* go on to say, "The relationship between the construct as measured by the selection procedure and the related work behavior(s) should be supported by empirical evidence from one or more criterion-related studies involving the job or jobs in question which satisfy the provisions [for criterion-related validation]" (EEOC et al., 1978, p. 38303). In short, despite some words supporting the use of construct validity arguments, this provision effectively rules out construct validity arguments, and we know of very few instances in which a test was defended on the basis of construct validity. Under discussions of descriptive and psychometric validity in Chapter 5, we give serious consideration to the nature of construct validity and of its role in establishing (or failing to establish) the job relatedness of a selection procedure.

Use of Valid Personnel Selection Procedures

"Transportability" of Validity Information. Do I always have to do a local validation study, whether content or criterion related? Perhaps surprisingly, the answer is no; the *Uniform Guidelines* allow for another type of validity referred to as transportability. Acceptable evidence of validity may be based on validation research done elsewhere, but only with severe restrictions. The question is whether the outside research generalizes to the user's situation; some people refer to such generalizing as "transporting" the validity evidence. This provision predates the development of validity generalization research, and discussing it now seems rather quaint. However, it is still a part of the legal context, and there are often situations where one wishes to generalize from a specific study rather than from a body of related studies. Many consulting firms appear to rely heavily on transportability or transfer studies as evidence of the validity of their selection procedures.

A generalized requirement, regardless of the nature of the validity evidence, is that the documentation and reporting be available in a form "similar to" the form required by the *Uniform Guidelines*. Other requirements for transporting criterion-related validity studies include the following:

1. There must be evidence of the similarity of the job at hand and the job in the original study, identified by the same methods of job analysis.
2. The criterion in the original study is relevant to the local job.

3. The demographic characteristics of the applicant pool or research sample in the original study must be similar to those in the new situation.

Certainly the key characteristics of the job—those for which criterion data will be sought—should match in the two situations. It is less certain that broader similarity is truly necessary, and very nearly certain (from research done in the 1970s) that demographic similarity is not necessary; nevertheless, these requirements still define part of the legal context in which personnel decisions are made.

Testing for Higher Level Jobs. If you watch old movies, you will find many plot lines where an employee starts in the mailroom and then moves up to the executive suite. Frequently, banks hire applicants into teller jobs with the hope that they can be promoted into managerial or vice president positions. So can an organization consider the potential for promotion during entry-level hiring? The basic principle underlying the *Uniform Guidelines* is that hiring should be for jobs, not for organizations or for advancement. However, employers frequently want to hire people who will advance in the organization. The *Uniform Guidelines* recognize this and further recognize that, for many jobs, advancement to higher levels is rare; in such jobs, hiring for the higher level may, in effect, be a pretext for discrimination. Employers are permitted to assess applicants for the higher levels only if (a) the majority of those still employed after "a reasonable period of time" (rarely more than five years) progress to the higher level job, (b) the higher level job will continue to require largely the same skills during that time, and (c) the original job is not likely to provide the development of the requisite knowledge or skill.

A related question is whether organizations can use selection to improve their workforce. Imagine the Director for Human Resources for a large city would like to improve the quality of its firefighters by requiring a two-year college degree in fire science. In the past, the city has required only a high school degree for hiring firefighters. Can the city do this? The answer is yes. As a general principle, an employer is free to improve the workforce, and the decision to change requirements is not an admission that previous procedures were inadequate. However, the new requirement will be judged on its validity and its susceptibility to disparate impact.

Use of Scores. Four methods of score use are recognized: ranking, banding, pass/fail with a cut score, and combination with other tests. Where there are differences in mean scores of demographic subgroups (and there usually are), and where variances are about the same, there is a necessary relationship between the level of a cutting score and the degree of disparate impact. That is, a cut score can be set high enough that virtually no one in the lower scoring group can pass it. The way to reduce disparate impact, therefore, is to lower the cutting score. How low? The *Uniform Guidelines* themselves do not say, but some enforcement agencies in some situations have argued that the cutoff should permit hiring people at

approximately the same score of the lowest scoring employee who is retained and on that basis presumed satisfactory. This position ignores the statistical realities of linear or other monotonic relationships between test scores and performance. It ignores Congressional intent in supporting, in Title VII, employers' rights to set qualifications. It ignores the fact that selection procedures are typically adopted for the sake of improving levels of proficiency in the workforce, not simply for maintaining what may be an unacceptable status quo.

In a different vein, it ignores the realities of selection procedures. In civil service jurisdictions, the typical pattern is to establish an eligibility list giving the names of all candidates who have exceeded a low cut score; selection is then done by ranking (top-down) those on that list until the list is "exhausted." An exhausted list typically still has names on it of people who have passed the test, but the passage of time and difficulties in finding people still interested in the job induce the authorities to initiate a new examination and start over with a new eligibility list. Even though a passing score is established, actual practice makes the de facto passing score somewhat higher.

In the private sector, the difference between a minimum cut score, if one is even considered, and the de facto score is even more pronounced. Hiring rates differ with the times. In a period of recession, for example, a company may do little hiring, and it will choose from the best of the many applicants presenting themselves for consideration; the lowest score among those hired may be quite high. When unemployment is very low, when virtually "any warm body" will do rather than leave a job unfilled, the de facto cutting score is reduced drastically. Such variability seems to be unacceptable to the authors of the *Uniform Guidelines*; they seem to assume that, unless ranking is justified, a fixed cut point will be established. Nothing is said about selection above that point. If more people score above the cut score than can be hired, how should new employees be chosen? At random? The *Uniform Guidelines* do not say.

Reporting and Record-Keeping Requirements

The *Uniform Guidelines* specify stringent record-keeping requirements, and most companies have to meet even more detailed reporting rules as specified by either the EEOC or OFCCP. Although not matters of psychological or psychometric principle, these are important in HR management. They are so important to litigation that any employer affected by the *Uniform Guidelines* should study them in great detail and with informed legal counsel.

Case Law From Selected EEO Court Decisions

A *statute,* such as the Civil Rights Act, is a set of words adopted after legislative debate, compromise, and amendment. Application of these words to a specific instance is not always clear. Each party in a dispute may honestly believe the words

to be on its side. The courts have the responsibility of applying the words and their legislative history to the specific case. In the United States federal courts, the dispute is first heard by the judge or jury in a district court; the judge or jury is the "trier of fact" who determines the facts of the case and interprets them in the light of the relevant statutes and prior court decisions. Attorneys' arguments, testimony from witnesses, and study of the law and interpretation developed in prior cases all contribute to the judge's decision.

When a jury is involved, the judge instructs it as to the law. In the end, one party prevails; the losing party may appeal the decision to a circuit court of appeals (the appellate level), which has jurisdiction over district courts in its geographical area. At the appellate level, lawyers present their cases to a panel of judges; these judges do not hear witnesses or determine facts but hear and study arguments to determine whether an error of procedure or of legal interpretation has occurred.

The decision of the lower court may be confirmed, reversed, or remanded for reconsideration or retrial. Decisions at the appellate level become binding precedents for the district courts of that circuit; that is, those decisions guide district court judges in future cases involving the same or similar legal issues. A district judge does not always follow precedent, but strong and compelling reasons, based on the facts of the case and their differences from the facts in the precedent case, are needed to justify deviation. The highest level of appeal is to the United States Supreme Court. Decisions at this highest level are binding precedents for all other federal courts—with the same possibility that the triers of fact in a new case may find important differences justifying a different legal path.

At all three levels, decisions rendered become part of ***case law***—the body of judicial interpretations of the statute. The relative weight of decisions in case law is greater at the higher judicial levels, so we concentrate on a few decisions rendered by the Supreme Court and some recent ones from lower courts. We do not give details of cases but will give implications for personnel practices (See Gutman, Koppes, & Vodanovich, 2011 for more on case law). As mentioned previously, we are not lawyers nor do we pretend to be qualified to give legal opinions.

Griggs v. Duke Power Co. (1971)

When the Civil Rights Act of 1964 was enacted, the Duke Power Company had 95 employees in a North Carolina facility, of which 14 were African American. The plant had five departments, including a labor department. The company had required a high school diploma in all departments except labor, the only department with African American employees. On July 2, 1965, the effective date of the Act, the company extended the high school requirement to the labor department and required acceptable scores on two aptitude tests installed at that time.

The lower courts ruled that the requirements were permissible as long as the company did not *intend* to discriminate on the basis of race. The unanimous

Supreme Court decision reversed the lower courts, and included many far-reaching provisions (*Griggs v. Duke Power Co.*, 1971):

1. Intent Versus Effect—The court stated unequivocally that good intentions cannot excuse the use of procedures that create obstacles, unrelated to performance, for minorities. It is the effect of a practice, not the intent behind it, which is important. The extension of this principle leads to the distinction between intentional discrimination, disparate treatment, and unintentional discrimination through a neutral device, or disparate impact.
2. Business Necessity—The Court said that the Act prohibits the use of practices that appear to be fair but have discriminatory effects. "The touchstone," it said, "is business necessity." Although the Court seemed to equate business necessity with job relatedness, other cases were needed to clarify the still-controversial concept.
3. Job Relatedness—Whether job relatedness is sufficient to show business necessity was not clear from this one decision; that it is a requirement for professionally justifying use of a selection procedure was not in doubt. When the decision was announced, many psychologists equated job relatedness with validity, but later decisions have shown distinctions.
4. Deference to *Guidelines*—When the case was heard, only initial *guidelines* that had been issued by the EEOC were available. The Court asserted that *Uniform Guidelines* issued by the EEOC were "entitled to great deference." That did not give the *Uniform Guidelines* the force of law, but *Uniform Guidelines* provisions are to be carefully considered in Title VII cases.
5. Use of Tests of Job Qualifications—The Court affirmed that the purpose of the law was to require that selection decisions be based on qualifications rather than on race or color. EEOC's endorsement of tests that are job related was entirely consistent with Congressional intent.

In a disparate impact case, the presentation of disparate impact data by the plaintiff establishes the ***prima facie case***; of course, it is never that simple and the presence of disparate impact is debated frequently by the defense. After the establishment of a prima facie case, the defense would respond by presenting evidence that the test or selection device is valid, job related, or a business necessity.

The impact of *Griggs* on selection and IO psychology has been immense, widespread, and long lasting. The implications and meaning of the case are still debated among experts. Along with *McDonnel-Douglas Corp. v. Green* (1973), the case established the two basic theories of discrimination and should be in the working memory of every selection specialist.

McDonnel-Douglas Corp. v. Green (1973)

Percy Green, an African American male, was fired by McDonnel-Douglas for illegal acts against the company. McDonnel-Douglas then advertised for mechanics and Green applied; after all, he was qualified—having held the position before he was fired. Green was not rehired and sued on the grounds of race discrimination. Although there was no direct evidence of intentional discrimination, no so-called "smoking gun," the Supreme Court ruled that Green had met the burden necessary for a *prima facie* claim under a disparate treatment theory. That is, Green had presented indirect rather than direct evidence of disparate treatment.

The criteria for establishing a prima facie claim are that the plaintiff: (1) belongs to a protected class, (2) applied and is qualified, (3) is rejected, and (4) the position stays open or is filled by someone from the nonprotected class, who is less qualified. Of course, merely establishing a prima facie claim does not mean the plaintiff will prevail. The defense is able to respond, and this is where it gets tricky, in terms of burdens of proof, persuasion, or production. But without getting too bogged down in the legal minutiae, at this point the defense, usually the organization or company, must rebut the prima facie case by articulating a legitimate reason. The plaintiff may then respond that the so-called legitimate reason is but a pretext for illegal discrimination, and then the defense must demonstrate it is not a pretext. Court cases rarely run in such an ordered fashion, but that is the basic logic of the presentation of evidence by the two sides. Interestingly, disparate impact data may be presented in a disparate treatment case as a form of circumstantial evidence and plays a much more prominent role in a special type of case referred to as a *pattern and practice* case (*Bernard v. Gulf Oil,* 1988).

> The court distinguishes two kinds of evidence under burden of proof: The ***burden of production*** involves putting forth enough evidence to provide a ***prima facie*** case. The ***burden of persuasion*** refers to "going forward" with the evidence. The burden of persuasion carries the risk of "nonpersuasion," meaning that failure to meet a preponderance of evidence standard, the issue is decided against the party that bears this burden. However, the exact application of these standards is subject to a great deal of debate in the legal literature.

Regents, University of California v. Bakke (1978)

This case (*Regents, University of California v. Bakke,* 1978) was heard under Title VI, the educational section of the Civil Rights Act, but it is one of the most important cases involving affirmative action (see the discussion later in this chapter) and has had implications for employment practices. California had two independent admissions programs, a regular program for most applicants and a special one for

minorities who claimed disadvantaged status. Allan Bakke, a White applicant to the Medical College of the University of California at Davis, was rejected in each of two years when minorities with substantially lower scores were admitted, and he sued successfully in California courts. The United States Supreme Court affirmed that the admissions system was unacceptable and that Bakke should be admitted, but it reversed the judgment that race cannot be legally considered. Its view was that racial diversity among medical students might be a legitimate consideration among others, but that the two-track system used at Davis violated constitutional protections.

Guardians v. New York (1980)

An often overlooked and underappreciated ruling, in *Guardians Association of the New York City Police Department, Inc. v. Civil Service Commission of the City of New York* (1980), the U.S. Court of Appeals for the Second Circuit "distilled from the *Uniform Guidelines* . . . five attributes" that tests should possess in order to be considered valid even where there is disparate impact (*M.O.C.H.A. Society, Inc. v. City of Buffalo*, 2009, 39–40):

1. A suitable job analysis
2. A reasonable demonstration of competence in the construction of the test
3. The content of the test must be related to the job
4. The content of the test must be representative of the job
5. The scoring system must result in the selection of better performers

This is a much shorter and simpler list than the one contained in the *Uniform Guidelines*.

Connecticut v. Teal (1982)

In the classic case involving a battery of tests, *Connecticut v. Teal* (1982), the Supreme Court ruled against using the "bottom line" as a safe harbor in a case involving a multiple-component promotion system. First, candidates for promotion were required to pass a written test. Those who passed were placed on an eligibility list from which selections were based on prior work performance, recommendations of supervisors, and seniority. Test results caused disparate impact, but on the bottom line more Black than White candidates were promoted. The Court's view was that Title VII sought to assure every *individual's* equality in employment opportunity, not to provide overall equality for racial groups. Given that disparate impact is specifically oriented toward groups, on the surface this seems like a very strange decision. Any component of the overall process that precludes further consideration is subject to disparate impact analysis and the subsequent requirement for evidence of job relatedness. This decision was extremely

important in organizations using a "multiple hurdles" approach to personnel decisions. It also demonstrates how difficult it is to simply avoid having disparate impact in a selection battery.

Watson v. Fort Worth Bank & Trust (1988)

Since the 1970 EEOC Guidelines, attempts had been made to regulate subjective assessments (e.g., unstructured employment interviews) as well as formal tests. *Watson v. Fort Worth Bank & Trust* (1988) examined the applicability of the disparate impact trigger to a case in which promotions were based primarily on supervisors' subjective recommendations. The Court was aware of its dilemma. Requiring disparate impact analysis for every unstructured consideration could lead to the adoption of surreptitious quotas to avoid litigation. Not requiring it could mask strongly discriminatory effects of apparently benign procedures. Even objective data such as test scores or diplomas could be combined with subjective interviews, the composite, therefore, being subjective, and the entire thrust of *Griggs* and its disparate impact trigger could disappear as a mechanism for enforcing Title VII.

Given these poles, what should courts do about subjective practices? On this question, the Supreme Court was divided. The decision of the plurality said that two standards of proof are required to show discrimination prima facie. First, a plaintiff must identify the *specific practice* being challenged—not easily done when the practice is a private, subjective judgment. Second, with the practice identified, the plaintiff must also present statistical data strong enough to convince the presiding judge that the practice has the effect of *causing loss* of equality of opportunity for members of a protected group. The decision argued that a "burden of persuasion" does not transfer to a defendant; as in other matters of evidence, the defendant has the opportunity to criticize or refute either the data or the causal inference.

The Court also said that the cost of alternative procedures is a factor to be considered; cost had not heretofore seemed to be a matter of much concern to the Court and certainly not to enforcement agencies. Similarly, for the first time, the Court also said that expensive validation studies were not needed, even for tests, when common sense and good judgment affirmed the job relatedness of the practice. Indeed, in matters of judging job relevance, lower courts were urged to defer in many matters to the greater expertise of employers in questions of business practice. This is a critical case in that it established that even subjective procedures, such as a supervisor interview or a homemade test administered by a manager, could be seen as a test and subject to the same standards laid out in the *Uniform Guidelines*.

Ricci v. DeStefano (2009)

Section 106 of the Civil Rights Act of 1991 prohibits altering the results of a test based on group status. But what if an organization or jurisdiction has reason to believe that a test it has given resulted in disparate impact? Can the organization

alter the results from the test or decide not to use the results from the assessment? These are the questions addressed by the Supreme Court in *Ricci v. DeStefano* (2009), which involved a promotional test for firefighters administered by New Haven, Connecticut.

After giving and scoring the promotional test, New Haven decided to invalidate the results, as it believed the resulting promotional decisions would result in disparate impact on the basis of race. Eighteen firefighters, 17 of which were White and 1 Hispanic, filed a reverse discrimination case on the basis of disparate treatment on the basis of race. The Supreme Court ruled against New Haven and for the plaintiffs, on the basis of the lack of a "strong basis in evidence." In the arguments accompanying the decision, the Court also suggested that the test would meet the standards for a valid test; one of the issues in the case was whether New Haven had attempted to determine whether the test was valid before deciding to ignore the results. From the standpoint of a selection expert, the case could be interpreted as reinforcing the need for validity and suggesting that a job-related test cannot be discarded simply because of the potential for disparate impact.

Wal-Mart Stores, Inc. v. Dukes (2011)

Big companies with a large number of employees make for dramatic court cases, and only a few court cases can match the size of *Wal-Mart Stores, Inc. v. Dukes* (2011). Although *Dukes* is primarily known for the ruling concerning class certification, it also involved consideration of issues including subjective decision making (see *Watson* earlier), stereotypes, unconscious bias, and the value and nature of expert testimony.

The lead plaintiff was Betty Dukes, who alleged on behalf of herself and approximately 1.6 million other women that females had been discriminated against across the board, including in pay and promotion. The tie in to *Watson* was that this sex discrimination occurred through subjective decision making, among other factors. One of the complicating factors in the case was whether and to what extent Wal-Mart had a centralized HR system; a cynic would argue that the plaintiffs were arguing that the lack of a centralized HR system was what led to the discriminatory subjective decision making, but for purposes of certifying the class were also arguing there was a centralized system. In a 5–4 decision, the Supreme Court ruled that the plaintiffs' class could not be certified on the basis of a variety of reasons, but primarily because of the lack of commonality among the members of the class.

Affirmative Action

Employers must not only avoid unlawful discrimination but also must take affirmative action to reduce the effects of prior discrimination; this is especially true for employers who accept money from the federal government and fall under the auspices of the OFCCP. Early examples of affirmative actions included recruiting

efforts, special training programs, direct mentoring, or extended probationary periods. Some affirmative action programs are voluntary, but many are imposed by court orders or consent decrees. Affirmative action is not a requirement under Title VII, although it is mentioned in the *Uniform Guidelines*. Since 1961, it has been a requirement for government contractors under the various Executive Orders (EOs), including the still-effective 11246. It has been controversial since the development of the Philadelphia Plan in 1969.

The Philadelphia Plan

The affirmative action requirement in EO 11246 posed a special problem for the building trades. Contractors do not generally have their own crews of skilled employees; they often hire those sent by unions. OFCCP investigations found few minorities in trade unions in the five-county Philadelphia area, despite a substantial minority population. The Secretary of Labor issued an order calling for increased proportions of minorities in each of six trades in each year of a four-year period. Any building contractor submitting a bid for a federal contract was required to submit with it an affirmative action program to show goals within these standard ranges and a plan for reaching them.

Contractors faced a dilemma when the Comptroller General of the United States issued an opinion that commitment to the plan was illegal and that disbursement of federal funds for a contract with such a program would be withheld as unlawful. An association of contractors sought help from the courts. The appellate court supported the plan; so did the Supreme Court, in effect, by declining to hear the case. Thus began the equating of affirmative action, once largely matters of recruiting and training, with numerical goals and timetables.

Reverse Discrimination

Affirmative action was initiated not to provide favoritism for groups of people, but to compensate partially for the effects of past discrimination. When courts find that an employer has a history of discrimination, affirmative action programs or even outright quotas may be mandated as remedies. When an employer independently sees evidence of disparate impact on a particular job or set of jobs, that employer may voluntarily establish affirmative action plans, goals, and timetables. Doing so, however, runs the risk of a reverse discrimination charge, and the plan must explicitly correct prior discrimination (see *Weber v. Kaiser Aluminum & Chemical Corporation,* 1977).

> ***Affirmative action*** is not the same as nondiscrimination, and the courts have not looked favorably upon *voluntary* affirmative action—in the absence of prior discrimination.

Employers still feel that they walk a fine line in the conflict between obedience to the EO and compliance with Title VII. The EEOC, at least, will not hold an employer liable (for reverse discrimination) for voluntary affirmative action programs if (a) facts show an actual or potential disparate impact from practices in existence or planned, (b) the plan corrects for prior discrimination as shown by discrepancy between the relevant proportion of the employer's work force and the relevant labor market, and (c) the available labor pool among protected demographic groups is "artificially" limited.

Developing Affirmative Action Plans

To establish a local affirmative action program, the employer should first identify jobs with evidence of either disparate impact or disparate treatment. If there are such jobs, the responsible practices should be identified and corrective plans developed. The plans need not be (and to be effective, probably should not be) restricted to hiring intentions. They may include special recruiting, educational or training programs, and plans for identifying and advancing those whose abilities are underutilized in their current positions. They must be limited, both in time and scope; they should not go beyond correction of prior disparate impact or disparate treatment either from a desire to "do good" or from fear of litigation.

Diversity as a Business Necessity

A Supreme Court decision in *Gratz v. Bollinger* (2001, 2003) concerned the University of Michigan's use of bonus points for minority applicants. Like *Bakke,* this case was heard under Title VI, but the implications for employment decision making are still present. In the mid-1990s, the University of Michigan's College of Literature, Science, and the Arts had instituted undergraduate selection guidelines under which every applicant from an underrepresented racial or ethnic minority group was automatically awarded 20 points of the 100 needed for admission. The university relied on the judgment in the *Bakke* case, suggesting that consideration of race as a factor in admissions might serve a compelling government interest in some cases. The university argued that the educational benefit resulting from a racially and ethnically diverse student body served such an interest. Large corporations such as 3M and General Motors filed briefs with the court, stressing the need to employ racially and ethnically diverse workforces.

The Court's decision in the case was mixed and seemed to ignore many of the well-established principles of testing and assessment for employment (Tenopyr, 2004). The court accepted the notion that diversity is a compelling state interest, justifying its consideration in undergraduate admissions. At the same time, it struck down the university's selection system that grants bonus points to minorities, arguing that such a procedure is tantamount to quota selection. In delivering the opinion of the court, Justice Rehnquist argued that consideration of race as a factor must

be done at the *individual* level, not at the group level. In other words, each applicant must be considered as a whole, and preference is to be given on a case-by-case basis. Justice Rehnquist used the example of an applicant with artistic ability so great as to equal the talent of Picasso, yet this student receives only 5 points under the university's selection guidelines. Every minority student, however, would automatically receive 20 points under the same system. The court's argument seems to suggest that, instead of receiving 20 points, minority applicants might be scored on the basis of *how much* diversity they would bring to the student body. The idea of individual differences in diversity seems a strange concept indeed.

In a dissenting opinion, Justice Souter gave credit to the university for its transparent selection system. According to Souter, "I would be tempted to give Michigan an extra point of its own for its frankness. Equal protection cannot become an exercise in which the winners are the ones who hide the ball." Like Justice Souter, we worry that the *Bollinger* decision may encourage organizations to "hide the ball" in their attempts to increase diversity. Doing so would imply using a selection system that does not have clearly defined standards regarding each qualification, and its relative importance in the hiring decision.

At this point, the status and legality of affirmative action programs remain unclear. Generally, affirmative action programs are more likely to be approved by the courts if there has been a history of past discrimination by the institution. An important factor in *Regents, University of California v. Bakke* (1978) was the absence of a previous history of discrimination by the University of California at Davis. In scrutinizing the use of affirmative action, greater deference is also likely to be given to softer procedures, such as including a subjective evaluation of hardships overcome, as compared with hard quotas, including the awarding of points for race or sex (Doverspike, Taylor, & Arthur, 2000).

Recent court cases have failed to do much to clarify how the courts view affirmative action. In *Grutter v. Bollinger* (2003), a court upheld the use of a diversity program with undergraduates at the University of Michigan. The *Ricci v. DeStefano* (2009) case is often seen as a blow against affirmative action in promotions. In *Fisher v. University of Texas* (2013), the Supreme Court allowed the University to use race in admission decisions as a way to achieve diversity. However, in *Schuette v. Coalition to Defend Affirmative Action* (2014), the Supreme Court voted to allow a state amendment that prohibited university admissions decisions based on race. This leaves us in a very uncertain place, especially because many of these cases deal with education rather than employment.

Age Discrimination

The Age Discrimination in Employment Act of 1967 (ADEA; Age Discrimination in Employment Act, 1967; Sterns, Doverspike, & Lax, 2005) prohibits discrimination against anyone 40 years of age or older. It encourages employment decisions about older people on the basis of ability, not age. It applies to hiring,

early retirement programs, promotion, benefits packages, and so on. It is enforced through the EEOC. Most ADEA litigation involves terminations—firing, reductions in force, or involuntary retirement. One hurdle to employees filing an ADEA claim has been the presumed need to show intentional discrimination. In *Smith v. City of Jackson* (2005), however, the Supreme Court ruled that evidence of disparate impact on older workers could be used in establishing a prima facie case. This is a difficult case to interpret and generalize though in that many age cases involve a string or series of judgments by independent decision makers; thus, in many cases a disparate impact theory may be difficult to prove. The ruling in *Smith,* along with the general aging of the workforce, indicates that more ADEA claims are likely to be filed in future years. Defense of an ADEA claim involves showing that factors other than age were determining considerations. For promotions, transfer, or terminations, these factors are usually performance ratings.

Discrimination Against Persons With Disabilities

The Rehabilitation Act of 1973, the Americans with Disabilities Act of 1990 (ADA), and the Americans with Disabilities Act Amendments Act of 2008 (ADAAA) prohibit discrimination against qualified people who have disabilities. As with amendments to civil rights legislation, the ADA was amended on the basis of a series of court cases that limited the application of the ADA. A *disabled person* is defined as one with a physical or mental impairment that substantially limits one or more major life activities, or who has a record of such impairment, or who is regarded as having such an impairment. "Major life activities" include caring for oneself, walking, speaking, seeing, hearing, and working. Impairment might be a physiological or mental condition, cosmetic disfigurement, anatomical loss, mental illness, or learning disability. The ADA does not protect people whose employment on a given job would threaten the safety or property of others.

The law requires employers to focus on what a candidate can do, not on disabilities. For a job to be filled, the employer must be able to distinguish essential functions of the job from those that, even if important, may not have to be performed by every incumbent. A clerical job, for example, may require operation of certain machines, reaching certain file drawers or shelves, and delivering occasional materials to people in other offices. If any one of these is deemed an essential function of the job, a qualified candidate must be able to do it. The ADA prohibits only discrimination against *qualified* candidates with disabilities. It does not require preferential hiring of qualified but disabled candidates; it explicitly encourages hiring the candidates *most* qualified to perform essential functions, irrespective of disabilities.

Reasonable Accommodation

Employers must offer reasonable accommodation to overcome barriers a disability may pose for an otherwise qualified candidate. An unusually short person may be

considered disabled. The disability may be a barrier to the filing function if file drawers or shelves are too high, but providing a stool may be enough accommodation to enable a short person to carry out that essential function. Thinking of accommodation as a major architectural change is often unwarranted; Jeanneret (1994) reported that about two thirds of all accommodation requests cost less than $500. Congress, EEOC, and the courts have stressed reasonableness; accommodation is not required if it would impose an undue hardship on the organization. A more difficult question is what constitutes reasonable accommodations when considering a person with a disability who is completing an examination as an applicant for a job.

General Employment Procedures

Candidates may not be asked on application forms or in interviews about disabling conditions, although questions about their abilities to perform essential functions are permissible and, as with the collection of sex and race information, there is the possibility that in the near future organizations may have to collect information on disabilities in order to comply with record keeping and numerical goals. Those with known disabilities may be asked to describe or demonstrate how they might (with or without accommodation) perform those functions, but they may not be questioned about the disability itself. Reasonable accommodation applies to application forms and interviews as well as to the job and job environment; accommodation might include providing application forms with large type, completing them orally while someone else fills in the blanks, providing an accessible interview location for people with mobility problems, providing an interpreter to sign for deaf candidates, or readers for blind ones, and so on.

Medical examinations or background checks are frequent parts of the employment process. Under the ADA, these procedures are permitted only after making a conditional job offer (i.e., conditioned on satisfactory results of these post-offer procedures). Putting off the medical examination until a tentative decision is made poses problems for some kinds of testing. Psychomotor tests might be used in medical diagnosis, and personality inventories might be used to diagnose other disabilities, but both are more often used as predictors of performance. When the evidence says that such tests have been evaluated for predicting performance on essential job functions, they need not be treated as part of the medical examination.

Negligent Hiring, Defamation, and Wrongful Discharge

EEO law has dominated the legal context for nearly half a century, but other kinds of laws also need attention. Among these are laws of *torts,* that is, wrongful acts resulting in injury. If an employee does something that results in injury to a coworker, a customer, or some other third party, the employer can be sued for

damages. The suit might be based on the doctrine that an employee carrying out assigned duties is an agent of the employer. More often, in states where they apply, the doctrines of negligent hiring and retention are being used. These hold that an employer can be found negligent in hiring or keeping an employee if (a) an injury was caused by an employee acting "under the auspices of employment," (b) the employee is shown to have been unfit for the job, (c) the employer knew or should have known about the unfitness, (d) the injury was a foreseeable consequence, and (e) the hiring or retention of the employee was the proximate cause of the injury.

Showing that an employee is "unfit" is not necessarily showing incompetence on the job. Much litigation in this area involves violence, so a person with a history of violent reactions to interpersonal frustrations may be deemed "unfit" for employment in jobs where potentially frustrating contact with others is likely. Being "unfit" includes (from case law) not only mental or personality disorders, but also more ordinary deficiencies. An employee's competence in driving may not be considered in determining fitness in a job in which driving ability is hardly a defining characteristic (e.g., a social worker), but it may be in a position in which the employee must drive from one site to another. It is unclear whether checking for a valid driver's license is enough to avoid liability for an employee at fault in an injury-producing accident between sites, but a finding of unfitness seems likely if the employee had a history of multiple at-fault traffic accidents.

The most common basis for dispute in negligent hiring and retention cases is whether the employer should have foreseen the possibility of unfitness. Employers may be held liable when a reasonable preemployment investigation would have revealed that the employee posed a threat to others. Employers should take steps to identify potential problems. In the previous example, perhaps checking for the license would be a sufficient precaution, but greater care would be shown by checking accident records, driving records, insurance papers, or perhaps giving a special driving test. Exercising prudence and identifying possible consequences is necessary if a person who is unfit in any specific way is put on the job.

Appropriate Methods of Assessment

Most writers on negligent hiring emphasize reference checks and background investigations—advice easier to give than to follow. Another legal doctrine, known as defamation, has made reference checks all but worthless. About the only information prudent employers give when asked about former employees is confirmation or disconfirmation of dates of employment and last job held, and some refuse even that. There is safety in the refusal. To be actionable under a charge of defamation, information given by the previous employer must be shown to be false, but the burden of proof falls on the employer, who must show that the statement made is true. Saying that an employee was discharged because of the supervisor's *opinion* that the employee was not trustworthy can be true if the opinion is a matter of record; it is, therefore, not defamatory. The same information given in a context of

innuendo permits the inference that the employee did, in fact, violate trust, without factual support for the inference, and it may be defamatory under the principle known as "slander *per quod*." Statements that do not hold up under legal scrutiny, whether false, partially true, or unsupported by evidence, may also serve as the basis for suit for wrongful discharge. All in all, the risks are usually deemed too severe to take on behalf of inquiry from outside the organization.

Background investigations run similar risks and may also violate a candidate's rights of privacy. Many kinds of public information can be tapped, but always with some risk that the information is erroneous. Questions of validity of references and background investigations are not new. Moreover, some resulting information cannot be used for employment decisions. Courts have repeatedly ruled against the use of arrest records, for example, to deny employment to those in demographic groups experiencing unusual arrest frequency—although convictions may be used. There is always a question of cost. Thorough background investigations are likely to be fruitless for young applicants and very expensive for older ones with more background to investigate.

Technology has introduced a new issue. Many companies today "Google" or cybervet for information on job candidates. Whole industries have emerged dedicated to both searching the Internet for data on candidates and assisting clients in deleting, hiding, and cleaning online information so as to create a better "brand" for the job applicant. Although there are dangers in such a search, the organization may discover that someone has a previously undisclosed disability, for example, many employers argue that they are more worried about possible negligent hiring claims or a public relations backlash if they fail to search for damaging information.

This is an evolving area of the law. Some states have already moved to restrict accessing or obtaining certain types of information, even if not illegal under federal law. There may be differences between using material discovered online to "screen in" versus "screen out" of jobs. As with many other areas, there are authentication issues, and people have already sued in instances where a "fake" online profile led to discipline or firing by an employer. At a minimum, organizations should be aware of this issue and take affirmative steps to begin to develop or create policies to deal with these types of online employment searches.

The International Association of Chiefs of Police Center for Social Media offers an excellent site dedicated to coverage of social networking issues including cybervetting. The site can be found at www.iacpsocialmedia.org/Resources.aspx.

A Final Comment

Not all aspects of the legal context for employment decisions have been described. Omissions include, for example, record-keeping requirements, rules governing immigrants, requirements for federal employment of part-time people, the polygraph protection act, state laws or laws of other countries. Moreover, what *is* described here is subject to changes in statutes, regulations, or court decisions.

We have tried to emphasize that personnel decisions must be made according to existing laws. The law is dynamic and ever-changing, and it varies by state or local jurisdiction. Changes follow or accompany (or are accompanied by) changes in the ideas and attitudes of society in general, whether emerging spontaneously or in response to leadership. Even imperfect law is an expression of, and an instrument of, social policy. Perhaps, then, the objective of this chapter is better described as trying to emphasize that personnel decisions must be made according not only to organizational policies and interests, but also to social policy and interest insofar as it can be understood.

Discussion Topics

1. What is meant by business necessity? What did the 1991 Civil Rights Act contribute to clarifying this issue?
2. Some organizations rely heavily on promotion from within. Would it be appropriate for these organizations to test for KSAs needed for a higher level job to fill a position that may not need these skills?
3. Should the ADEA protect people under 40 from age discrimination?
4. What do you think of the practice of cybervetting job candidates? Should there be limits, and what should the limits be? If an employer was to search through your online history, what would he or she find?

Note

1 Of the three authors, none of us is an attorney. We do not provide legal advice. Nevertheless, effective HR management requires some knowledge of applicable legal principles.

References

Age Discrimination in Employment Act of 1967, 29 U.S.C. Section 621 (1967).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51,* 201–238.

Bernard v. Gulf Oil Corporation, 841 F.2d 547 (5th Cir., 1988).

Civil Rights Act of 1964, 42 U.S.C. Section 2000e (1964).

Civil Rights Act of 1991, 42 U.S.C. Section 1981A (1991).

Connecticut v. Teal, 457, U.S. 440 (1982).

Doverspike, D., Taylor, M. A., & Arthur, W., Jr. (2000). *Affirmative action: A psychological perspective.* Huntington, NY: Nova Science Publishers.

EEOC v. Joe's Stone Crab, 136, F.2d 1311 (2001).

Equal Employment Opportunity Commission. (1970). Guidelines on employee selection procedures. *Federal Register, 35*(149), 12333–12336.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290–38315.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, Department of Justice (1979). Interpretation and clarification of the Uniform Employee Selection Guidelines. *Federal Register, 44,* 11996–12009.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, Department of Justice (1980). Adoption of additional questions and answers to clarify and provide a common interpretation of the Uniform Guidelines on Employee Selection Procedures. *Federal Register, 45,* 29529–29531.

Fisher v. University of Texas, 570 U.S. ___ (2013).

Gratz v. Bollinger, 135 F.2d 790 (2001).

Gratz v. Bollinger, 539 U.S. 244 (2003).

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Grutter v. Bollinger, 539 U.S. 306 (2003).

Guardians Association of the New York City Police Department, Inc. v. Civil Service Commission of the City of New York, 630, F.2d 79 (1980).

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gutman, A., Koppes, L. L., & Vodanovich, S. K. (2011). *EEO law and personnel practices* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Highhouse, S., & Gutman, A. (2011, January). Was the addition of sex to Title VII a joke? Two viewpoints. *The Industrial Organizational Psychologist, 48,* 102–110.

Hosanna-Tabor Evangelical Lutheran Church and School v. EEOC, 565 U.S. ___ (2012).

International Brotherhood of Teamsters v. United States, 431 U.S. 324 (1977).

Jeanneret, P. R. (1994, July). *Accommodation: State of the research and practice when complying with the Americans With Disabilities Act.* Address to the American Psychological Society, Washington, DC.

Johnson et al. v. City of Memphis, LEXIS 20644 (2014).

Latuga v. Hooters Inc., WL 164427, (1996).

McDonnel-Douglas Corp. v. Green, 411 U.S. 792 (1973).

M.O.C.H.A. Society, Inc. v. City of Buffalo WL 604898 (2009).

Regents, University of California v. Bakke, 438 U.S. 265 (1978).

Ricci v. DeStefano, 557 U.S. 557 (2009).
Schuette v. Coalition to Defend Affirmative Action, 572 U.S. ___ (2014).
Smith v. City of Jackson, 544 U.S. 228 (2005).
Sterns, H. L., Doverspike, D., & Lax, G. A. (2005). The Age Discrimination in Employment Act. In F. J. Landy (Ed.), *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives* (pp. 256–293). San Francisco, CA: Jossey-Bass.
Tenopyr, M. L. (2004, April). *The University of Michigan cases: Promises and problems.* Presentation at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.
Wal-Mart Stores, Inc. v. Dukes, 131 S. Ct. 2541 (2011).
Wards Cove Packing Co. v. Atonio, 109 S. Ct., 2115 (1989).
Watson v. Fort Worth Bank & Trust, 108 S. Ct. 2777 (1988).
Weber v. Kaiser Aluminum & Chemical Corporation, 563 F.2d 2126 (1977).

This page intentionally left blank

PART II

Knowing How to Assess

This page intentionally left blank

5

MINIMIZING ERROR IN MEASUREMENT

Measurement Theory, Reliability, and Validity

Assessment is a broader term than measurement. Measurement is a special case of assessment. It is based on a more defined scale along which scores can be ordered with relatively fine gradations (e.g., measurement of mechanical ability). Measurement seeks precision. In contrast, many other assessment procedures are ad hoc or used for specific practical purposes where precision is not useful or perhaps not possible. The topics of assessment and measurement, as well as topics such as reliability and validity, are usually covered in textbooks and courses dealing with psychometrics (we defined *psychometrics* in Chapter 1). A comprehensive discussion of psychometrics would require a book in itself, so in this chapter we focus on the topic of assessing reliability and minimizing error in measurement. In other words, our focus will be on enhancing measurement precision.

Background

Adolphe Quetelet, a Belgian astronomer and mathematician, noted that, if the center of a distribution of human observations were correct, or represented perfection, then nature erred equally often in either direction. He later found that distributions of social and moral data also followed this "normal law of error." That law was important to Francis Galton's studies of the inheritance of genius, and he used a crude index relating offspring and parent ability to test the proposition that ability is inherited. His index eventually led Karl Pearson to develop the product–moment coefficient of correlation. It treats the standard deviation of a more or less normal distribution (the "normal law") as a useful unit of measurement. It continues to be the unit in most psychological measurement.[1]

Cattell (1890) and others of his era developed several perceptual and sensory tests and tests of memory, which Hull (1928) considered academic aptitude tests. Employment tests were developed by Hugo Münsterberg at Harvard, clinical tests by Emil Kraepelin in Germany, and intelligence tests by Alfred Binet, Théodore Simon, and Victor Henri in France. Most mental ability tests of the early 20th century used Binet's question-and-answer approach. The same period saw projective personality tests and standardized school achievement testing. By mid-century, a specialized group of test experts, concerned about the proliferation of tests used with or without clear measurement properties, developed a set of "Technical Recommendations" for the development and evaluation of tests and test use (American Psychological Association [APA], American Educational Research Association [AERA], & National Council on Measurement in Education [NCME], 1954).

Different kinds of psychological measurement all have emphasized individual differences. Some have emphasized theories of psychological processes, but few have offered theories of the attributes measured. Usually the technique came first, followed later by questions of what the measures mean. Reliability was the dominant topic in measurement during much of the second quarter of the 20th century; later, validity became the dominant concern. In employment practice, validity is often equated with effectiveness of prediction, but in psychometric theory it refers more generally to score meaning.

Reliability: Concepts of Measurement Error

People differ. So do measures, for many reasons: flaws in measurement, the vagaries of chance, or traits measured—including traits not intended to be measured. Flaws, chance, and unintended traits are *measurement errors*. The concepts of reliability and validity both involve error, although in different ways.

Measurement Error and Error Variance

Errors happen in measurement. Two people using the same yardstick to measure the same kitchen table may get different results. A chemist using the most sophisticated equipment available may weigh a crucible several times, with results apparently differing only trivially, and settle for the average as the "true" weight. Mental measurements are even more error prone. Intelligence is an abstract, complex concept, nearly defying definition, yet it is routinely measured with tests. There are always measurement errors. Yet scores typically reflect fairly well the level of the trait being measured. It is sensible to assume, despite error, that one who scores high on an arithmetic test really is pretty good at arithmetic. The basic assumption of psychological testing is that any measure contains an element of error and an element of correctness, or "truth."

What accounts for differences in test scores? A measuring instrument, whether a yardstick or a test, is a constant stimulus; variance in measures stems from people's

responses to it. Table 5.1 shows why people get different scores on the same test. The first category of reasons in Table 5.1 suggests that scores will differ because of those general, long-term characteristics of applicants that will influence scores on virtually any test. This category includes such characteristics as general understanding of language, terms, expressions, instructions, and the skills used in test taking. The second category suggests that scores will differ because of individual differences in the characteristics measured by the test. It also suggests that some

TABLE 5.1 Reasons for Individual Differences in Test Performance

Reason	*Examples*
I. Reasons that are more or less permanent and that apply in a variety of testing situations.	A. Some traits (e.g., reading ability) influence performance on many different kinds of tests. B. Some people are more test-wise than others. C. Some people grasp the meaning of instructions more quickly and more completely than others.
II. Reasons that are more or less permanent but that apply mainly to the specific test being taken.	A. Some of these reasons apply to the whole test or to any equivalent forms of it. a. Some people have more of the ability or knowledge or skill being measured by the test. b. Some people find certain kinds of items easy while others may be more confused by them. B. Some reasons apply only to particular items on a test. If the test happens to contain a few of the specific items to which the person does not know the answer, that person will have a lower score than someone else who is luckier in the specific questions asked.
III. Reasons that are relatively temporary but would apply to almost any testing situation.	A. A person's health status may influence the score. B. A person may not do as well when he or she is particularly tired. C. The testing situation is challenging to some people; others may feel less motivation to do well. D. Individuals react differently to emotional stress. E. There may be some relatively temporary fluctuations in test-wiseness. F. A person varies from time to time in readiness to be tested. G. People respond differently to physical conditions (light, heat, etc.).

(*Continued*)

TABLE 5.1 (Continued)

Reason	*Examples*
IV. Reasons that are relatively temporary and apply mainly to a specific test.	A. Some reasons apply to the test as a whole (or to equivalent forms of it). a. People differ in their understanding of a specific set of instructions. b. Some people may "stumble" sooner into certain insights useful in tackling a particular test. c. Differences in the opportunities for practicing skills required in test performance. d. A person may be "up to" a test or "ripe" for it more at one time than at another. B. Some reasons apply only to particular test items. a. Momentary forgetfulness may make a person miss an item that might otherwise be answered correctly. b. The same thing can be said of momentary changes in level of attention or carefulness.
V. For measures involving interactions between examiner and examinee, for measures using open-ended responses to be evaluated on a complex basis, for measures involving ratings (e.g., performance evaluations, evaluations of work samples)—for all of these, scores may be influenced by characteristics of someone other than the examinee.	A. Conditions of testing may vary in conditions intended to be standard or controlled. B. Interactions between examiner and examinee characteristics (e.g., race, sex, age, personality) C. Bias or carelessness in rating or other evaluations of performance.
VI. Some reasons just cannot be pinned down after everything else has been taken into account.	

applicants with limited vocabularies may be lucky or unlucky enough to find specific words in the test that they do or do not happen to know. With a different set of words, these applicants would have scored differently.

The third category lists temporary characteristics of the applicants that could influence scores on any test; a person who is very nervous or distracted might do much better on a test under better circumstances. The fourth category includes reasons that are temporary and specific to the test or some part of it, such as temporarily getting stumped by a word usually known or recognized. The fifth category describes reasons for differences among scores that reflect conditions of administration of the test, such as interaction with the examiner or idiosyncrasies of scorers. Some of these have nothing to do with the

examinee, and others may be applicant reactions to the conditions. The sixth category reflects pure chance.

Notice that category II-A is the one of primary interest in measurement; other categories reflect unwanted variance in scores. Classical psychometric theory begins by assuming that any measure X (obtained score) is the algebraic sum of a true measure (true score) t and a measurement error (error score) e:

$$X = t + e \tag{1}$$

Further assumptions are (a) that true scores and error scores are not correlated, (b) error scores in one measure are not correlated with error scores in another, and (c) error scores in one are not correlated with true scores on something else. Together, these assumptions say that error scores are truly random. In fact, however, some errors are not purely random. A *true score,* if really true, contains no error, but the theory defines it as *the mean of an infinite number of a person's obtained scores on parallel measures of the same trait* (Thurstone, 1931), that is, measures with the same means, standard deviations, and distributions of item statistics. But if every obtained score in that infinite set contains the same error, the mean is the score one would intuitively consider "true," plus or minus that repeated, constant error. The theoretical error score, in short, does not include errors the person makes constantly over repeated testing; it includes only unpredictable, random error. If errors were only random, the mean of repeated measures would approximate an intuitively "true" score. The constant error across repeated measures for each person influences the mean of repeated measures precisely as it influences each individual measure.

One example of **systematic error** that differs from person to person might be a tendency for some people to prefer multiple-choice items, while other people find them confusing and instead prefer an essay question.

Distinguishing systematic, repeatable errors from errors that vary randomly across repeated measures allows us to rephrase the basic equation as:

$$X = s + e \tag{2}$$

Instead of t (true score), Equation 2 considers a person's actual score X to consist of a systematic score s (a composite of an intuitive true score and any systematically repeated error), and e, the random error. Equation 2 describes the score of just one person; the s score includes that person's own private constant error. These personal errors differ for different people, so a set of them has some variance.

A different sort of error is constant for everyone in the set. It influences all measures in the set equally and, therefore, has no variance. Classical reliability theory is concerned with data sets and variances, so the equation is expanded from describing a person's score (X) to one describing variance across people's scores:

$$\sigma_x^2 = \sigma_s^2 + \sigma_e^2 \tag{3}$$

where σ_x^2 is the total variance (i.e., differences from person to person) in a sample of scores, σ_s^2 is the variance or differences caused by systematic causes, and σ_e^2 is variance caused by random error.

Reliability

Technically, reliability is consistency in sets of measures or items. Equation 3 shows where the consistency comes from: the trait being measured and individual systematic (nonrandom) errors, with little variance caused by any kind of random errors. As a basic concept, then, ***reliability*** *is the extent to which a set of measurements is free from random-error variance*. In equation form, the conceptual definition of a reliability coefficient is as shown:

$$r_{xx} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \tag{4}$$

where r_{xx} is the theoretical reliability coefficient, σ_e^2 is the variance of the random sources of error, and σ_x^2 is the total variance. The smaller the error variance relative to total variance in obtained scores, the more reliable the measures in the distribution.

Keep in mind that this is a conceptual discussion; many different reliability coefficients can be computed from the same data, but each is simply an estimate of the theoretical reliability coefficient. In discussions of reliability, "measurement error" refers to random sources of error. Thus, we can think of reliability as defined in terms of the proportion of total variance attributable to systematic sources, but recognizing that it is *not* defined as the proportion of total variance caused by "true" variance is important. Such a definition would imply a specific trait, a specification irrelevant to an understanding of consistency (Tryon, 1957).

Reliability as a Necessary Condition for Validity

Reliability is often termed the sine qua non of mental measurement; if a test is not reliable, it cannot have any other merit. Imagine if every time you stepped on the bathroom scale it produced a different reading of your weight. Such a scale would be worthless (surely your weight would not change as a result of stepping on and

off the scale!). However, evidence of reliability is not in itself sufficient evidence that a measure is a good one. There is still the very important question of whether systematic sources of variance are relevant to the purpose of measurement. If systematic sources of variance on the vocabulary test were due to test-wiseness, and not ability in vocabulary, then those systematic sources are not relevant to the purpose of measurement. This is a matter of validity. Validity is the major consideration in test evaluation. Reliability is important because it imposes a ceiling for validity. The theoretical relationship of reliability to validity is shown by the formula:

$$r_{x\infty y\infty} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}} \tag{5}$$

where $r_{x\infty y\infty}$ is the *theoretical* correlation that would exist if predictor x and criterion y were perfectly reliable, r_{xy} is the validity coefficient actually obtained, and r_{xx} and r_{yy} are the respective reliability coefficients. This is known as correcting the validity coefficient for *attenuation,* that is, for unreliability.

It may be important for theoretical purposes to ask what the correlation would be if the two variables were measured with perfect reliability. That question is rarely important in personnel research. When we have an imperfect employment test, we use it anyway, use something else, or improve its reliability; in any case, we use a less than perfectly reliable test. There is little value in dreaming about the validities that might have been if only we had a perfectly reliable test.

In some situations, however, it is useful to know the level of validity with a perfectly reliable criterion, that is, to know how much of the *reliable* criterion variance is associated with predictor variance. We can find out by correcting for criterion unreliability only:

$$r_{xy\infty} = \frac{r_{xy}}{\sqrt{r_{yy}}} \tag{6}$$

where y is the criterion, y^{∞} is the perfectly reliable criterion, and x is the test. This is the correction for attenuation in the criterion.

Assume a validity coefficient of .40, better than many, but not noteworthy. Assume also a criterion reliability coefficient of .25, a terribly low reliability. Substituting in Equation 6, $r_{xy\infty} = .40/\sqrt{.25} = .40/.5 = .80$, the estimated correlation with a perfectly reliable criterion. The coefficient of determination for this hypothetical correlation is .64; 64% of the total *explainable* variance is accounted for by the test. Clearly, this offers reasonably effective prediction, given the limits of criterion unreliability. A validity coefficient expressed as the relationship of the predictor to the explainable criterion variance is a more standardized statement than the uncorrected coefficient, is less subject to the vagaries of random criterion variance, and generally makes more sense.

Reliability Estimation

Reliability is traditionally estimated by computing a correlation between two sets of measures, presumably measuring the same thing in the same sample of people in the same way. The two sets of scores could be scores on two different but equivalent forms of the same test, scores on the same test given at different times, or scores on two halves of a test. For each person, the two systematic scores are expected to be the same; systematic variance, therefore, causes, improves, or at least maintains correlation. Error scores are *not* the same, being random, so error variance comes from conditions or personal characteristics that differed in the two sets of scores. Random error variance inhibits or lowers correlation coefficients. If the effect of a source of variance is consistent in the two sets of scores, it is treated as a source of systematic variance. If it differs, it is treated as a source of error variance. Different estimates of reliability differ in the sources of variance treated as *systematic* (correlation-causing) or as *error* (correlation-reducing).

Each method for estimating reliability is a specific set of procedures for defining what is meant by reliability—an operational definition of reliability. Different operational definitions emphasize different sources of error variance. Giving a test a second time in exactly the same form and manner as before, the so-called test–retest method, considers stability the source of consistency. This method treats errors caused by particular items on the test, for example, as variance contributing to individuals' true scores (i.e., as a systematic source). Another method correlates scores obtained from equivalent forms of the same measuring device. This method treats peculiar items as a source of error variance. Either of these methods may be varied by allowing different amounts of time to elapse between measurements. If the two measures are obtained at pretty much the same time, a short-term effect on test performance (e.g., alertness or tiredness of test takers) is treated as a systematic source, but it is treated as error when the time interval between is large. Dividing a test into two equivalent halves is another method. With this technique, even very temporary characteristics, ordinarily considered sources of error in other procedures, may enhance correlation.

Coefficients of Stability

Stability means scores are consistent over time. ***A coefficient of stability*** defines random error as individual differences in score change (inconsistency) over an appreciable time period. *Test–retest* (using the same test) is useful if item sampling is not a problem. Test–retest correlation may be spuriously high if previous responses are remembered. Testing with an *equivalent form* (defined in the next section) after the time interval increases variance attributed to error, eliminating memory as an irrelevant source of systematic variance.

Coefficients of stability are useful for psychomotor or sensory tests if intervals are long enough to counteract practice or fatigue effects. Longer time intervals are

needed for cognitive tests. The appropriate time interval depends on how long people remember particular content and how often the content is practiced. A lot of skill practice or information use produces overlearning; those with initial high scores would surely repeat them, but over time those with low scores would improve. Differences in benefits of practice are treated as error variance.

Coefficients of Equivalence

Two test forms with different items are equivalent if (a) they have matching content (each has the same number of each kind of item), and (b) their means and standard deviations do not differ significantly. Equivalent forms are developed by specifying logical and statistical properties (item type and content; item difficulties, validities, and intercorrelations; test means and standard deviations) to which each of them will conform. Such item matching should yield correlated forms with essentially the same "true score" distributions. A ***coefficient of equivalence*** defines reliability as the extent to which a set of measures is free from errors because of sampling a test content domain. Actually, because genuine equivalence is hard to achieve, reliability estimates computed as correlations between equivalent forms are rather conservative (i.e., low). The conservatism is not so great with tests of well-defined content like vocabulary or arithmetic, but it may seriously distort reliability estimates for less well-defined areas such as temperament and motivation or for measurement by ratings.

Coefficients of Internal Consistency

Coefficients of internal consistency treat variance caused by variations in item content as a major source of error variance, and they show how much the variance is systematically based on a common concept measured by the test as a whole. In other words, they indicate how much the items in the test are getting at the same thing. Coefficients of internal consistency are widely used because of their convenience; they need only one administration with just one test form (if there is no time limit).

Kuder–Richardson Estimates. Techniques involving analysis of item variance are estimates of internal consistency. The most common of these methods were presented by Kuder and Richardson (1937) in a series of formulas; these formulas require the assumption of homogeneity (Cureton, 1950). The Kuder–Richardson formulas may be considered averages of all the split-half coefficients that would be obtained using all possible ways of dividing the test. The preferred formula (Richardson & Kuder, 1939) known as Kuder–Richardson Formula 20 (K–R 20) from the numbering of equations in the original derivation, is as shown:

$$r_{xx} = \left(\frac{n}{n-1}\right) \cdot \left(1 - \frac{\sum pq}{S_x^2}\right) \tag{7}$$

where n is the number of items in the test, p is the proportion of correct responses to a given item, $q = (1-p)$, and S_x^2 is the total test variance. Note that error variance is given as the sum of item variances, pq. This is a harsh assumption and may indicate why this formula gives a lower bound estimate of reliability (Guttman, 1945).

Finding references to Cronbach's ***coefficient alpha*** (Cronbach, 1951), a more general version of the K–R 20 equation, is now more common:

$$\alpha_n = \left(\frac{n}{n-1}\right) \cdot \left(1 - \frac{\sum S_i^2}{S_x^2}\right) \tag{8}$$

where α_n is $r_{xx'}$, the reliability coefficient called alpha for a test of n components (items or sets of items), S_i^2 is the variance of item responses or other component scores, and S_x^2 is the total score variance. If item responses are dichotomous, then $\sum S_i^2 = \sum pq$, and the equation for alpha is the same as K – R 20. Alpha can be used for items with response scales, ratings, or scores on small sets of dependent items such as a set of items based on a single passage or illustration.

Useful as it is, the alpha coefficient should not be used merely for convenience, and it should be interpreted only as internal consistency, not confused with equivalence or stability (Cortina, 1993; Schmitt, 1996). It is appropriate for most norm-referenced tests of abilities because these are typically constructed to provide homogeneous sets of items. It is not appropriate for domain-referenced tests constructed to represent a not necessarily homogeneous content domain.

Interrater Agreement

Two different observers seeing the same behavior or product may evaluate it differently—a source of error variance. With tests and rating scales scored by observer judgments, such a source of error can be large. The score depends on not only the behavior of the person observed or rated, but also scorer or rater responses and characteristics. ***Interrater reliability***, like other operational definitions, is often expressed as correlation. If there are several raters or scorers, a correlation matrix can be computed and an average determined, or intraclass correlation can be used. With dichotomous ratings, it may be expressed as the percentage of agreement between pairs of raters.

Comparisons Among Reliability Estimates

Estimates calculated by various methods turn out to be similar. If variance in a set of measures is generally systematic, with little of it attributable to random error, different operational definitions of reliability should agree fairly well. For this general statement to be so, the systematic variance has to be attributable to long-term, general characteristics of examinees, including the characteristics one is trying to measure. However, different methods make different assumptions, procedural

and mathematical, and define error differently. Researchers, test developers, and test users should use estimates that make sense for the circumstances they face. When a test is used to predict performance over a long period of time, stability is more important than internal consistency. If retesting is common enough to justify equivalent forms, coefficients of equivalence are needed. If production should be consistent month in and month out, an alpha coefficient over a period of several months is appropriate. The absolute values of stability, equivalence, and alpha coefficients may not differ very much, but small differences in reliability can make great differences in the appropriateness of decisions about individual people (Schmidt, Le, Ilies, 2003; Wainer & Thissen, 1996).

Standard Error of Measurement

So far, reliability has been defined and discussed in terms of distributions or sets of measurements. However, the basic datum is always a single measure, and the *reliability of an individual score* may be important—increasingly so as the ADA and selection for single positions preclude use of large data sets. The standard error of measurement, expressed in test score units, serves that purpose. Rearranging the definitional equation for reliability, we get the following:

$$s_e^2 = s_x^2 (1 - r_{xx}) \tag{9}$$

or,

$$s_e = s_x \sqrt{1 - r_{xx}} \tag{10}$$

where s_e is the *standard error of measurement*.

Using some simple math, you can quickly determine that the standard error of measurement is quite large, even for highly reliable tests. For example, consider an intelligence test with an estimated reliability coefficient of .90 and a standard deviation of 15, which is very close to the values we might find in the real world. Can you determine the standard error? You should be able to do so relatively easily by simply inserting the obtained values into Equation 10. If you substitute the values in the equation, you should find that the standard error of measurement equals approximately 4.74. Of course, that is only one standard error of measurement. Often we want to find a 95% confidence interval, which would be approximately 1.96 standard error units. If we multiple 4.74 by 1.96, we find that the 95% confidence interval would be approximately 9.29. Thus, if someone had a true score of 110 on an intelligence test, we would expect that 95% of the time their observed score would fall between 100.71 and 119.29.[2] On the surface, this would seem to be a wide range of scores, and this would be for an intelligence test with high reliability.

Standard errors of measurement have three uses in personnel decisions: (1) to determine whether scores for two people differ significantly, (2) to determine whether a person's score differs significantly from a hypothetical true score, and (3) to determine whether scores discriminate differently in different groups, such as different demographic groups or groups defined by different score ranges. The latter use should be more common than it is.

> If you were to take a test repeatedly, with no change in your standing on the attribute being measured, scores would vary around your true score on the attribute. The *typical* distance between your true score and your observed score is the ***standard error of measurement***.

In mass employment, it is important to know whether test scores distinguish people reliably in those regions of the distributions where hard decisions are made. They should; one evaluation of a test can ask whether the range with the minimal standard error of measurement is the crucial range for decisions. Standard errors may be computed independently for different regions, given enough cases.

Interpretation of Reliability Coefficients

Some people simplify reliability interpretation by stating a minimally satisfactory coefficient. It is not that simple. The interpretation must consider other information, including the intended use of the measures. For basic research, high reliability may not be critical. Decisions about individuals, however, require highly reliable measures. A reasonably sought level of "highly reliable" may depend on the history of a particular kind of measurement; "high" for interviews is lower than for standardized tests. Several other factors need to be considered in interpreting coefficients, including sample and item characteristics (see Guion, 2011).

One particularly important factor influencing reliability is the number of items in the test. With some exceptions (e.g., Li, Rosenthal, & Rubin, 1996; Wainer & Thissen, 1996) reliability generally is influenced by the length of a test or period of observation. Determining how long a test must be for adequate reliability is expressed in this formula:

$$n = \frac{r_{nn}(1 - r_{xx})}{r_{xx}(1 - r_{nn})} \tag{11}$$

where n is the number of times the existing test must be multiplied for a desired level of reliability, r_{nn} is that level, and r_{xx} is the reliability coefficient before lengthening the test. Use of the equation assumes that increments are equivalent to the

existing procedure. It may be applied only to coefficients of equivalence or of internal consistency.

Reliability improvement improves validity. Properly increasing test length by a specific value of n, estimated validity will be shown by:

$$r_{xny} = \frac{r_{xy}}{\sqrt{\frac{1}{n} + \left(1 - \frac{1}{n}\right) r_{xx}}} \tag{12}$$

where r_{xny} is the validity expected for the lengthened test x, and n is the factor by which the test is to be lengthened (Thorndike, 1949). Using selected values in this equation will show that, where a test is reasonably reliable to begin with, not much added validity will be gained through lengthening the test. Where, however, a low validity coefficient is due to low reliability, lengthening the test can be useful.

Psychometric Validity

The classical notion of validity used criteria only to judge the excellence of tests as trait measures. A test, it was generally said, "purports" to measure something, and validity is the degree it measures "what it is intended or purports to measure" (Drever, 1952, p. 304). This view differs from a later view of validity as the effectiveness of test use in predicting a criterion measuring something else and valued in its own right. The early concept of validity evaluated test scores as measures of a trait of interest; the later one evaluates test scores as predictors of something else. Investigations of both ideas have been called validation, results of either are called validity, and data collected for one of these evaluations may (but may not) be useful for the other. The distinction has not been commonly recognized.

People tend to use verbal shorthand, referring to "test validity" as if validity were a property of the method of measurement. It is not. It is a property of the *inferences* drawn from test *scores*; the inferences (interpretations) may be descriptive or relational.

> It is the ***inferences*** we make from test scores that are either valid or not valid—not the tests themselves. If we believe that a test should predict job performance, we would validate that inference or belief by correlating scores on the test with job performance measures. If we believe that a test should measure perceptual accuracy, we could correlate its scores with a different, preferably better, measure of the same sort of perceptual accuracy.

Three Troublesome Adjectives

Early attempts to clarify the validity concept (APA, AERA, & NCME, 1954; 1966) described criterion-related, content, and construct validity as aspects of validity;

however, no general definition of validity was offered. Criterion-related validity was shown by the relation of test scores to an external criterion. Content validity was a matter of the fidelity of sampling a content domain in the construction of the test. Construct validity was more complex, requiring a showing of reasons both for inferring a particular construct from the test scores and for not inferring alternative constructs. The three came to be treated as if they were three different *kinds* of validity, not *aspects,* an error of interpretation forcefully criticized by Dunnette and Borman (1979) and by Guion (1980) as psychometric trinitarianism. At least since Cronbach (1971), validity concepts have emphasized the meaning of scores: how a score can be interpreted, or what can be inferred about a person with that score. *Inferences are constructs,* and the "unitarian" view that has emerged treats the notion of validity, with no modifying adjective, as an expanded view of what was called construct validity (Messick, 1989, 1995).

Descriptive and Relational Inferences

Descriptive inferences interpret scores in terms of the characteristics revealed by the measurement procedure itself. For example, a high score on a cognitive ability test suggests that the test taker is high in intelligence. ***Relational inferences*** interpret scores in terms of different but correlated characteristics. For example, a high score on a cognitive ability test suggests that the test taker will perform well on a job that is mentally challenging. These are not wholly independent. The validity of descriptive inferences depends on several sources of information, relational data among them; it is associated closely with the idea of construct validity. Relational inferences are not well understood without understanding the descriptive properties of the related variables. Nevertheless, the distinction emphasizes different demands on validation: the difference between evaluating the success with which a construct is measured and evaluating its use. Both are important, but so is the distinction. *For personnel decisions, the distinction is between interpreting scores as traits and interpreting them as signs of something else.*

A relational inference is made when one infers from a score a corresponding level of performance on a criterion; it is usually evaluated by correlations. There is almost always more than that to be inferred from a well-understood test score. Validity is more than a correlation coefficient. To be sure, a test can be designed to do no more than predict a criterion—having no meaning at all if the criterion changes. A change in the job or technology or context can destroy the validity of such a limited relational inference, and no one will know why.

Usually, several constructs can be offered as plausible descriptive interpretations. One of them may be intended by the developer or user; others may be unwanted contaminants. If scores can sensibly be interpreted in terms of the intended construct or meaning, but not in terms of the intrusive others, then the intended descriptive inferences are surely valid, apart from any relational inferences that may also be valid.

Psychometric Validity Defined

Validation for descriptive inferences seeks confirmation of the meaning of test scores intended by the test developer (or some subsequent meaning intended by a test user) and disconfirmation of plausible alternative meanings. Because such validation is procedurally different from traditional employment test validation, we distinguish evaluating (validating) descriptive inferences from validating relational ones. For personnel assessment, with apology for yet another adjective, we call the result of the former *psychometric validity*. The result of the latter we call *job relatedness,* at least in personnel decision contexts.

> Evidence for descriptive inferences is *psychometric validity,* and evidence for relational inferences is *job relatedness*.

This chapter emphasizes psychometric validity. It is intended to examine classical psychometric theory and look beyond the comfortable limits that corral validity within a coefficient. Validity is itself an inference—a conclusion reached from an abundance of information and data.

The simple, fundamental question of psychometric validity is, "How well has the intended characteristic been measured?" More precisely, the question asks, "With what confidence can the scores resulting from the measurement be interpreted as representing varying degrees or levels of the specified characteristic?" There is never a simple answer. Answers are judgments, not numbers, and they are to be supported by data and logical argument. They depend on the relative weight of evidence—the weight of accumulated evidence supporting an interpretation relative to the weight of accumulated evidence opposing it. One looks not at single bits of information but at the preponderance of the evidence.

Varieties of Psychometric Validity Evidence

Evidence Based on Test Development

Did the developer of the procedure have a clear idea of the attribute to be measured?
This is a question of intentions; the developer of the procedure must have had something in mind to be measured. It may have been a thoroughly established construct or little more than a vague idea of a continuum along which people or objects could be ordered. It may have been a theoretical construct such as latent anxiety, something empirically tangible such as the smoothness of a machined surface, or something observable such as coordination of motor responses to visual stimuli. These are all abstractions, attributes of people or objects of concern. A small but positive sign of validity is if development followed a clear conception of the

attribute to be measured. A large, negative piece of evidence is if the developer has not bothered or is unable to describe the attribute measured, how it matches or differs from other attributes under other names, or whether it is an attribute of people, of groups of people, or of objects people do something with or to.

Are the mechanics of measurement consistent with the concept? Most psychological measurement is based on the responses people make to standard stimuli presented according to standard procedural rules. If the developer had a clear idea of what was to be measured, this idea should have guided a plan for procedures, and further questions like these need answers:

- Is the presentation medium appropriate? Does showing the test on an iPad fit the definition of the attribute to be measured?
- Are there rules of administration, such as time limits? If so, were they dictated by the nature of the trait being measured, or were they chosen for convenience?
- Are response requirements appropriate? It is not appropriate to use a recognition-based, multiple-choice item type for a construct defined in terms of free recall; it is not appropriate to use verbal questions and answers for constructs defined as physical skills.

Satisfactory answers to such questions provide only slight evidence of validity, but unsatisfactory answers—or no answers at all—are reasons for questioning assertions of validity.

Is the stimulus content appropriate? The content of a measurement procedure should certainly fit the nature of the attribute to be measured. This is more than so-called content validity. If the attribute to be measured implies a specific content domain, such as knowledge of the content of a training program, then content-oriented test development—with its insistence on domain definition and rules for domain sampling—constitutes useful and strong evidence of validity. But the principle applies also to more abstract constructs such as those developed by factor analyses. For tests of factorial constructs, item types defining the factor in prior research should be used, or evidence should show that the item type chosen taps the factor satisfactorily.

Was the test carefully and skillfully developed? When judgments are required by the test developer (and they nearly always are), they and the reasons for them should be a matter of record. Useful evidence also comes from answers to questions like these:

- Were pilot studies done to try out ideas, especially if they are unusual, about item types, instructions, time limits, ambient conditions, or other standardizing aspects of the test?
- Was item selection based on item analysis? Were appropriate item statistics computed and used?

- Did the data come from an appropriate sample? Was the sample large enough to yield reliable statistics?
- Does the final mix of selected items fit the original plan, or is there some imbalance? Was the item pool big enough to permit stringent criteria for item retention?

Evidence Based on Reliability

Is the internal statistical evidence satisfactory? Classical item analysis looks for two item characteristics: (1) ***difficulty*** level, usually expressed in the reverse as the percentage giving the correct item response, and (2) ***discrimination*** index, typically expressed as the correlation of item responses to total scores. Item statistics can be examined for spread and average difficulty or discrimination indices to see if they are appropriate for the anticipated measurement purposes. A test that is too easy or too hard for the people who take it will not permit valid inferences. Item statistics should be evaluated, of course, in the light of the circumstances that produced them. Their usefulness may depend on such things as sample size, appropriateness of the sample to the intended population, and probable distributions of the attribute in the sample. No universally correct statement of the most desirable item characteristics can be made. Ordinarily, one might consider a variety of item difficulties a sign of a "good" test. For some kinds of personnel decisions, however, a narrow band of difficulties might enhance precision in a critical region and be considered better evidence of validity than a broad band.

Responses to individual items should be somewhat correlated with the total score on all items; otherwise, no clearly definable variable is measured. Usually, a rather high level of internal consistency is wanted. A high coefficient alpha does not provide positive evidence that the item set as a whole is measuring what it is supposed to measure, but it *does* offer assurance of systematic content. If other information (such as meritorious care in defining the construct and in developing items to match the definition) makes it reasonable to assume that most items have, indeed, measured the intended concept, then a satisfactory alpha is reasonable evidence that the scores reflect it without much contamination. Constructs vary widely in specificity. Some are very narrowly defined, such as any one of the 120 constructs in the Guilford structure of intellect model (Guilford, 1959); others are very broadly defined, such as a construct of creativity, which may include many narrower constructs. *Tightly defined constructs require high internal consistency coefficients.*

Are scores stable over time and consistent with alternative measures? Stability over some not-too-brief time period seems essential, especially if internal consistency is relatively less important. If equivalent forms of the test have been developed, they should at least meet the minimal requirements of equivalence (common means and variances) and correlate well. If scoring is done by observers rating performance or its outcome, then certainly interrater agreement is an essential ingredient of

reliability. Because reliability limits validity, evidence of high reliability suggests good descriptive validity, but consistency may be due to consistent error.

Evidence From Patterns of Correlates

Correlating scores on a measurement procedure to be evaluated with other measures may yield evidence of validity. Such research provides information of two kinds, both equally important to conclusions about validity. One is ***confirmatory evidence,*** evidence that confirms (or fails to confirm) an intended inference from test scores. It is evidence that relationships logically expected from the theory of the attribute are, in fact, found. The other is ***disconfirmatory evidence,*** evidence ruling out alternative inferences or interpretations of scores—evidence that relationships *not* expected by the nature of the construct are, in fact, *not* found. Confirmatory evidence that does, in fact, confirm the intended interpretation is a necessary but insufficient condition for accepting scores as valid measures of an intended trait. Evidence is also needed that plausible alternative interpretations can be rejected (i.e., disconfirmed).

Multitrait–multimethod matrices are commonly used to study correlates with other variables. Two or more traits are identified (one measured by the assessment method at hand), and they are each measured by two or more methods. Confirmatory evidence of validity exists if correlations among measures of the same trait across methods are higher than correlations among traits within common methods (Campbell & Fiske, 1959).

One form of statistical evidence that has pleased some people remarkably well is a high *validity coefficient* (i.e., the correlation between a predictor and some criterion). Many people place far too much faith in a single validity coefficient. A high validity coefficient might stem from a common contamination in both the instrument being validated and the criterion. Suppose that performance ratings of school principals are contaminated by a general stereotype that a good principal is physically tall, is imposing in stature, looks like a scholar, and speaks in a low, soft voice. If the measure to be validated is an interview rating of administrative potential, and if these ratings are influenced by that same stereotype, there will be a high validity coefficient. It does not follow that the interview ratings are good indicators of administrative ability.

Another problem with a single validity coefficient is that it seeks only evidence confirming (or failing to confirm) a particular inference. It says nothing to confirm or to disconfirm alternative inferences. Validity coefficients are, of course, valuable bits of evidence in making judgments about validity, but *one should not confuse validity coefficients with validity,* and one should not base judgments about validity on validity coefficients alone.

Does empirical evidence confirm logically expected relations with other variables? The theory of the attribute will suggest that good measures of it will

correlate with some things but not with others, and at least some of these hypotheses can be tested. Evidence supporting them also confirms the validity of the scores as measures of the intended attribute. Traditional criterion-related validation procedures follow this logic. One might hypothesize from one's theoretical view of mechanical aptitude that those who score high on such a test will do better in an auto mechanics school than will those with low scores. To test the hypothesis, scores are correlated with grades in the school. A significant, positive correlation is evidence of validity for both relational and descriptive inferences. Testing other hypotheses, perhaps showing correlations with the number of correct troubleshooting diagnoses in a standardized set of aberrant pieces of equipment, or the speed with which a bicycle is taken apart and reassembled, gives further evidence of psychometric validity. Every such hypothesis supported provides further confirming evidence for the validity of interpreting scores as measures of mechanical aptitude as it has been defined. Failure to support hypotheses casts doubt on (a) the validity of the inference or (b) the match of the theory of the attribute with the operations (i.e., of the conceptual and the operational definitions of the attribute).

Does empirical evidence disconfirm alternative meanings of test scores? In practical terms, this means ruling out contaminations. Work sample scores should not be biased by the particular equipment an examinee happens to use. Performance ratings should not be biased by differential stereotypes among raters. Work attitude scores should not be biased by a social desirability response set. A test of spelling ability should not be biased by a printed format that requires excellent visual acuity. Many such problems can be guarded against during test development, but some need empirical study. Failure to disconfirm the more plausible contaminants may suggest validity problems.

Are the consequences of test use consistent with the meaning of the construct being measured? A *theory of an attribute* should identify outcomes or consequences of test use relevant to the construct. For example, if the attribute to be measured involves flexibility in thinking about problems, the theory of the attribute may include flexibility in solving problems of malfunctioning equipment. If a test of the attribute is used to select mechanics, then one consequence of its use is that high scorers are likely to think of and try alternative explanations for mechanical problems—hence, to solve more of them. If high scorers actually do solve more problems than do low scorers, it is evidence of valid measurement of the construct as well as evidence supporting the predictive hypothesis.

Beyond Classical Test Theory

Classical psychometric theory has served well and is sufficient for many practical uses. Extensions of classical theory, and alternatives to it, have been developed. They are useful, but they may sometimes require more resources (e.g., extremely large number of test takers, opportunities for repeated measurement, time, money)

than most personnel researchers will have. Even so, they offer concepts worth considering. These discussions are brief and elementary, but some awareness of these methods is needed by anyone even slightly involved in personnel assessment.

Factor Analysis

In principle, factor analysis identifies dimensions underlying test scores by finding clusters of highly correlated variables, which are minimally correlated with other clusters of variables. All of these methods, from the simplest to the most complex, use a common set of variables, all measured in the same sample, providing a matrix of correlations or covariances.

A small correlation matrix, good only for illustrative purposes, is shown in Table 5.2. The variables are four different tests. Each cell in the matrix shows the correlation coefficient computed for a pair of tests, such as the correlation of .65 between Tests A and B. Let Test A be a hypothetical test of general mental ability, B a vocabulary test, C a test of reading speed, and D a perceptual speed test. What factors account for scores on them? It is easier to understand what happens in factor analysis if we work backwards from an answer to the question. Possible answers are in Figure 5.1, showing underlying, latent factors and their contributions to score variance on each test. In this example, Test A should correlate well with Test B because most of the total variance in each is due to individual differences in language ability. The correlation between Tests A and C may be even higher; more of the total variances from these tests stem from common dimensions—in this case, two of them. Test D correlates only slightly with A and C (only small common sources of variance) and not at all with B (because they measure nothing in common). Error, of course, refers to the unreliable component of scores attributed to random sources which, by definition, should be uncorrelated across the four tests.

This backward approach is, of course, unreal. In practice, we know only the correlations between the tests and we draw inferences about the factorial structure of the tests from the correlations. A matrix like Table 5.2 would

TABLE 5.2 Correlation Matrix Showing Hypothetical Relations Among Four Tests

	A	*B*	*C*	*D*
A	—	.65	.63	.15
B		—	.52	.00
C			—	.17
D				—

Test	*Factors Actually Measured*	*% of Total Variance*
A: Mental Ability	Language Ability	55
	Perceptual Speed	40
	Error	5
B: Vocabulary	Language Ability	70
	General Reasoning	28
	Error	2
C: Reading Speed	Language Ability	35
	Perceptual Speed	50
	Word Fluency	10
	Error	5
D: Perceptual Speed	Perceptual Speed	6
	Finger Dexterity	70
	Visual Acuity	10
	Error	14

FIGURE 5.1 Dimensions contributing to total text variance in each of four hypothetical tests.

not allow us to determine the full structure of these tests. Of course, in this matrix, fancy statistical analysis would be unnecessary anyway. One can simply look at the correlations and know that Tests A, B, and C are all measuring the same thing to some degree, and that Test D does not measure it. Knowing something about these tests, we can simply look at the matrix and draw some inference about the nature of that "same thing." Obviously, Tests A, B, and C all require test takers to understand verbally expressed ideas. It is, therefore, plausible to infer that the ability to satisfy this requirement is one underlying cause of the correlations observed; it can be tagged "language ability," and it is a "factor."

Three of these tests, according to Figure 5.1, require at least some perceptual speed; an actual factor analysis (if it could be done on a 4-variable matrix) would identify perceptual speed as a factor. The matrix is too small, however, to identify some of the dimensions in Figure 5.1. The data for a factor analysis must include at least two variables for each anticipated factor. General reasoning ability, word fluency, finger dexterity, and visual acuity are also sources of variance in the matrix, but each of these influences scores on only one test and they could not be identified by a factor analysis of this matrix. Although specified as systematic sources of variance in creating the example, factor analysis of such an inadequate matrix would treat them only as sources of error variance—as if the variance they produce were random—because they are not systematic sources across two or more variables. At a minimum, four more tests would have to be included in the set for these factors to be identified.

Generalizability Theory

Generalizability theory examines the limits or boundaries within which score meanings generalize. Scores may be influenced by particular circumstances and be useful only if the assessment generalizes to other circumstances (e.g., other times, other behavior samples, other test forms, other raters or interviewers). An assessment is a valuable aid to decisions only if inferences drawn from it are like those drawn under other conditions. Generalizability theory as developed by Cronbach, Gleser, Nanda, and Rajaratnam (1972) described principles for testing the limits of the generalizability of inferences from trait measurements. Here we present only their basic logic.

Assessments are done in given sets of circumstances—by a certain person, on a particular day or time of day, in a certain room or other location, with specific ambient temperature or noise or distractions, and so on. If score interpretations were limited to any single combination of these specific circumstances, they would have little interest to anyone. Such circumstances are usually expected to be matters of indifference, variables that have at most a trivial influence on the outcome. We want to generalize the assessment inference to the one we would have made from assessment of the person on another occasion, in another setting, or with another administrator or ambience.

Traditional reliability estimation inquires into limited kinds of generalizability. Internal consistency coefficients refer to generalizing across the various items or observations. Stability coefficients tell whether inferences generalize across testing occasions. A generalizability study can answer both kinds of questions by collecting data in an analysis of variance design.

> A ***generalizability study*** simultaneously estimates variance attributable to sources such as persons, items, and occasions.

If all items are used for all persons on all experimental occasions, the design is "fully crossed," expressed as $p \times i \times o$, and shown by Venn diagrams as part *a* of Figure 5.2. Part *b* depicts a design in which two (presumably equivalent) sets of items are used on two occasions, as if using equivalent forms in a test–retest reliability study. Items would be "nested within" occasions, expressed as a $p \times i{:}o$ design. These are only two of the many different designs one might choose for studies of two sources of variance other than the traits of the people assessed. Clearly, either design provides more information than does a single reliability coefficient.

Generalizability theory would appear to provide an answer to many classic reliability problems. So why has it been used infrequently? There would appear to be two primary reasons. First, although the basic principles of generalizability theory are easy to understand, designing, carrying out, and analyzing the results of

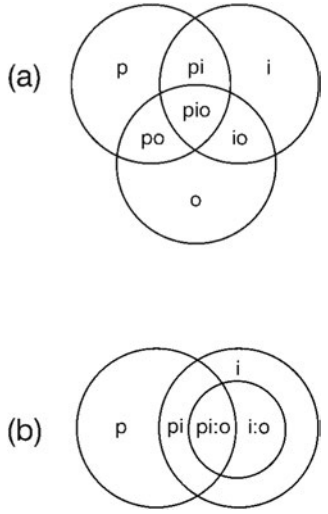

FIGURE 5.2 Two designs for person, item, and occasion generalizability studies.

generalizability studies tend to be quite difficult. Second, this has been exacerbated by a scarcity of simple to use software for calculating generalizability coefficients. (For one of the more accessible discussions of generalizability theory, see Cardinet, Tourneur, & Allal, 1976).

Item Response Theory

Statistics used in classical test theory—item statistics, distribution descriptive statistics, and the various correlations of reliability and validity estimation—depend on the distribution of the measured construct in the sample providing the data. If the sample is truly representative of the population from which it is drawn, then these sample statistics can be taken as reasonable estimates of population values. However, if sample distributions differ markedly from population distributions, they may be poor descriptors of population values. This problem led to the development *item response theory (IRT)* and procedures for developing relatively *sample-free* estimates of population values.

> With classical test theory, your score on a test reflects your standing relative to the rest of the sample of test takers. With *item response theory,* your score reflects your standing on the latent factor, or trait, measured by the test.

IRT is based on the commonsense idea that people with a lot of a specific ability are more likely to give the right answer to an item requiring that ability

than are people with less ability. A systematic relationship can be assumed between levels of the trait and the likelihood of a specified item response. The relationship can be modeled as a mathematical function, or equation, defining an *item characteristic curve (ICC).*

Figure 5.3 shows a sample ICC for an item. The X-axis corresponds to the latent trait; for example, it could be your knowledge of psychology. The measure or score on the latent ability is designated *theta,* or θ, and it is usually represented as a standardized score with a range from –3 to +3. The Y-axis is the probability of a correct response to an item, which can range from 0% to 100%. In this case, it could represent the probability of your correctly responding to an item on a psychology test. The resulting curve, or ICC, provides a way to estimate the probability of a correct response to an item, given a certain level of the latent trait.

An ICC may have many forms, but the most common is a positive, monotonic curve, as depicted in Figure 5.3. With a monotonic curve, the probability of a correct response continually increases (or continually decreases) with ability level, but not at a constant rate. People in a low-ability range have little likelihood of giving a correct or keyed response; regardless of specific levels of low ability, the slope of the curve in the low range is very slight. In a middle range of ability, the change in probability increases sharply with increasing increments in ability—up to a certain point—after which there are further but progressively smaller changes. That is, in a middle range of ability, the slope of the curve is increasingly steep up to a point.

The slope of the ICC is also referred to as the *a* parameter. The location of the curve, which corresponds roughly to the point along the theta scale where there is a 50% chance of a correct response, is referred to as the *b* parameter. Finally, the Y-intercept corresponds to the probability that someone with very little of the

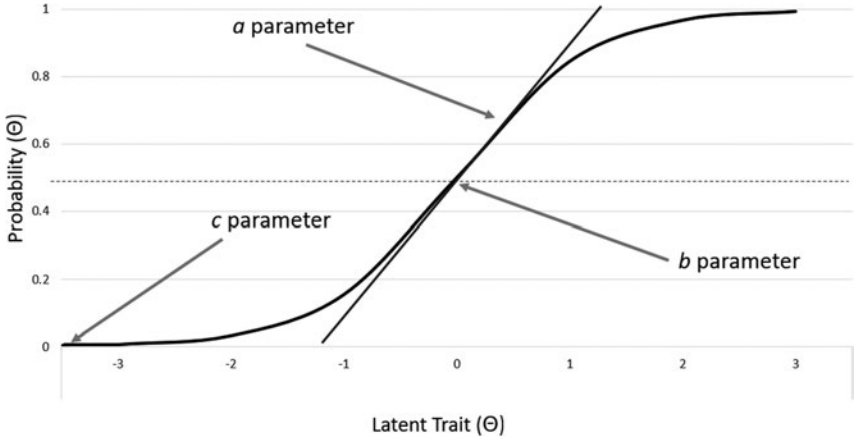

FIGURE 5.3 A sample item characteristic curve (ICC).

latent trait will guess the correct response, and is referred to as the *c* parameter. Although it is possible to use computer software to estimate all three parameters, a very popular approach in education and psychology is the *Rasch* model, which involves finding a solution for only one parameter, the *b* parameter.

Computerized Adaptive Testing With IRT. Think of an examinee on one side of a table and a test administrator, with a very large deck of cards, on the other. Each card has on it a test item and, visible only to the administrator, a difficulty level. The administrator chooses an item of low difficulty first. If the examinee answers it correctly, the administrator chooses a second, harder item. If the first answer is wrong, the second item chosen is easier. A few more items are similarly chosen to identify the likely region of θ. A few more items are chosen most appropriate for information in that region. The result can be a very precise estimate of θ for that person with only a few, carefully chosen items. The scenario is unlikely. The idea of someone sitting with a large deck of cards picking out items is, frankly, boring. The scenario is not at all unlikely, however, if the examiner is a computer and the cards are entries in its data bank. A computer program can do it almost instantly. The result is called *computer adaptive testing*. This is discussed further in Chapter 10.

Analysis of Bias. IRT is useful for EEO concerns because item parameter estimation is independent of the ability distribution in the sample studied. If the trait measured is not itself correlated with sex, race, or idiosyncracies of a particular culture, then subgroups based on sex, race, or culture should yield the same invariant ICC parameters within linear transformations. IRT analysis is, in fact, sometimes used to identify items that function differently in the different subgroups. As a technique for identifying bias, differential item functioning is discussed in more detail in Chapter 9.

Discussion Topics

1. How would you explain reliability and validity to a high school student? Why is reliability a necessary condition for validity?
2. Why is it inappropriate to speak of a test as being valid or not valid?
3. You have been given the task of designing a new test of emotional intelligence. Discuss how you would build the case for the psychometric validity of your new measure of emotional intelligence. What confirmatory evidence could you offer? What disconfirmatory evidence could you offer?

Notes

1 Boring (1961) pointed out, first, that one cannot assume that a mathematical function such as the "normal law" applies to a particular variable until it has been demonstrated empirically, which Galton and most of his followers failed to do. He then went on to say,

"The *a priori* assumption that the normal law applies to biological and psychological variables, and, therefore, provides a device for changing ordinal scales into equal intervals has continued well into the present century. The scaling of mental tests in terms of standard deviations . . . in some ways preserves this ancient fallacy" (p. 123).

2 The interpretation of the confidence intervals associated with the scores and the standard error of measurement is more complex and controversial than presented here. For a more in-depth discussion, you may want to ask your instructor to discuss the topic or consult a good psychometrics text.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51,* 201–238.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association.

Boring, E. G. (1961). The beginning and growth of measurement in psychology. In H. Woolf (Ed.), *Quantification: A history of the meaning of measurement in the natural and social sciences* (pp. 108–127). Indianapolis, IN: Bobbs-Merrill.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cardinet, J., Tourneur, W., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement, 13*(2), 119–135.

Cattell, J. M. (1890). Mental tests and measurements. *Mind, 15,* 373–380.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78,* 98.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Cureton, E. E. (1950). Validity. In E. F. Lindquist (Ed.), *Educational measurement.* (pp. 621–694). Washington, DC: American Council on Education.

Drever, J. (1952). *A dictionary of psychology.* Baltimore, MD: Penguin.

Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and classification. *Annual Review of Psychology, 30,* 477–525.

Guilford, J. P. (1959). *Personality.* New York, NY: McGraw-Hill.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11,* 385–398.

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions.* New York, NY: Routledge.

Guttman, L. (1945). A basis for analyzing test–retest reliability. *Psychometrika, 10*, 255–282.

Hull, C. L. (1928). *Aptitude testing.* Yonkers-on-Hudson, NY: World Book.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2,* 151–160.

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods, 1,* 98–107.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education & Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Richardson, M. W., & Kuder, F. (1939). The calculation of test reliability coefficients based upon the method of rational equivalence. *Journal of Educational Psychology, 30,* 681–687.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods, 8,* 206–224.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350–353.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques.* New York, NY: Wiley.

Thurstone, L. L. (1931). The reliability and validity of tests. Ann Arbor, MI: Edwards.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin, 54,* 229–249.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22–29.

6

PREDICTING FUTURE PERFORMANCE

Criterion-Related Validation, Regression and Correlation, and Significance Testing

Because identifying the best person for a job is a probabilistic dilemma, the primary objective of employee selection is to reduce error in prediction. Over the last century, psychologists have made considerable gains in reducing prediction error by developing standardized procedures for collecting and combining information on job candidates. Predictions, like measures, should be evaluated by comparing, in an accumulated record, the match of explicitly predicted and actual performance—that is, the likelihood that an assessment-based prediction is true. Traditionally, the relation is determined and evaluated statistically through criterion-related validation.[1]

Criterion-related validation seeks answers to two basic questions. First, what *kind* of relation exists between a predictor and the criterion predicted? This question is answered by a regression line, straight or curved, or an equation. Second, what is the *degree* of relationship? Is there any relationship at all? How strong is it? Is it significant? How accurately can predictions be made? Answers can be based on validity coefficients, usually a Pearson product–moment correlation coefficient, showing the strength of relation specified.

> Criterion-related validation provides evidence for job relatedness. In other words, it is a technique for validating our relational inferences.

Validation as Hypothesis Testing

Criterion-related validation directly tests the hypothesis that criterion Y is a mathematical function of predictor X. It is not the only way to test a predictive hypothesis, but it offers a prototype. It specifies a criterion Y worth predicting and a way

to assess a predictor *X*. It points out that a time lag is inherent in prediction, at least conceptually. The information on which predictions and decisions are made is available in advance, sometimes far in advance, of the time when the criterion information becomes available. A close look at the problems and procedures of criterion-related validation can help in understanding other ways to evaluate bases for personnel decisions.

The first essential requirement for good criterion-related validation is a well-chosen, well-measured criterion. It must be important to the organization and to the decisions to be made, have substantial variance, and be measured reliably and validly. Conceptualization and measurement seem obviously important, but the habit of using whatever criterion lies at hand is so strong that these obvious requirements are often overlooked. Statistical validation of a predictor should not merely assume that the criterion measure is valid. Its psychometric validity should be evaluated using the same principles used to evaluate other measures.

Generalizing from a research or validation sample to an applicant population requires caution. A research sample hardly ever is a representative sample of an applicant population, a fact often overlooked. Only those selected can provide criterion data, so a research sample is usually a biased sample of an applicant population. Researchers should try to specify and match as well as possible the population to which their results should generalize, but they must also acknowledge some imprecision in the match.

Bivariate Regression

Regression refers to the clustering of measures around a central point. A *scatterplot,* graphically showing a point for each pair of *X* and *Y* values, will show a distribution of *Y* values for any given value of *X*. Values of *Y* in each *X* column are distributed about a central point; usually more of them are near that central point than are far away from it. It is convenient to think of the distribution as normal, around the column mean or some other designated central point.

If the two variables are related, the central point in each column changes systematically with changes in the predictor variable. The pattern of change can be shown graphically with a smoothed regression line or curve that describes the relation. The pattern can also be described algebraically with a functional equation, $Y = f(X)$. Many functions are possible, but some may fit the data better than others. One can usually predict that, most of the time, performance of those who score high on the predictor will be better than that of those whose scores are low.

This general statement is based on the usually reasonable assumption that the relation is *positive* and *monotonic.* A relation is positive if higher predictor scores are associated with higher criterion scores. It is monotonic if that statement (or the converse negative statement) is true throughout the predictor score distribution. It

is both positive and monotonic if the central points in the criterion distributions are consistently higher for successively higher values of the predictor—if the smoothed curve always goes up, even if only a little bit in some places.[2] If the functional relationship is both positive and monotonic, more of *X* implies more of *Y* throughout the *X* range.

If an actual criterion level is to be predicted (rather than relative level), the regression pattern—the kind of relationship between predictor and criterion—must either be empirically determined or assumed. Two different kinds of positive, monotonic relationships are shown in Figure 6.1. The equation for the linear (straight line) relationship is $Y = 0.6X + 1.0$. In a linear relationship, the incremental difference in predicted values of *Y* for adjacent values of *X* is constant throughout the range of scores in *X*. In the straight line in Figure 6.1, a 1-point difference in *X* is always matched by a difference of .06 in the predicted value of *Y*.

The curve in Figure 6.1 describes a different kind of relationship. It is a simple freehand curve, drawn to represent a smoothed pattern approximating the mean values of *Y* (one definition of the central points) for narrow intervals of *X*. With such a curve, predicted values of *Y* differ very little for different scores in either the low or the high end of the *X* scale, but they differ a great deal in the mid-range of the *X* scale. An equation could be computed for the curve, but it would have little practical value.

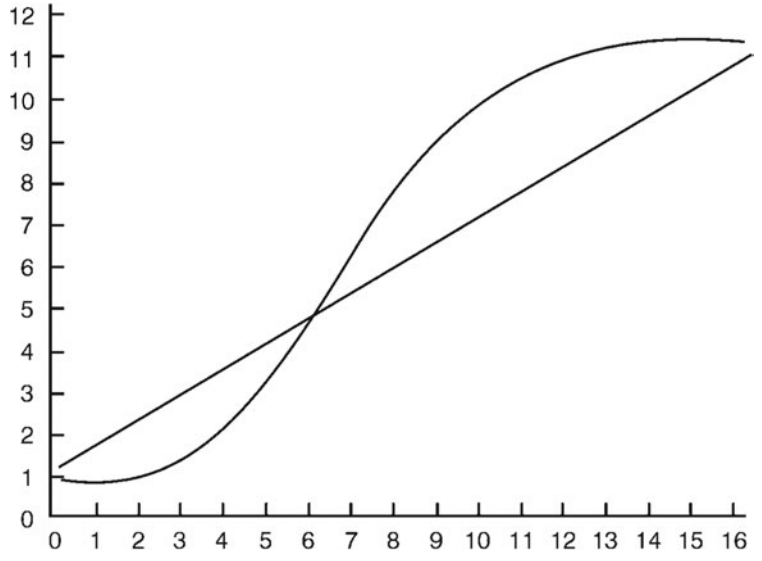

FIGURE 6.1 Straight line or curve for use in predicting a criterion from assessment.

Linear Functions

The general linear regression is $Y = a + bX$. The constant b is the *slope* of the line, the incremental increase in Y with each unit increase in X. a is the Y *intercept,* or the expected value of Y when X equals zero.

Why Linearity Is Typically Assumed

In personnel research, linear regression is typically assumed and rarely questioned, maybe because it is easily computed. Technically, more important justifications for assuming linearity include the following:

1. In computations based on the same data set, the linear regression constants, a and b, and the associated statistics such as correlation coefficients, are more reliable than those in nonlinear equations. (A more reliable statistic is one with less variability from sample to sample.)
2. Linear regression is "robust"; its relevant statistics (a, b, r, etc.) do not seem to depend much on the fit of data to the basic assumptions.[3] To say a statistic is robust may suggest only that it is not particularly sensitive to violations of assumptions.
3. Evidence of nonlinear relations is relatively rare. Hawk (1970) and Coward and Sackett (1990) found them with about chance frequency in studies using the Generalized Aptitude Test Battery (GATB); similar results were found for the relation between conscientiousness and job performance (Robie & Ryan, 1999).
4. Departure from linearity can be statistically significant without being important.
5. Some nonlinear functions can easily be transformed (e.g., with logarithmic transformations) to linear ones.
6. Correlation coefficients based on the linear assumption are required in many statistical analyses following bivariate validation. Multiple regression, factor analysis, meta-analysis, and utility analysis are a few examples of procedures that usually need linear coefficients.

Despite the arguments favoring it, it is unwise to assume linearity automatically, without further thought. Scatterplots should be examined routinely for regression patterns and outliers. A nonlinear pattern may fit better and make more sense. Ghiselli (1964) reported a nonlinear regression that withstood several cross validations. Where a specific form of nonlinear regression is superior to linear regression in repeated replications, there is little reason to use a repeatedly inferior linear regression, especially if the curve makes sense and makes substantially different predictions in the score ranges where decisions are made. This may be especially important for personality tests.

Measures of Correlation

A coefficient of correlation describes how closely two variables are related. It is based on the tightness with which criterion values cluster around the central points that define the regression function. Various kinds of correlation coefficients describe degrees of relation; they may differ on the kinds of relations assumed, on data distributions, or on kinds of measurement scales, but they have important common characteristics.

Basic Concepts in Correlation

Any coefficient of correlation is based on a specified regression pattern. If the pattern does not fit the data very well, but is assumed in computing a coefficient, the coefficient understates the relation. The degree of understatement can range from trivial to dramatic.

If correlation is perfect, the applicants would have identical rank orders on predicted and actual performance, and the scale distances between measures of any pair of people is the same on both scales. Perfect correlation is rare; departures from perfection are expected. The lower the correlation, the greater the prediction error. *Regression functions permit prediction; correlation coefficients permit inferences about the degree of prediction error based on the specified regression function.*

Residuals and Errors of Estimate. A *residual* is the difference between the observed value of Y for an individual case and Y_c, the predicted criterion level for the value of X in that case; Y_c may be found from the regression equation or from a graph of it. If a less than optimal regression pattern is used, the mean and variance of the residuals will be relatively large. When differences in Y are, in fact, related to those in X, the variance of the residuals is necessarily lower than the variance of Y itself. This is what is meant when it is said that X "accounts for" some of the variance in Y.

> It is often useful to think of *variance* conceptually, as well as statistically. Variance reflects differences in scores from person to person. How much of those differences on Y can be accounted for by differences on X? The *residuals* are the differences on Y that are *not* accounted for by X.

A Generalized Definition of Correlation. The basic defining equation for all correlation is as shown:

$$\text{Coeff} = \sqrt{1 - \frac{s^2_{res}}{s^2_y}} \tag{1}$$

where "Coeff" is used in place of a more specifically identified coefficient to emphasize the generality of the equation, s^2_{res} is the variance of the residuals, and s^2_y is the total variance of Y. Most coefficients of correlation can range between 0

and 1.0; for monotonic relations, the range can be from +1.0 to –1.0, depending on whether high scores are associated with good or poor performance. (A negative slope can be changed to positive by the simple expedient of reversing the scale of one of the variables, so this discussion of basics is limited to positive values.) A coefficient of 1.0, then, indicates a perfect relation in which every data point falls directly on the regression line or curve with no residuals at all. The ratio of residual variance to total variance indicates the degree of imperfection in the strength of relation. If s^2_{res} equals s^2_y, that ratio is 1.0 and the coefficient is 0.0.

Coefficients of Determination. If Equation 1 is squared (i.e., the square root is not taken), the result is called the *coefficient of determination*. It estimates the proportion of shared variance in the two variables, typically expressed by saying that the proportion of variance in one of them (usually *Y*) is "accounted for" by the variance in the other. This means common or associated variance, but the usual parlance includes terms like "variance explained by" or "variance accounted for" despite their unwarranted causal implication. Even the term itself, determination, inappropriately implies causation.

Validity coefficients of .30 are not uncommon: The corresponding coefficient of determination for that value would be .09, or 9% common variance. Expert witnesses and attorneys in litigation are fond of intoning in such a case that "less than 10% of the criterion variance is explained by the predictor," slurring over the word variance as if it were unimportant. Note that flipping a coin accounts for 0% of the variance in the criterion. Thus, the 10% should be viewed as the amount of gain in explained variance over and above random selection.

> The ***coefficient of determination*** or r^2 is typically interpreted in terms of how much of the criterion variance can be explained by predictor variance; for example, the amount of the differences from person to person in graduate school performance explained by their differences on the GRE. Although this is true in general, utility (see Chapter 8) is a function of *r*, rather than r^2. Thus, although it can be argued that for a validity coefficient of .30, only 9% of the criterion variance is explained by the predictor, the utility of such an increase can still be substantial.

Variance is an important statistical concept. Variances can be added together; standard deviations cannot be. The standard deviation is the closer description of variability because it is a kind of average of individual differences expressed in the same units as the measurement scale. However, it has limited mathematical usefulness. You cannot add (or subtract, or multiply, or divide) the standard deviation of one measure to the standard deviation of another because standard deviations are square roots of other numbers. It is obvious that 3 + 3 = 6; it is equally obvious that $\sqrt{3} + \sqrt{3}$ is *not* equal to $\sqrt{6}$.

A common variance statement simply is not a useful description of a co-relation; an unsquared correlation coefficient is directly useful. An even better descriptive

statistic is the slope of the regression line; it is more meaningful because it gives the expected change in *Y* associated with a change in *X*. This is only the first of many caveats about bivariate coefficients of correlation and their derivatives. Historically, the validity coefficient was the end product, virtually the only product, of criterion-related validation. Researchers in psychometrics and personnel decisions are increasingly skeptical of a lone correlation coefficient as an index of the value of a predictor.

Third Variables. A second caveat is familiar: Correlation says nothing about causation. It is easy to presume that a variable obtained first somehow produces the second one. To do so is to forget the third variable problem. Both the *X* and the *Y* may be effects of some common third variable or collection of variables. Gulliksen (1950) gave a delightful example. He said that the number of storks' nests built each year in Stockholm correlated .90 with the annual birth rate there! Few people believe that storks bring babies, or vice versa, in Stockholm or elsewhere. If the correlation is reliable, one might speculate about third variables that may explain it, such as economic variation or perhaps the coldness of winters. Other speculation is possible, but the only sure thing is that a causal interpretation of the correlation is wrong.

The Null Hypothesis and Its Rejection. To be useful for prediction, a predictor's correlation with the chosen criterion should be greater than zero—preferably substantially greater, but at least statistically significantly greater. The significance question is discussed later, as this has been the topic of great debate in recent years.

The Product–Moment Coefficient of Correlation

Nearly all statistical computer packages include procedures for computing product–moment coefficients, also known as Pearsonian coefficients. Different programs use slightly different equations, but all are derived from the basic product–moment definition:

$$r_{yx} = \frac{\sum z_x z_y}{n} \tag{2}$$

Statistics such as correlation coefficients and regression lines were once calculated by hand. Today, we have access to a variety of software programs that greatly ease the burden of calculating the desired statistics. Although there are a variety of specialized software programs designed for use with tests, most practitioners and researchers use popular software packages such as SPSS (IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.) or SAS (SAS Institute Inc., Cary, NC: SAS Institute Inc., 2014). More recently, there has been a movement to the use of a package of programs known as *R*. (R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria).

This basic equation looks simple but is too complex for practical purposes. It requires transforming every value of X and Y to z-scores (once called the "moments" of a distribution), multiplying each pair of z-scores, and finding the mean of the products. A useful computational equation uses raw scores:

$$r_{yx} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \tag{3}$$

where r_{yx} is the product–moment coefficient for the regression of Y on X. X and Y are the raw scores, XY is the product of the raw scores for each person, and n is the total number of cases. Several things influence a product–moment correlation coefficient.

Nonlinearity. A product–moment correlation coefficient assumes linear regression. To the degree that this assumption is violated, the coefficient will underestimate the degree of relationship, but where evidence of nonlinearity is questionable or trivial, the linear assumption is still preferred.

Homoscedasticity and Equality of Prediction Error. It also assumes *homoscedasticity,* that is, equal residual variances in different segments of the predictor distribution. If the outline of the scatterplot is approximately an oval, the assumption may not be violated seriously. Serious violations, however, cause r_{yx} to understate the relation seriously. Heteroscedasticity, the opposite of homoscedasticity, or when there are different residual variances at different points of the predictor distribution, can be a more serious problem than usually recognized because it may result in correlation coefficients that markedly understate the value of a predictor. The average correlation may be poor, but if the lowest residual error is in that part of a distribution where the most critical decisions are made, the predictor may be more useful than the coefficient suggests. It may also work the other way. If decisions are to be made at the extremes of the distribution (e.g., if only top candidates are to be accepted), and if residual error at the top scoring levels is great, the predictor may not be useful despite generally high correlation.

Correlated Error. Measurement errors are assumed to be uncorrelated with each other and with the two variables. If the assumption is limited to random errors, violations have little effect on the correlation of reasonably reliable measures. Safeguards against major influences of correlated random error are (a) maximizing reliabilities of both measures, and (b) replicating studies in new samples.

Unreliability. As described in Chapter 5, unreliability, in either variable, reduces correlation. The effect is systematic and, therefore, correctable.

Predictor unreliability is simply a fact in the decision context as well as in research. Criterion unreliability, on the other hand, influences research findings but not individual decisions. Coefficients should, therefore, be corrected only for criterion unreliability. This estimates the population coefficient for the predictor as it is:

$$r_{y\infty x} = \frac{r_{yx}}{\sqrt{r_{yy}}} \tag{4}$$

where $r_{y\infty x}$ is the *expected correlation between a perfectly reliable Y and the fallible predictor X.*

As demonstrated in Chapter 5, the impact of correcting for unreliability can be substantial. Thus, realism dictates two other actions. First, overestimate criterion reliability so that the resulting correction is an underestimate of the population value. Spuriously high corrections "may not only lead one into a fantasy world but may also deflect one's attention from the pressing need of improving the reliability of the measures used" (Pedhazur & Schmelkin, 1991, p. 114). Second, correct only coefficients that are statistically significant. Adjusting possibly zero correlations can be seriously misleading and is a bad practice. It is, fortunately, an uncommon practice, but it happens often enough to warn against it. Professional practice and guidelines also dictate the reporting of both the uncorrected and corrected correlation.

Reduced Variance. If variance on either variable is substantially less in the sample than in the population, the sample coefficient underestimates population validity. Reduced variance is commonly called ***restriction of range,*** associated with truncation of one or both variables.

With a high correlation, the generally elliptical or football-shaped scatterplot in (a) in Figure 6.2 is narrow relative to its length. If the correlation is low, the ellipse is wider, as in (b). If the correlation is zero, the scatterplot is outlined by a circle. Removing either end of the ellipse, as illustrated in (c) reduces variances and makes the remaining portion wider relative to its length, that is, reduces the correlation, as in (d) in Figure 6.2.

The problem cannot be solved by meddling with the scale, such as turning a 5-point rating scale to a 9-point scale. The problem is not the measurement scale but the *disparity in scale variance between the sample and the population.* Anything that truncates the sample distribution reduces variance and, therefore, correlation. Several things can happen to produce a research sample with lower than population variance, and corrections are available for some of them:

1. The predictor distribution can be directly truncated, for example, by accepting all those above a cut score and rejecting those below it. Variances are known both for the unrestricted group (an estimate of variance in the applicant

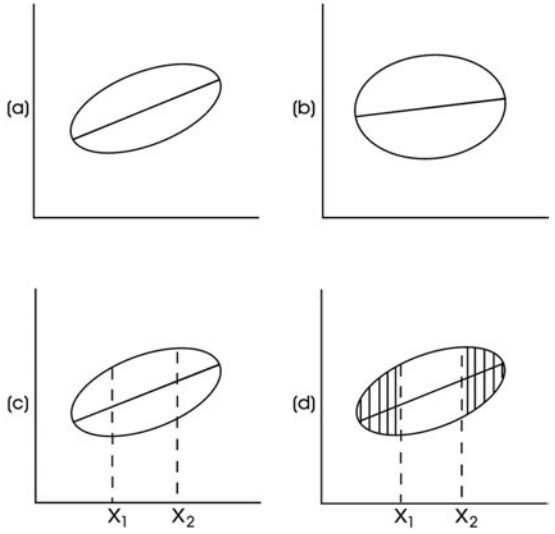

FIGURE 6.2 Elliptical scattergrams showing effect of restriction of range on correlation: (a) scatterplot of high correlation, (b) scatterplot of low correlation, (c) plot of high correlation with X_1 showing where low scorers tend to be lost to the research sample and X_2 showing where high scorers are likely to be lost, and (d) the changed shape of the scattergram when the low and high scoring cases are in fact lost.

population) and the restricted group (those hired), so the corrected correlation coefficient can be obtained by the equation:

$$r_n = \frac{r_o \cdot \frac{s_{xn}}{s_{xo}}}{\sqrt{1 - r_o^2 + r_o^2 \cdot \frac{s_{xn}^2}{s_{xo}^2}}} \tag{5}$$

where r_n is the new estimate of the coefficient for an unrestricted sample, r_0 is the old (obtained) coefficient for the available restricted sample, and s_{xn} and s_{xo} are the predictor standard deviations for the unrestricted and restricted groups, respectively. Sometimes the "old" standard deviation is not known. In this equation and the two that follow, an estimate can be based on available national norms. Sackett and Ostgaard (1994) recommended an estimate 20% lower than national norms.

2. The organization may accept all applicants on probation and then terminate or transfer people below some criterion cut point. Then a test may be given and concurrently validated. The direct restriction is on the criterion, not on

the predictor, but the estimated unrestricted correlation coefficient can be found by reversing the roles of predictor and criterion in Equation 5.

3. Indirect truncation of the predictor occurs if prior selection is based on a correlated third variable. If selection has been based on one test, and another test is being validated, variance on the new test is restricted to the extent that it is correlated with the old one. Guion (1998) provides the formulas for estimating the unrestricted coefficient and for correcting simultaneously for unreliability and range restriction.
4. Sample variance may be lower than population variance just by chance. But one would not know, and no correction is available.
5. Unknown factors may have reduced variance indirectly. Again, no correction exists.

Range restriction is *direct* when people are selected on the basis of scores on the test being validated. It is *indirect* when people are selected on the basis of some other predictor that happens to be correlated with the test being validated.

Correction equations can be used in cases of reduced variance even with no clear point of truncation. Instead of an explicit cut score, for example, there may be a region—a score interval with fuzzy boundaries—below which no one was hired, above which most applicants were hired, and within which decisions were mixed.

Correcting for restriction of range is complex, controversial, and requires a great deal of knowledge of the potential applicant pools and assumed populations. For an excellent introduction to the topic of corrections for restriction of range, including multivariate range restriction, we refer the reader to an overview provided by Sackett and Yang (2000; see also, Beatty, Barratt, Berry, & Sackett, 2014; Hunter, Schmidt, & Le, 2006).

Distributions. Product–moment coefficients require no assumption about distributions, but some interpretations assume an underlying normal bivariate surface. Extreme skewness in one variable but not the other produces nonlinearity and consequent correlation reduction. In fact, any time the two distributions differ markedly in shape, the potential range of obtained correlations is markedly reduced.

Group Heterogeneity. A large sample sometimes seems like a Holy Grail people will do anything to find, such as combining small, disparate samples. Samples may then include groups of people that differ in systematic ways. Combining them

may hide important differentiating characteristics. Subgroups in the overall sample may have different means on one or both variables or different correlations.

Questionable Data Points. Plotting data sometimes shows one or more ***outliers***. An outlier is "an unusual, atypical data point—one that stands out from the rest of the data [and] may lead to serious distortion of results" (Pedhazur & Schmelkin, 1991, p. 398). An outlier can reduce correlation if it is included with the mass of data; in a small total sample, it could even turn an apparent positive relationship to a negative one. Some outliers, on the other hand, would inflate the correlation.

A Summary Caveat. We have identified several things that can influence or distort a product–moment correlation coefficient. Sometimes the direction of error is predictable, but some influences may lead to unknown or unknowable error. Some with knowable effects can be corrected, but many of them are like incurable aches and pains: You simply have to live with them. Living with them, however, should induce caution. One should not place undue faith in a single bivariate validity coefficient. It can offer some evidence—even good evidence—of validity, but potential distortions should be considered in evaluating that evidence. One may need to gather new data, either through replications or studies of possible explanations.

Statistical Significance

Research reports typically say something like, "the correlation was not statistically significant," or "it was significant at the 5% level of confidence." These terms refer to the probability that the reported coefficient differs from zero only by chance. If it differs more than expected by chance, the "null hypothesis" of no relationship is rejected. What researchers know and how they behave are not always the same. Too often they act as if, having rejected the null hypothesis, they can virtually equate the sample correlation coefficient with the population value, or as if mere rejection is enough to assure that the population correlation is *usefully* nonzero. Neither is so.

The Logic of Significance Testing. Statistical validation begins in a sample where both predictor and criterion data are known. Suppose $r = .20$. This is not a very strong relation, but it can be useful. Can a similar relation be expected in a later sample where decisions have to be made without prior knowledge of criterion performance?

Part of the answer depends on the quality of the research. We cannot have confidence in the generalizability of poorly designed or conducted research where research subjects are inappropriately chosen, data collection is haphazard and inconsistent, criteria suffer contaminations, or data recording or analysis is careless. However, any sample statistic is subject to error, no matter how carefully the research was

conducted. Some error may be due to idiosyncratic characteristics of people in that sample. Part of the answer may lie in inevitable violation of statistical assumptions. Correlation analysis assumes that measurement errors are not correlated. In any given sample, however, the errors will, in fact, have *some* nonzero correlation, even if small. Part of the answer may lie in sampling error. The smaller the sample size, in absolute number or relative to the population, the greater the likely error.

Clearly, the unimpressive but potentially useful coefficient of .20 is to some degree in error, even if negligibly. Different samples from a common population would provide a distribution of different coefficients; the mean of a big enough set of them would match the population correlation. Can the population, and future samples from it, be counted on to give coefficients of about the same size as the one at hand? That is a useful question, but it is not the question answered by significance testing. Significance testing goes at it in a reverse process; it tests the null hypothesis that the correlation coefficient in the population is precisely zero. Now, rejection of the null hypothesis does not imply that the sample coefficient is a good estimate of the population coefficient, and failure to reject the null hypothesis does not mean that it is true. Literally, the null hypothesis "is *always* false in the real world" (Cohen, 1990, p. 1308). Specifically, it estimates the probability *(p)* that, if indeed the population correlation is zero, a sample would capitalize enough on error to provide a correlation as large or larger than that obtained, just by chance. Significance testing asks not what that probability is, but only whether it is lower than some prestated level. It answers with a yes or no dichotomy, not with a probability level.

As with many other topics, the debate over statistical testing is complex, involving practical, legal, and philosophical issues. What is the alternative? Finding an acceptable alternative is the major issue. The calculation of confidence intervals does not avoid the problem of significance testing and introduces additional issues. Bayesian approaches are offered as an alternative, but Bayesian statistics lead to other controversies. In 2015, the journal *Basic and Applied Social Psychology* banned the use of traditional hypothesis testing (Trafimow & Marks, 2015). Perhaps surprisingly, the journal took a less than positive view of confidence intervals and Bayesian approaches. The proffered solution was large sample sizes; however, the sample sizes that would be required to eliminate the needs for statistical inference are not feasible for most real-world validation studies.

Type I and Type II Errors and Statistical Power. In strict significance testing, a researcher either rejects or fails to reject the null hypothesis. If it is true and the researcher does not reject it, or if it is false and the researcher does reject it, the choice is correct. The choice is erroneous if the null hypothesis is true but is rejected, or if it is false but not rejected. These two types of errors are known as *Type I* and *Type II* errors, respectively. The chosen level of confidence is called *alpha,* α, the probability associated with Type I errors. The lower the α probability, the lower the probability of a Type I error.

The lower the likelihood of Type I error, the greater the likelihood of Type II error. Which is the more serious error can be determined only in the full context of a particular situation. As the probability of Type I error increases, so does the probability of hiring people on the basis of an invalid assessment. As the probability of Type II error increases, so does the probability that a valid assessment procedure will be discarded.

Statistical power "is the probability that a statistical test will lead to the rejection of the null hypothesis" (Cohen, 1977, p. 4). Power is a function of three things: (1) the size of the sample used, (2) the *effect size* (e.g., correlation) in the population, and (3) the alpha level chosen. A judgment of significance, then, is made more likely by increasing sample size, by working with intelligently developed predictive hypotheses that are very likely to result in substantial correlations, or by relaxing α. Some ambivalence is justified; we like to reduce error, but we do not ordinarily like to lose power. The complement of power (1–power) is the *beta* probability, β, the probability of Type II error—that is, the failure to reject a false null hypothesis.

Published studies often report "*p*" values, such as $p < .05$. This means that there is less than 5% (the *alpha* level chosen) chance of a Type I error. *Power* is the probability of rejecting the null. One minus power (the *beta* level) is the probability of a Type II error.

Concepts of power and Type II error received little attention in early personnel research. Usually, failure to reach traditional levels of significance (i.e., .05) was not seen as a serious problem if the results "approached" it, the sample was small, and the correlation was fairly large. If the predictor were badly needed, an "almost significant" finding was likely to be used in decision making. It was not good science, and it was certainly not orthodox, but it might have made good business sense.

The advent of litigation under the Civil Rights Act of 1964 changed matters. Validity under the *Uniform Guidelines* became virtually synonymous with significance at the .05 level, and lack of a statistically significant validity coefficient was reason enough to abandon a predictor, regardless of other lines of evidence. Issues of statistical power became important; with insufficient power, one could lose the use of a good predictor. Type II error took on importance not earlier recognized.

A Comment on Statistical Prediction

In regression analysis, the predicted value is a specific point on the criterion scale, a central point in the *Y* distribution for the *X* value. More accurately, the prediction is that on average, people with a certain score will perform at the predicted

criterion level. A range based on the standard error of estimate can be specified within which the criterion value for an individual may be expected at a given probability. Identifying such a range acknowledges that most people are not going to perform at that precise point. Most researchers know these things, but in their statistical zeal, they tend to forget them. Together, these specifications are saying that the predictor variable itself leads us to expect a certain criterion performance, but that chance or other things may intervene to lead any given person to perform at a better or poorer level.

Nearly all statistical analyses are based on assumptions that (a) are rarely if ever satisfied in real data, (b) generally can be violated noticeably without seriously affecting results, but (c) can be violated in any single situation with serious effects on results and their interpretations. One such assumption is the assumption of a normal distribution. There is no such thing in real data. Micceri (1989) examined 440 large sample distributions of test data and other distributions gleaned from published articles or reports of various kinds. *All* were in some respect nonnormal. He concluded that the normal curve, like the unicorn, is an improbable creature.

If the normal curve is improbable for one variable by itself, a normal bivariate surface is more so. To be sure, statistical analyses based on the assumption of the normal bivariate surface have, on average, been useful in analyzing real data. That fact is not enough, however, to justify the blithe assumption that violations of assumptions never matter, that a prediction has not been affected by them, or, indeed, that the prediction has not been affected by other considerations not in the equation. Statistical prediction, as surely as predictions without a statistical basis, is subject to informed professional judgment.

Discussion Topics

1. Discuss the problems caused by restriction of range. Give examples of the problem as well as steps you could take to correct it.
2. Come up with examples of predictors that would have nonlinear relations with criteria. Be sure to consider criteria other than job performance ratings (e.g., turnover, sales).
3. What size samples do you believe would be required to eliminate the need for some type of statistical inferences such as confidence intervals or null hypothesis testing? Are those sample sizes realistic or attainable in a typical validation study performed by an organization?

Notes

1 Occasionally, someone will abbreviate the term criterion-related validity and speak or even write about "criterion validity." This should be avoided; logically, criterion validity refers to the validity of a criterion, a psychometric evaluation of the criterion measure, *not* to the statistical concept of the degree of relationship between a criterion and another variable (or collection of variables) that predicts it.

2 The emphasis on the smoothed curve is because random variations occur in the pattern of column means as central points; literally connecting them ordinarily yields a jagged pattern.
3 The assumptions are, at least for linear correlation, linearity of regression and homoscedasticity, meaning equal *Y* variances in the different values of *X*. Homoscedasticity (and its opposite, heteroscedasticity) will be defined more fully later in the chapter.

References

Beatty, A., Barratt, C. L., Berry, C. M., & Sackett, P. R. (2014), Testing the generalizability of indirect range restriction corrections. *Journal of Applied Psychology, 99,* 587–598.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York, NY: Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Coward, W. M., & Sackett, P. R. (1990). Linearity of ability-performance relationships: A reconfirmation. *Journal of Applied Psychology, 75,* 297–300.

Ghiselli, E. E. (1964). Dr. Ghiselli comments on Dr. Tupes' note. *Personnel Psychology, 17,* 61–63.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gulliksen, H. (1950). *Theory of mental tests.* New York, NY: Wiley.

Hawk, J. A. (1970). Linearity of criterion-GATB aptitude relationships. *Measurement and Evaluation in Guidance, 2,* 249–251.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*, 594–612.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Robie, C., & Ryan, A. M. (1999). Effects of nonlinearity and heteroscedasticity on the validity of conscientiousness in predicting overall job performance. *International Journal of Selection and Assessment, 7,* 157–169.

Sackett, P. R., & Ostgaard, D. J. (1994). Job-specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology, 79,* 680–684.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: an expanded typology. *Journal of Applied Psychology, 85,* 112–118.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37,* 1–2.

7

USING MULTIVARIATE STATISTICS

Multiple Regression, Cutoff and Multiple Cutoff Models, and Validity Generalization

Most jobs are complex. Effective performance requires several traits, not just one. Practitioners rarely make a decision from just one test or one score. Practitioners usually rely upon a selection battery, which is composed of multiple measures. Often, what appears to be a single test may actually have multiple dimensions generating several scores. In order to minimize costs, multiple tests may be given over time, with a segment of the applicants eliminated at each step.

As performance comprises both abilities and motivation, multiple predictors will be needed to account for it. Predictions of performance require combining or sequencing predictors in some way. The method chosen should be based on both statistical considerations and professional judgment.

Compensatory Prediction Models

Scores on predictors can be combined in any of several models. In a linear, additive model—the most common—scores are summed to form a composite, maybe with different weights for different variables. The several predictors are assumed to be linearly related in the composite, which is linearly related to the criterion. Summing scores is *compensatory*; a person's strength in one trait may compensate for relative weakness in another. For example, lower ability might be compensated for with more motivation. Consider Table 7.1. Candidate A has equal strength in all three traits. Candidate B is weaker than A in Trait 1 but may have enough added strength in Trait 3 to compensate. Candidate C is extremely deficient in Trait 2 but strengths in the other two may compensate. All three form the same composite score by adding the three component scores. If one trait is more important than the others, its scores get more weight (i.e., multiplied by a larger value) than the

TABLE 7.1 Composite Scores for Three Traits for Three Hypothetical Candidates

Candidate	*Trait*			*Sum*
	1	*2*	*3*	
Without different weights				
A	10	10	10	30
B	8	10	12	30
C	15	0	15	30
With different weights				
Weights	2	3	1	
A	20	30	10	60
B	16	30	12	58
C	30	0	15	45

others, as in the lower half of Table 7.1. If Trait 2 is considered so important that Candidate C's deficiency in it is unacceptable despite other scores, an additive model is inappropriate.

Regression Equations

Multiple regression analysis finds optimal weights for the several predictors, multipliers that form a composite having the best possible correlation with the criterion in the sample studied. The composite of weighted predictors estimates the expected criterion value for each person. Those optimal weights are optimal *only* in the research sample. In a different sample, different optimal weights would be found. Ordinarily, the weights computed in one representative sample will approximate the optimal weights in most other samples.

> A weighting method should be based on *rational* and *theoretical* grounds rather than on computations alone. Often, psychometric and statistical assumptions are not met in applied settings; it is not wise to take excessive pride in an impressive weighting system. It is wise to see if effective weights make sense.

Weights may be computed for either standardized or unstandardized scores. In conventional notation, the letter beta, β, stands for standardized weights used with standard *z*-scores, and the letter *b* refers to unstandardized weights used with raw scores or deviation scores. Both kinds of weights depend on correlations with the criterion and other predictors. Unstandardized weights also depend on relative variances.

The case of two predictors illustrates more general principles. In raw score form (with simple subscripts), the 2-variable regression equation is as shown:

$$Y = a + b_1X_1 + b_2X_2 \tag{1}$$

where a = the Y intercept and b = regression coefficients for multiplying predictors as identified by subscripts. If the composite score C is the sum of b_1X_1 and b_2X_2, the equation can written in the familiar $Y = a + bC$ form, where $a = Y$ intercept and b = the slope of the regression of Y on the composite score C.

Computing Regression Coefficients

Regression coefficients can be computed directly from the relevant correlation coefficients and standard deviations:

$$b_1 = [(r_{yx1} - r_{yx2}r_{x_1x_2}) / (1 - r^2_{x_1x_2})] \cdot (s_y / s_{x_1})$$

and (2)

$$b_2 = [(r_{yx2} - r_{yx1}r_{x_1x_2}) / (1 - r^2_{x_1x_2})] \cdot (s_y / s_{x_2})$$

where the values of r = a correlation coefficient, specified by subscripts, and s = a standard deviation of the criterion or of a predictor.

If $r_{x_1x_2} = 0$, the regression weight of either predictor is its validity coefficient reduced by the ratio of the criterion standard deviation to the predictor standard deviation. If raw score distributions are standardized, all standard deviations are 1.0, so standardized regression weights equal the validity coefficients. If $r_{x_1x_2} > 0$, β weights are lower than the validity coefficient. If the two validity coefficients differ, the predictor with the higher validity has the greater weight, and the disparity increases as the intercorrelation increases. If $r_{x_1x_2} = 1.0$, one predictor is enough; the other adds nothing.

Multiple Correlation

Sometimes the size of the multiple coefficient of correlation, R, is of more interest than the regression equation. It is an index of the strength of the relation of the predictor composite and the criterion. It can be computed as a bivariate r, with the optimal composites as X, or from existing correlation coefficients. For the two-predictor case,

$$R^2_{y \cdot x_1x_2} = \frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r^2_{x_1x_2}} \tag{3}$$

where $R_{y \cdot x_1 x_2}$ equals the coefficient of multiple correlation for two X variables predicting Y; the various product–moment correlations are defined by the subscripts. The equation shows general principles of multiple correlation: (a) the validity of the composite is proportional to the validities of the components, and (b) the validity of the composite is inversely proportional to the intercorrelations among components.

The *multiple correlation, R,* will equal the sum of the individual correlations only when the predictors are uncorrelated (highly unlikely). If the predictors are correlated, then $R < r_1 + r_2$.

Suppressor and Moderator Variables

Suppressors. By those principles, each test in a well-developed battery is a valid predictor of the chosen criterion and has low correlations with other variables. A valid predictor may contain an invalid, contaminating variance component. A variable that does not predict the criterion but is correlated with the contamination may actually improve prediction. To see how this works, look again at Equation 3. If $r_{yx_2} = 0$, but if both of the other two correlations are not zero, then the numerator of that equation becomes simply r_{yx_1} (the other two terms being zero). The denominator is less than 1.0 (because $r_{x_1x_2}$ is *not* zero); therefore, $R_{y \cdot x_1 x_2}$ is *greater* than the validity of the one valid predictor alone. The reason is that variable X_2 removes from the composite (suppresses) the unwanted variance in X_1 not associated with the criterion. In a regression equation, it has a negative weight.

Consider, for example, a case in which a paper-and-pencil test of law enforcement knowledge is used to hire security guards. This test is valid, but it requires a relatively high level of reading ability to complete—a level of ability not necessary for a security job. A reading ability test would correlate with the law enforcement knowledge test, but not with ability to perform the security job. The reading ability test would, therefore, receive a negative weight in the regression equation. Although it may slightly improve prediction to add a reading test to the security guard selection system, it would be hard to explain to the company (and the test taker) why a candidate is rejected for scoring too high on it!

Moderators. Moderator variables influence the relation between other variables; they are correlated with correlation. Frederiksen and Melville (1954) found prediction of academic performance from interests better for noncompulsive students than for those classed as compulsive. Although it is easier to think about validities in subgroups, validity should change systematically and continuously as the level of the moderating variable changes. A regression equation for one predictor and one moderator has the following form:

$$Y = a + b_1x + b_2z + b_3xz \tag{4}$$

where Y = the criterion, x = the predictor, z = the moderator variable, and xz = the product of x and z scores, the interaction or moderator term, weighted in the composite by b_3. In Equation 4, variables X and Z are expressed in deviation score units (e.g., $x = X—M_x$) with means of zero, unlike the linear Equation 1. Moderated regression is an additive (compensatory) model, but it is not linear because of the multiplicative term. A significant interaction term says that, for every value of z, there is a different slope of the regression of y on x, even though the difference may be small and gradual.

Like suppressor variables, examples of moderator effects for personnel decisions are rarely reported and rarely replicated. The initial surge of enthusiasm led to sweeping searches for moderators in whatever data pool was available; such reliance on exploration and serendipity was not often rewarded. Enthusiasm for demographic moderators (as solutions to fairness problems) was no more fruitful. Searches for moderators in the validity generalization paradigm have turned up only a few. As a result, many selection specialists have given up on moderators.

Such pessimism is unwarranted. Some failures to find moderators are methodological (e.g., use of raw scores). Many more are due to inadequate logic. Research agenda should abandon serendipity. Moderators seem more likely to be found after serious thinking, hypothesizing, and theory formation than after searches among variables for which there is no useful rationale. For example, Witt and Ferris (2003) hypothesized that, for jobs where interpersonal effectiveness is an important part of performance, conscientiousness would predict performance only for socially skilled people. Indeed, the authors found evidence across four studies that social skill moderated the conscientiousness–performance relation. The field of Industrial and Organizational psychology is big. Many activities and concepts expected to influence a variety of work outcomes have been studied and promoted extensively. A systematic consideration of such variables surely should be a rich source of reasoned moderator hypotheses.

Other Additive Composites

Multiple regression is but one way to form a composite, and reasons for forming composites are not limited to criterion prediction. Reasons might include, for example, weighting predictors to promote an organizational policy. Consider an organization that is interested in moving from a competitive to cooperative work environment; predictors that worked under the old competitive system might not work for predicting teamwork and cooperation. A weighting method should be based on rational, theoretical grounds rather than on computations alone. One important rational principle is simplicity.

"Unit" weighting means simply adding scores or standard scores, literally multiplying by 1.0, as in the top of Table 7.1. Dawes and Corrigan (1974) insisted, and demonstrated, that use of more complex models offers no more than slight

improvement over simple weights, whether equal or differential, in accounting for criterion variance; their finding held even with randomly chosen weights. Subsequent research has supported the finding (Bobko, Roth, & Buster, 2007). However, regression weights may predict better than unit weights if (a) patterns of intercorrelations among the predictor variables differ, (b) the regression-based multiple *R* is high, (c) different predictors have substantially different weights, and (d) the ratio of respondents to variables is large. In general, carefully computing weights to several decimal places may give only the appearance of precision; simpler nominal weights may do as well or better if variables are carefully selected, are positively correlated with each other and with the criterion, and do not differ greatly in validities or reliabilities. Overall, weighting every predictor equally is often an attractive option (Bobko et al., 2007), although it may be difficult to explain to test takers and other stakeholders in the assessment process that every predictor is equally important.

Noncompensatory Models Based on Cutoffs

For a moment, we will return to the situation where we have only one predictor. In this situation, selection decisions can be made based on either ranking or a cutoff. Consider a situation where we are hiring for 10 clerical openings. We give a test to a large group of applicants, but have to select only 10 for the job. One simple approach is to take the 10 highest scores, which is a decision that is very consistent with the idea of a linear relation between the predictors and the criterion. Throughout this book, we more often than not assume a linear relation between the test and job performance, accompanied by the selection of applicants in a top-down fashion or through ranking. In a set of candidates, those with higher scores at any level are preferred over those with lower scores.

There is, however, another choice; we could establish a cutoff and select those who score above that cutting point. Hiring the best of a poorly qualified lot is poor management. In a test of prerequisite job knowledge, if every examinee should have a very high score, it is not helpful to say that someone with a very low score still has more job knowledge than a lot of other people and should, therefore, be chosen. For that, and a number of other reasons, we may choose to use a cutoff score rather than do top-down ranking.

For example, we might determine that to be considered a qualified candidate, the applicant for the clerical job must score about 70%. Everyone who scores above 70% is considered qualified and put into a hiring pool. If we end up with only 5 people passing the test, then we know we have to improve our recruiting efforts and also attempt to attract better applicants. Of course, if we end up with 30 people passing the test and need only 10 hires, we still need to choose among those who are designated as qualified. One approach would be to select at random from the 30 qualified applicants. A second approach would be to allow a manager or supervisor to choose among the people scoring above the cutoff.

A misguided tendency in personnel selection, and also in education, has been to simply accept a passing cutoff of 70%. Psychometric evidence is necessary in order to support and document the use of an appropriate cutoff. Two broad classes of approaches have been used for establishing a cutoff for a psychological measure. The first involves using the published, normative data available on a test and setting the cutoff to correspond to some logical percentile value. The second approach relies upon local information involving either empirical studies or judgments made by SMEs.

Norm-Referenced and Domain-Referenced Cutoffs

Test scores are often ***norm-referenced***, that is, interpreted relative to the scores of people in a comparison (norm) group. Whether a score is considered good or poor depends on the distribution of scores in the norm group. Figure 7.1 shows percentile ranks associated with raw scores in three hypothetical distributions.

	Percentile Rank in		
Raw Score	Group A	Group B	Group C
24			
23	99.9		
22	99.4	99.9	
21	97.7	99.6	
20	94.3	98.5	
19	88.4	96.4	
18	79.9	93.0	
17	70.2	88.6	
16	60.0	83.1	
15	50.0	76.9	
14	40.8	69.9	
13	32.3	62.1	99.9
12	24.6	54.0	99.2
11	18.1	45.5	97.2
10	12.7	36.7	94.1
9	8.5	28.2	89.3
8	5.4	23.9	82.6
7	3.1	16.7	73.9
6	1.7	10.6	63.2
5	.8	5.7	51.0
4	.3	2.5	37.5
3		.7	23.7
2		.2	11.9
1			4.0
0			.4

FIGURE 7.1 Differences in interpretations of a given test score with different norm groups; a raw score of 12 is in the bottom quarter of the distribution in Group A, slightly above average in Group B, and outstanding in Group C.

An examinee with a score of 12 has answered half of the items correctly. It is a magnificent score compared with those in Group C, better than more than 99% of the scores in that group. Compared with those in Group B, it is about average, neither very good nor very bad. It is not good at all—in the bottom quarter—in Group A, the group with the best set of scores.

Now, imagine a situation where you are working as a consultant for a sales organization. You determine that to be successful, an applicant for a sales position must score at least at the 50th percentile for Group A, or achieve an average score on our ability test. Inspection of Figure 7.1 reveals that a cutoff should be set at a score of 15 in order for the applicant to be considered suitable.

In truth, however, norm tables are rarely consulted in employment testing. First, except for some widely available and frequently used assessments, occupation specific norm tables are unlikely to be available. Second, as discussed in Chapter 8, expectancy charts are more likely to be useful in setting cutoffs; expectancy tables provide the probability of success for individuals at various score levels.

An alternative to normative interpretations was originally called criterion-referenced interpretation. In it, scores are interpreted relative to the content domain being tested; we prefer domain-referenced interpretation.[1] Under either term, the basic idea is that a domain of accomplishments is identified and defined. It should be defined clearly enough that people, even those who disagree about the domain, generally can agree on whether a specified fact or achievement is in or outside of it. Measures of the domain should fit the definition, and scores should be explicitly interpretable in terms of it. In ***domain-referenced*** testing, the domain, not a point in a score distribution, is the criterion for referencing or interpreting an obtained score. If we could develop tests that were truly domain referenced, then we could set cutoffs easily by simply determining the types or level of behavior required. An example might be offered by the recent cooking contests that are now prevalent on cable television. Our test might consist of requiring an applicant to cook an appetizer, entrée, and dessert in less than an hour using provided ingredients. If an applicant can use all the ingredients, do it in the time allowed, and produce a tasty product, we would assign that person a passing grade on our test.

An example of a *norm-referenced* test might be a test of mechanical aptitude. Those with higher scores would be seen as better candidates for mechanical careers than those with lower scores. An example of a *domain-referenced* test might be a test designed to certify a mechanic as competent to work on a certain automaker's vehicles—there is a well-defined domain of knowledge that the person must master. Note, this approach should be differentiated from a work sample, as such tests more often than not are interpreted in terms of the score achieved, rather than on an interpretation of test performance in terms of the underlying domain.

In theory, any test can be used for either norm- or domain-referenced interpretation. In practice, tests may be developed differently for these differing purposes. Clarity of test purpose, always important, is especially so in domain-referenced testing. It is not enough to say that a test's purpose is to measure knowledge of computer repair procedures. Defining "knowledge of repair procedures" requires clarity about component content areas; components should be assigned relative weights, and the kinds of items to be used for each component should be specified. Unfortunately, developing domain-referenced tests turns out to be much more difficult than one would think, and, therefore, this approach is rarely used in selection testing.

Cutoffs Based on Local Information

Instead of using national or published norms, cutoffs can be based on data obtained from the test construction or validation process. In the case where empirical validation data are available, a number of statistical techniques can be applied to determining an appropriate score corresponding to minimally acceptable job performance.

Contrasting Groups. One simple procedure involves identifying two contrasting groups of high performers and low performers. The test scores of the high and low performers are then recorded, and the distribution of scores is then inspected in order to identify natural points of discrimination between the two groups. One technique of doing this that relies upon visual inspection is to graph the histograms for the high and low performing groups and then look for the point where the two graphs intersect. This point of intersection can then be used as a cutting score.

Predicted Yield Method. Distributions of candidate qualifications fluctuate from week to week. Availability of openings also varies. The two may not coincide; the best applicants may present themselves when there are no immediate openings. One large company in a small town had such a problem in hiring skilled clerical workers. The best applicants graduated from high school and community colleges in the spring and usually moved away. The solution was to hire good applicants when available, place them in clerical pools, and promote or transfer employees as positions opened.

The plan required fairly accurate prediction of the number of openings likely over the coming year and knowledge of the probable distributions of qualifications. A cut score could then be found to permit hiring enough people at graduation to meet the organization's needs for that year. This kind of cut score is not a costly dichotomization; it is based on a top-down policy. In effect, it is an answer to, "If all these people were available when we wanted them, and if we hired from the top-down as positions opened up, how far down the distribution would we go?"

Regression-Based Methods. Cut scores can be established by working *backward* through a regression equation, thus:

$$\textit{Desired Performance Score} = a + b\text{X} \tag{5}$$

A manager can specify a performance level that is desired of all new hires. The equation is then solved for X, the score needed on the predictor to achieve the desired performance level.

Judgmental Methods. In practice, it is more common to see assessment professionals rely upon judgmental methods based on SME judgment. Probably the most commonly used approach to setting cutoffs is the Angoff method (Angoff, 1971). The attractiveness of the Angoff method lies in its simplicity and the ease with which it can be applied to different types of tests. In using the Angoff method, SMEs are asked to read each item on the test. Then, for each item on the test, the SME indicates the proportion of minimally qualified test takers who will respond to the item correctly. The responses are then averaged across SMEs to obtain the probability that a minimally qualified test taker will answer each item correctly. Those probabilities are then averaged across all items, which leads to a proportion or probability that represents the percentage score that will be obtained by a minimally qualified applicant. So if the average of judgments across 100 items is a proportion of .65, then we can infer that a minimally qualified applicant will obtain a percentage score of 65%. Therefore, 65% is the appropriate cutoff for our test. Although the Angoff method is not without its critics, the popularity of this judgmental method of setting cutoffs can be traced to it being easy to understand, quick to complete, reliable, and having decent validity.

Multiple Cutoff Methods

A multiple cut (multiple hurdles) approach uses a cut score for each of two or more tests. An applicant scoring below the cut score on any of them is rejected; each test is a "hurdle" to clear. Two situations may justify the method: (1) if each trait is so vital to performance that other personal strengths cannot compensate for weaknesses in them (e.g., good eyesight may be required of an airline pilot), or (2) if their variance is too low to yield significant correlation.

A truly ***noncompensatory*** trait—one so vital to performance that no other strength can compensate—is unlikely. Psychologically, people learn to live with deficiencies and make up for them. Statistically, the idea suggests a discontinuous function with no functional relationship on either side of the point of discontinuity. We know of no such finding. Even so, some researchers have found nonadditive, noncompensatory prediction models useful.

Generally, however, objections to cut scores in bivariate prediction apply even more to multiple predictors, where even very low cut scores can result in rejecting too many candidates. More hurdles mean more rejections. Many of those passing all of the hurdles will do so with scores too low to suggest any genuinely useful qualifications at all. Cut scores high enough to ensure people qualified on each trait may find no one qualified on all of them. A multiple cutoff approach is justified only when predictors are perfectly reliable. If practical considerations demand it, it can be modified by a partially compensatory model. Selection effects using

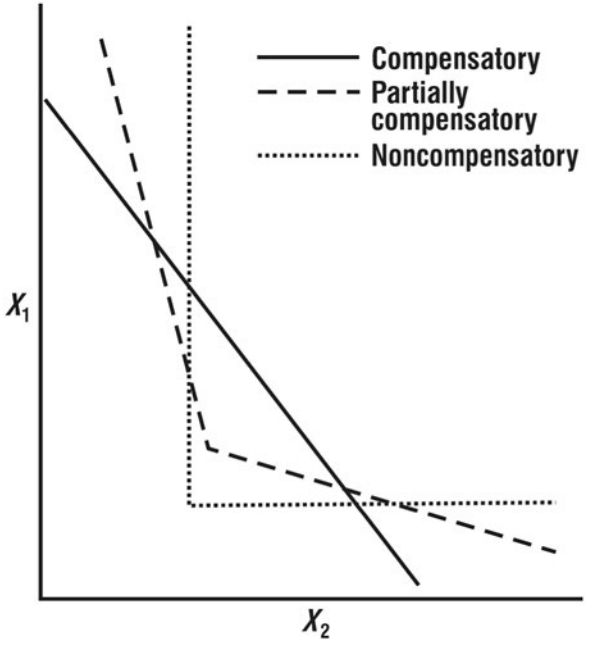

FIGURE 7.2 Areas of decision within two-predictor scattergrams for compensatory, partially compensatory, and noncompensatory decision models; in each case, the area above or to the right of the line is the area within which selection is the appropriate decision and those below and to the left of the line are rejected. From Guion (1965).

compensatory, noncompensatory, and partially compensatory models are shown in Figure 7.2; those selected score above (or to the right of) the lines applicable to both variables.

In a *sequential hurdles* approach, those who "pass" one or more preliminary steps are assessed later on other characteristics. The early cut scores are often intended to reduce the size of the group to be assessed by costlier methods. There may not be a fixed cut score; a fixed number of candidates more qualified than others may move to the next stage. Fixed cut scores transform scores to a dichotomy, 1 *(pass)* or 0 *(fail)*. This is important in validating later assessments in a candidate population limited to those passing the earlier hurdles.

Cut Scores Caveats

A cut score effectively dichotomizes a score distribution, loses information, and, if not near the mean, substantially reduces validity. Dichotomization is rarely recommended. Some situations, however, justify and even require a cut score:

1. Civil service jurisdictions commonly give a test to masses of candidates at one time and do not test again for a year or more. Candidates are listed in

an "eligibility list" ordered from those with the highest score to a minimum score. The minimum is a cut score below which examinees may not be listed and no one will be hired.

2. Licenses or certification are intended to certify a useful level of knowledge or skill, a degree of competence presumed to protect the public against incompetence. Certification is not limited to governments. Private organizations, including trade associations, may elect to certify the competence or knowledge of sales people, technical advisors, repairers, or others whose work affects customers or the public.
3. Hiring may be cyclical. For example, if there is a policy of hiring new graduates from high schools or colleges to work as trainees, most hiring will be done at about graduation time in the spring. Openings may arise at any time through the year. By forecasting the number of openings likely to be needed before the next hiring phase, and with a fairly accurate notion of the score distribution, one can establish a cut score that will provide the necessary number of trainees who can then be assigned to more permanent positions that become available.
4. Assessment may be sequential; an assessment may be scored on a pass–fail dichotomy to decide who gets to the next step. Where many candidates compete for one or a few positions, preliminary screening may be used for all candidates, saving complete assessments (e.g., assessment centers or complex simulations) for the most promising ones. For some jobs, the preliminary assessment may look for intrinsically disqualifying considerations (e.g., poor spelling among proofreader candidates).

Cut scores are too often established merely for convenience. With them, managers getting a candidate's test score need make no judgment more taxing than whether it exceeds the cut point—and no HR person need try to explain more valid decision processes to the managers. This bad habit would not be worth mentioning were it not so common, so unnecessary, and so costly in terms of assessment usefulness.

Replication and Cross Validation

A simple additive combination can give a large validity coefficient in one sample that is never again repeated in another. Results of validation, especially a multivariate one, need to be repeated—replicated—in a new sample when feasible.

Multiple regression requires *cross validation*. Loose use of language sometimes treats cross validation and replication as interchangeable, but they are different. Cross validation applies multiple regression weights obtained in one sample to data obtained in a different one to see whether the multiple *R* found in the first sample holds up in a second or whether it was inflated by sample-specific error. Replication refers to a repetition of an original study, with or without some systematic change in measures or procedures, to see if independent results are similar.

Cross validation is required in multiple regression studies because the composite-forming regression equation developed in one sample has the highest possible correlation with the criterion in that specific research sample. In another, independent sample from the same population, using the same equation, the new correlation is almost always lower. *Shrinkage,* the reduction in the size of the multiple coefficient of correlation, is expected. It is the difference between the R^2 in the original validation sample and the R^2 in the cross validation sample. If shrinkage is negligible, the weights are considered stable; if large, the weights are not reliable and the composite is not recommended.

> Think of *shrinkage* like going to a tailor to have your tuxedo fitted; the tailor modifies the jacket and pants to fit the unique contours of your body. When your same-size sibling asks to borrow the tuxedo for a wedding, however, the tuxedo is only going to be an approximate fit on your sibling.

An alternative is to estimate shrinkage from a single sample by formula estimation. Formula estimates consider only sampling error, not measurement error, either random or systematic, but shrinkage from random error is lower in the large sample than in the smaller one resulting from dividing your sample in half. Wherry (1931) offered the most commonly used equation for estimating the shrunken coefficient from a single sample. As presented by Claudy (1978) but with notation used earlier, and in squared form, it is as follows:

$$\overline{R^2} = 1 - [(n-1)/n-k-1)] \cdot (1 - R^2) \tag{6}$$

where $\overline{R}$ = the estimate of the shrunken coefficient, R = the computed coefficient, n = sample size, and k = the number of predictors in the equation. Given the option, replication is preferable.

Validity Generalization

Validity generalization is a specific form of meta-analysis. Meta-analysis looks quantitatively for conclusions that have been drawn using independent research on the same basic hypothesis. Traditional literature surveys had the same objective but were verbal rather than quantitative, often imprecise, and subjective. Subjectivity remains in meta-analysis, primarily in coding information, but procedures are systematized and results are quantitative. Of the many approaches to meta-analysis, the one known also as validity generalization is the most directly appropriate to personnel testing.

Validity generalization (Schmidt & Hunter, 1977) assembles correlation coefficients from independent validation studies of the same hypothesis. The mean of

the resulting distribution is an estimate of the mean in the population from which the samples came. The variance of the distribution exceeds zero only to the degree to which results in the samples come from different populations, stem from different systematic influences, or are subject to different sources of error.

The validity generalization approach begins with the idea that the criterion-related validity coefficient is the same in all tests of the research hypothesis—or would be if not for artifactual influences on the results of individual studies. Coefficients can be corrected statistically for some artifacts, such as sampling error, criterion unreliability, and range restriction; corrections for others are applied to the estimated variance of the distribution of corrected coefficients. If that variance can be explained largely by these artifacts, then validity is said to generalize across the diverse situations from which individual coefficients came. If not, then systematic characteristics of different studies are examined as potential moderating influences.

Validity generalization tests two hypotheses in addition to the substantive hypothesis. The ***situational specificity*** hypothesis is that criterion-related validity depends in part on unknown influences within research settings; it can be rejected if corrections substantially reduce the variance of the validity coefficient distribution. Corrections cannot be made for unknown or unreported artifacts, so Hunter and Schmidt (1990) advocated a rule of thumb that rejects situational specificity if 75% or more of the variance is explained by known artifacts. Unknown artifacts may account for the rest so that the corrected mean correlation may be treated as the population value across all studies.

The validity generalization hypothesis is not simply the obverse of situational specificity, although rejecting the hypothesis of situational specificity is a necessary first step. Validity generalization is supported when nearly all of the validity coefficients in the distribution are at or above a nontrivial level and in the same direction (all positive or all negative). Reports usually identify the point in the distribution above which 90% or more of the corrected validity coefficients lie. If validity generalizes, the mean of the distribution of coefficients (after correction for statistical artifacts) is the best single estimate of validity in the job or job family sampled in the accumulated research.

Three different results occur in validity generalization research. A study may (1) refute (or support) the situational specificity hypothesis by showing (or not) that the variance of the distribution of corrected coefficients approaches zero, (2) support (or refute) the validity generalization hypothesis by showing (or not) that all or nearly all validity coefficients across diverse situations are nontrivial in size and in the same direction, and (3) (if situational specificity is rejected and generalization is supported), give an estimate of population validity in the form of the mean of the corrected coefficients ($\bar{r}_c$). The quality of these findings depends on how many of the artifacts the analysis has been able to correct and on how well the corrections have been made. Research reports rarely give all the information needed for the corrections, but meta-analytic results are usually more dependable than single-study results.

Schmidt and Hunter (1981) said, "Professionally developed cognitive ability tests are valid predictors of performance on the job and in training for all jobs" (p. 1128). This seems an overwhelming, even reckless, generalization, yet there is support for it, and it is important to note what it does *not* say. It does not suggest that all cognitive tests are *equally* valid predictors across all jobs, all criteria, or all circumstances. Validity in this sense means only that the correlation is nonzero across settings. For example, cognitive ability has been found to predict performance better for jobs higher in autonomy versus jobs that are routine and structured (Barrick & Mount, 1993). Indeed, keeping workers happy and preventing turnover in boring jobs would seem to argue against a heavy reliance on cognitive ability in selection. Note that the criterion has shifted from performance to satisfaction and turnover. Considerable thought needs to be given to what outcomes are most important, given the organization's needs and priorities.

Comments on Statistical Analyses

Chapters 6 and 7 have offered many equations relevant to the evaluation of predictors. Personnel researchers need extensive training in data analysis. A much wider variety of data analytic techniques are available to statistically well-trained people than described here, and different situations may favor different methods of analysis. The conventional statistics mentioned here are descriptive; they permit inferences of statistical reliability, but they are not well suited to seeing how well real data fit organizational needs or theoretical models. Research related to personnel decisions, perhaps due in part to the freezing of the field in the EEO era, has (like this chapter) given relatively little attention to newer, theory-confirming statistical methods. Those who will improve the empirical evaluation of assessment-based personnel decisions surely will develop a larger repertory of confirmatory techniques and models.

Researchers need an inclusive knowledge of statistical procedures, but there is an important caveat: *Statistics is a tool, not a religion*. Too often, researchers appear to have a blind faith in the results of statistical analysis. Statistics is a guide to judgment, not an alternative to it; results of statistical analysis merit thoughtful evaluation, not automated acceptance.

Discussion Topics

1. When selecting weights to apply in a multiple-predictor situation, what are the advantages and disadvantages of relying on each of the following: unit weights, regression analysis, SME judgments or the results of a job analysis unit weights, and finally, theory?
2. Why is it important to cross-validate results in personnel research?
3. What are some problems associated with using a multiple hurdles process in selection?

Note

1 Not everyone shares this preference. Linn (1994) considered "domain-referenced" to require domain specifications too rigid to be feasible for any but extremely narrow, finite domains; he said that "criterion-referenced" refers to "broader, fuzzier, but more interesting achievements" (p. 13). Glaser (1994), who introduced criterion-referenced testing (Glaser, 1963; Glaser & Klaus, 1962), prefers the original term, pushing aside the barnacles of misinterpretations of his idea that occurred over the years.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Barrick, M. R., & Mount, M. K. (1993). Autonomy as a moderator of the relationships between the big five personality dimensions and job performance. *Journal of Applied Psychology, 78,* 111–118.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods, 10*, 689–709.

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement, 2,* 595–607.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95–106.

Frederiksen, N., & Melville, S. D. (1954). Differential predictability in the use of test scores. *Educational and Psychological Measurement, 14,* 647–656.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 519–521.

Glaser, R. (1994). Criterion-references tests: Part I. Origins. *Educational Measurement: Issues and Practice, 13*(4), 9–11.

Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. Gagné (Ed.), *Psychological principles in system development* (pp. 421–427). New York, NY: Holt, Rinehart, & Winston.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis.* Newbury Park, CA: Sage.

Linn, R. L. (1994). Criterion-referenced measurement: A valuable perspective clouded by surplus meaning. *Educational Measurement: Issues and Practice, 13*(4), 12–14.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128–1137.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2,* 446–457.

Witt, L. A., & Ferris, G. R. (2003). Social skill as moderator of the conscientiousness-performance relationship: Convergent results across four studies. *Journal of Applied Psychology, 88,* 809–821.

8

MAKING JUDGMENTS AND DECISIONS

Intuitive Prediction, Judgment Aids, and Utility Analysis

The purpose of assessment is to provide a basis for decisions. Decisions about candidates are made by managers, not researchers. Good decisions depend on valid assessments. Whether the assessments lead to valid decisions, or are even considered in making decisions, depends partly on the way assessment data are presented. Decision aids help; managers should be trained in their use. Wise decisions require not only data and information, but also their integration into a broader experiential framework of information and knowledge. Researchers and HR specialists should not merely pump data—assessment scores—to their managers; they should make sure that the assessments are informative and fit into a broader scheme of managerial knowledge about people, jobs, the organization, and the position at hand.

Judgments of Validity

Validity of prediction is inferred, not merely computed. It is a judgment, to be inferred only if the preponderance of evidence supports the intended prediction. A validity coefficient is accepted as evidence of valid prediction only if the data and analysis are judged adequate. If local validation is not feasible, and no relevant meta-analysis exists, job relatedness can be based on two sequential judgments in an option emphasizing psychometric validity. First, a trait must be judged related to performance of important aspects of the job—the predictive hypothesis. Second, the assessment device must be judged a valid measure of that construct. If logic and data support both judgments, the assessment is judged a valid measure of a job-related trait.

Evaluating a Measure

1. Did the developer of the procedure seem to have a clear idea of the attribute (construct) to be measured? Was development informed by at least a rudimentary theory of the attribute?
2. Do you consider the measurement methods—including presentation, procedures, and response requirements—consistent with that idea?
3. Do you think the stimulus content is appropriate? Is the content domain unambiguous? Is it relevant to the measurement purpose? Was it properly sampled? Can responses be scored, observed, or evaluated reliably?
4. Can you infer care and skill in the development of the assessment instrument? Were pilot studies and item analyses done, and done well?
5. Is the score intended to reflect a single attribute or to sample a heterogeneous domain? If the former, are items internally consistent? If the latter, was the domain well defined and sampled systematically, and does it have at least a modicum of internal consistency?
6. Are scores stable over time?
7. Do the scores relate to other variables in a way consistent with the relationships expected from the theory of the attribute?
8. Do relationships disconfirm alternative hypotheses about the meaning of the scores?
9. Does the predictive hypothesis sensibly relate the attribute to job performance? Do job experts consider the attribute relevant? Is there prior research suggesting or even demonstrating its relevance?
10. Does a well-formed predictive hypothesis require other attributes of equal or nearly equal importance? If so, can the job relatedness of the attribute at hand be evaluated on its own?
11. Is there any reason to suspect that a nonmonotonic relationship exists? If so, is there any evidence suggesting the points in the assessment distribution where the relationship changes from positive to zero to negative?
12. Are criteria measured validly and predicted with reasonable accuracy? The question assumes criterion-related validation but requires judgments about criterion validity and possible contaminants, adequacy of research design, sufficiency of sample in size and composition, and others.

Answers to questions 1–8 in this list can be drawn from manuals or other documents or from local research. Favorable answers form a basis for inferring psychometric validity. Positive responses to the remaining items provide evidence of job relatedness—even where the criterion-related evidence

implied by question 12 is missing. Answers to some of the questions are data based, but they require judgment, if only to judge the adequacy of the data. If an overall judgment of job relatedness is based on good reasons for favorable responses to most of questions 1–11, it is probably better than that based on a single, local, unreplicated criterion-related validity study.

Managerial Use of Assessments

Managers—not researchers, test developers, staff psychologists, or HR specialists—make staffing decisions. Managers, despite wanting the best people, usually want to fill a vacancy satisfactorily as quickly as possible; testers generally want to maximize performance and compliance with government regulation. Most managers have no training in psychometrics or test theory, they may not understand the constructs assessed, and they may hold unwarranted views about tests. Some managers distrust tests and place little reliance on test scores. Perhaps worse is a manager who believes tests are great, who defers to test scores even when evidence shows them invalid, and who simply does not hear warnings or qualifications about them. To deal with both kinds of unwarranted views, some staff psychologists establish rules for using tests or other assessments in making personnel decisions. The rules might specify preferred score levels or patterns, circumstances to justify overlooking poor scores, or further information to consider along with test scores or other systematic assessments. Some managers may decide for themselves whether to use test information and, if so, how to use it. That seems to be an odd policy. Developing and validating systematic, standardized assessment programs requires an investment. It is strange to let individual whim determine how or whether the results of the investment will be used. Those responsible for the assessment programs should take active steps to gain program acceptance and to assure proper use of scores.

Judgments as Predictions and Decisions

Personnel selection decisions are judgments that constitute a prediction about future success on the job. Too often, the people making these decisions are not evaluated seriously beyond vague statements like, "Our personnel director really knows how to size people up." We can do better.

Many judgmental predictions are not even recognized as such; that is, no clear statement identifies the basis for judgment or hypothesizes that it is somehow related to a predictable outcome. One might, of course, formally frame and test a hypothesis that a firm handshake, direct eye contact, or some other form of body language indicates that the candidate will work hard or be conscientious. More often, such cues are not even recognized as the basis for judging that "this person is a good bet."

Clinical and counseling psychologists make "clinical predictions"—judgments—of likely future behavior. In evaluating a convict being considered for parole, a psychologist's duty may be to predict whether the person, if paroled, will be a repeat offender. Such predictions are not made lightly; the psychologist responsible for them gathers much data about the person, considers much data about recidivism in general, and gives these data much thought before making the prediction.

So also the personnel decision maker may make informed judgments about candidates and the performance expected of them if hired. Candidates might be tested, interviewed, and evaluated in assessment centers; their backgrounds might be checked, and people who have known them in various contexts might be interviewed for still more data. Much of the data about a candidate might be useless, and the decision maker may not know the value of specific pieces of information. Yet a decision must be made. It can be made on the basis of informed and explicit judgments, and those judgments are more or less well-informed predictions.

Statistical analysis can be misleading, too, particularly when data are poor or greatly violate statistical assumptions. Nevertheless, Meehl (1954) demonstrated long ago that statistical prediction is consistently superior to clinical (judgmental) prediction. Later, he suggested six circumstances that might, perhaps, favor clinical prediction. Among them was the idea that optimal prediction might be based more on patterns of relationships among predictors than on the linear, additive relationships assumed in the multiple regression equations (the most common statistical prediction); perhaps well-informed clinicians (judges) could identify salient patterns better than arithmetic processes could. People making judgments might be using information in a "configural" way (i.e., using algorithms that may be nonlinear, nonadditive, or even noncontinuous). It was an interesting idea but, in subsequent research, it did not pan out. It has been a well-accepted view that statistical prediction is almost always, some even say necessarily, better than prediction by human judgment. Table 8.1 outlines a number of lay assumptions about judgmental prediction (vs. statistical formulae) that are considered conventional wisdom—even though none of them hold up against the evidence.

Prediction and Decision Without Statistics

Statistical prediction is feasible when common criterion and predictor data can be collected for a lot of people. In these cases, employment decisions can be evaluated in terms of mean performance of those selected. However, many decisions must be made without the luxury of research data. For unusual jobs, many high-level jobs, or lower level jobs in small organizations, many candidates may be assessed, but only one (or a few) may be chosen. The cost of error in these cases, and the reward for being right, may be greater than in those where statistical predictions are feasible. If only one person is chosen, that person's performance is the crucial evaluation of the decision. In small organizations or large, the higher the

TABLE 8.1 Lay Assumptions About Judgmental Prediction

Lay Assumption	*State of the Evidence*
Assessors can take into account constellations of trait and ability data.	Just as astrologers are unable to conduct "whole chart" interpretations to render their professional judgments, there are far more unique configurations of data than can be cognitively processed by assessors.
Assessors can identify idiosyncrasies that formulas ignore.	The problem is that assessors overrely on idiosyncratic cues, not distinguishing the useful from the irrelevant. Assessors find too many "broken legs."
Assessors can "fine tune" predictions made by formulas.	Intuition could be used to alter the formula-based rank-ordering of candidates. We have yet to find evidence, however, that this results in an improvement in prediction of job performance.
Some assessors are better than others.	Although training can improve prediction, experience has no demonstrated impact. There are experts in many domains, but evidence for expertise in intuitive prediction is lacking.
Candidates for technical jobs don't differ much on ability and personality.	Research has shown that managers and executives are more variable in ability and motivation than conventional wisdom suggests. Test scores can predict for technical jobs.
Formulas become obsolete.	Assessors are likely to rely on implicit theories developed from past training and experience, and these have likely become resistant to change. Formulae may be updated on the basis of new information and empirical research.

Note: Adapted from Highhouse, S., & Kostek, J. A. (2013). Holistic assessment for selection and placement. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology*. (pp. 565–577). Washington, DC: American Psychological Association.

organizational level, the fewer the incumbents. Even with dozens of keepers of accounts, there is but one comptroller. When that comptroller retires, moves to a different organization, gets fired, or dies, another one person must be chosen to fill the position; perhaps a dozen people or more may be considered. There is no choice between judgmental and statistical prediction; judgment is the only option.

Communicating Statistics to Lay People

Many managers lack the prerequisite knowledge about statistics and probability to fully grasp the impact of data-based assessment practices. Big data exacerbate this problem by creating even more complex algorithms. One possible solution to this problem is to incorporate visual aids in presenting data to make data more user friendly.

In the past decade, there have been several endeavors aimed at presenting data to the public using visual displays ranging from interactive graphs (Gapminder.org) to sophisticated infographics (InformationIsBeautiful.net). InformationIsBeautiful.net combined design principles with data visualization tools to create a large library of engaging graphs across a variety of contexts. Certain visual aids, such as Iconarrays (Iconarray.com), create charts with human avatars as icons to illustrate the benefits of medical treatments (e.g., Figure 8.1). There are several advantages to "humanizing" the data. First, graph comprehension requires people to relate visual features of the graph to conceptual relations represented in the graph (Kosslyn, 1989; Pinker, 1990; Shah & Hoeffner, 2002). Human silhouette or stick figures automatically evoke the concept that each icon represents a person, making the graph comprehension process easier (Cleveland, 1993; Larkin & Simon, 1987). Second, human icons can reduce the demand of working memory because they reduce the need for the viewer to keep track of the meaning of graphical elements (Kosslyn, 1994). Finally, human elements can evoke imagery about individual persons in the data, which can lead to stronger affective responses (Slovic, Peters, Finucane, & MacGregor, 2005).

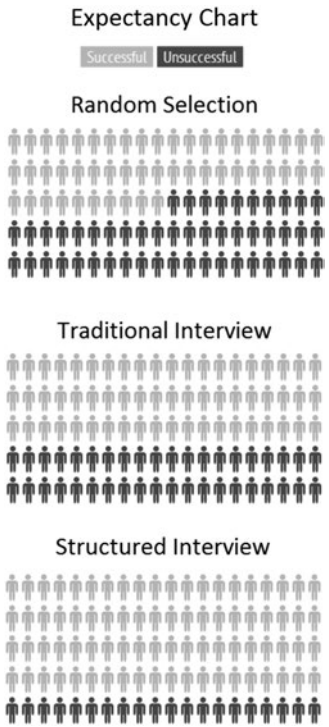

FIGURE 8.1 Expectancy chart using icon arrays.

Judgment Aids

Often, statistics used to describe the validity of hiring methods underestimate real-world utility. Although a correlation of .40 may seem large to a scientist, it is less impressive to a hiring manager (Muchinsky, 2004). There are several alternatives to traditional effect sizes (i.e., r and r^2). The alternative methods entail transforming the traditional effect sizes to either a ratio or a percentage (Breaugh, 2003). Effect sizes can be rewritten as a relative risk (RR), which is the ratio of the probability of an outcome (e.g., good employee) in one group (high cognitive ability) compared with another (low cognitive ability). The relative risk ratio translates the esoteric effect size statistic into natural language that is easy to understand.

Effect sizes can also be described as percentages. The first method is describing the effect size as improvement in the percent of correct decisions over a random process or existing procedure (Kuncel & Rigdon, 2013). In a hiring context, one can say: "Compared with your interview, the use of an algorithm produces a 20% increase in the number of successful future employees." The Common Language Effect Size (CLES) represents the probability that a randomly sampled score from one distribution will be greater than a score sampled from a separate distribution (McGraw & Wong, 1992). In a selection context, it is the probability a randomly selected person who passed the hurdle will perform better than one who did not pass. Finally, the Binomial Effect Size Display (BESD) displays in tabular form the change in the percent of a particular outcome between groups. In Table 8.2, you will find a BESD that illustrates the change in successful outcomes when moving from random selection, to a traditional interview, and finally to an unstructured interview.

The BESD combines the ease of interpretability of percentages and the visual appeal of tables. Brooks, Dalal, and Nolan (2014) found that both the CLES and the BESD were easier to understand, more useful, and more effective than traditional effect size metrics.

Expectancy Charts. In Chapter 7, we noted that assessment professionals are more likely to use expectancy charts or expectancy graphs rather than normative data. In the situation where local data is available, it is a relatively simple matter to

TABLE 8.2 A Binomial Effect Size Display (BESD)

		Employee Future Performance Outcome	
		Unsuccessful	*Successful*
Hiring Method	Random Selection	50	50
	Traditional Interview	40	60
	Structured Interview	20	80

construct expectancy charts and graphs. Even in those situations where tests are used but not empirically or locally validated, theoretical expectancy charts and graphs can be useful aids, if an acceptable validity coefficient is available or can be estimated. They can be developed using a validity coefficient (a) reported in a manual or other research report based on a comparable situation, (b) estimated from an appropriate meta-analysis, (c) estimated less formally from a body of prior research where each study fits only part of the situation at hand but the accumulated data fill it reasonably well, or (d) estimated by panels of experts.

Expectancy charts show the percentage of those at given test score intervals that could be expected to be successful on the job. Expectancy charts are for decision makers, not for researchers. They provide a sense of the kind and strength of relationship found but without the precision of a regression pattern or correlation coefficient. They do, however, promote understanding of the usefulness of predictors and help in making decisions about applicants. People with little statistical training can make predictions using expectancy charts.

Consider a situation where the assessment plan for a specialized sales position includes assessment by tests of two traits, general intelligence and surgency. Suppose that, for both psychometric and theoretical reasons, the Watson-Glaser *Critical Thinking Appraisal* (*CTA*; Total Score) and the *Hogan Personality Inventory* (*HPI*; Sociability Score) are used. Assume that no relevant validity coefficient was found for the *CTA*, but that an expert panel linked component scores to job duties and concluded that a total score validity coefficient, appropriately corrected, would not be less than .35. A validity coefficient of .51 reported in the Hogan Manual (Hogan & Hogan, 1992) for advertising sales, a position making interpersonal demands similar to those in the position at hand, was rounded down to .50. These estimates permit theoretical expectancy charts for these two tests, shown in Figure 8.2. Score intervals on the *CTA* were derived from the full norms for upper division college students; those for the *HPI* are based on total sample norms. Examination of Figure 8.2 shows that someone who scores 63 on the *CTA* would have approximately 58% chance of success on the job, whereas someone who scores 23 on the *HPI* has approximately 78% chance of success on the job.

Expectancy Graphs. Theoretical expectancy graphs can be developed from empirical data by assuming linear regression and a normal bivariate correlation surface, the characteristic oval-shaped outline of a bivariate scatterplot when both distributions are approximately normal. If each distribution is dichotomized, as in Figure 8.3, the proportion of superior workers above or below a specified score can be estimated. The proportion of the total area above the horizontal line (quadrants a and b) represents the proportion of employees classed as superior. The area to the right of the de facto cut score (a and d) represents the proportion of applicants who were hired: the ***selection ratio***. The shaded quadrant a represents people who are both hired and superior. The proportion of those accepted who are also superior can be increased by relaxing

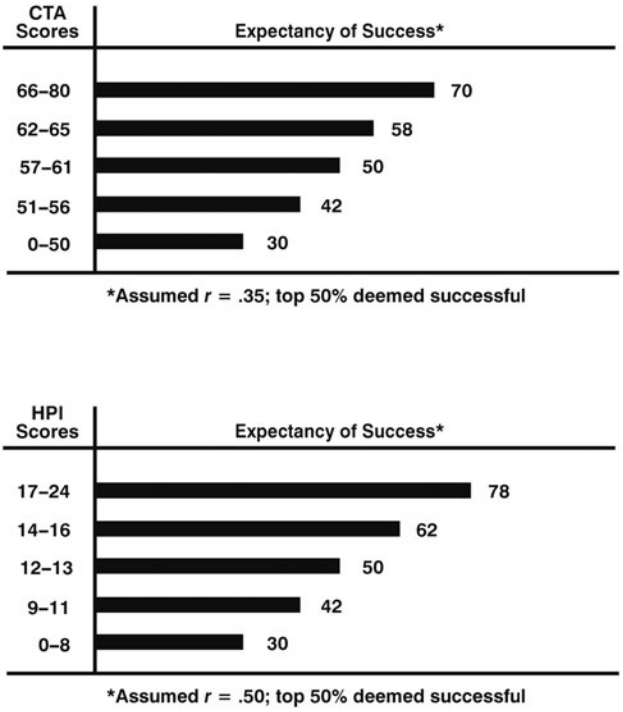

FIGURE 8.2 Theoretical expectancy charts for two tests for a hypothetical specialized sales position.

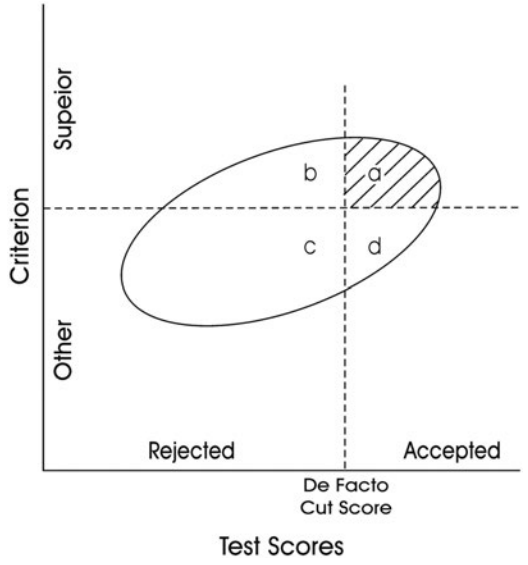

FIGURE 8.3 Schematic diagram of relationship between validity, percentage superior without testing, and de facto cutting score.

one's definition of success (i.e., by lowering the horizontal line dichotomizing the criterion), or by decreasing the selection ratio (moving the vertical line to the right), or by making the relationship stronger (i.e., increasing correlation that narrows the elliptical outline of the data). Reducing the performance level defining success does little to improve organizational functioning. Raising the de facto cut score requires more recruiting.

The concept of the use of expectancy charts is usually associated with Taylor and Russell (1939). Expectancy graphs, along with expectancy charts, are also seen as an early attempt at operationalizing utility.

The use of expectancy graphs can help to illustrate to managers the costs involved in moving the cut score up (more chance of overlooking a good applicant), and moving it down (hiring more people who will not succeed). It is also a way of demonstrating the impact of validity on the selection errors. Imagine if the oval-shaped outline of the scatterplot is changed to be much narrower (as happens with increased validity); the scatterplot goes from looking like a football to looking like a cucumber. Now the percentage of people who are hired and fail, along with the percentage of people who are overlooked, drops considerably. The graph demonstrates clearly how the selection ratio, validity, and definition of performance has an impact on the success of a selection program.

A staff psychologist's or researcher's responsibility includes assuring that decision makers are trained in (a) the nature of the constructs being assessed, (b) why they are important, (c) the fundamental principles by which the assessment of the constructs was evaluated, (d) the nature of defensible and indefensible inferences from scores, and (e) acceptable limits of individual judgments to override ordinarily defensible inferences. Expectancy charts and graphs can help. They help teach that prediction of either success or failure is rarely certain but is instead probabilistic. They usually show that the probability of success is greater at higher score levels. Good training would also teach the limits of predictions such as those imposed by the criterion chosen; an expectancy of a superior level of production gives no clue about probable performance on a criterion the decision maker might have preferred, such as a dependability or ingenuity.

Utility Analysis

There are many potential indices of the value of a test. We have already discussed the use of expectancy graphs and charts in assessing decision accuracy. Other simple indicators of the potential usefulness of a test include reliability

and validity. However, because tests have been used in selection, an obvious question has been whether we can determine the value added from the use of a test in dollars. Being able to convert validity into dollars would provide a simple metric for comparing tests and also for comparing selection to other organizational interventions (Cascio, 1993, 2000). Estimating the value of a test in terms of dollars is the goal of "utility analysis," although utility analysis has now been expanded to include a consideration of a wider set of costs and benefits (Boudreau, 1996).

Utility analysis is a formal, analytic study of usefulness that can serve as a decision aid when statistical information is available on predictor–criterion relations—even if only from meta-analyses or expert judgments of validity. Utility analysis can serve many organizational purposes; we mention just three. First, it can aid decisions about using a particular procedure, comparing the benefits of its use to the costs incurred in installing and using it. Second, it offers a means for choosing between alternatives. Many assessments are based on paper-and-pencil aptitude tests. An alternative might be a hands-on assessment of existing competencies. Utility analysis might determine the relative utility of each form of assessment. Where considerations like costs differ greatly, as they do in comparing paper-and-pencil tests with hands-on performance tests, utility analysis can be an important decision aid. Third, utility analysis is a tool for the internal marketing of a proposed program. Modern history includes many great ideas abandoned or never implemented because of a lack of compelling evidence of their worth. As a case in point, Johnson and Johnson (1975), describing a book on the famous Hawthorne studies and subsequent counseling program, said that the Hawthorne plant had 5 counselors on the staff in 1936, a peak of 55 in 1948, and down to 8 in 1955: "There came a time when new management . . . began to ask questions about justifying the cost of it. Under the impact of this questioning, the program declined" (Johnson & Johnson, 1975, p. 275). Perhaps, but by no means certainly, utility analysis might have saved the program.

As early as 1946, Brogden laid out the basic theory of converting a validity coefficient into an estimate of utility in dollars. Cronbach and Gleser (1957, 1965) built on the ideas of Brogden and offered a formula for the calculation of utility in dollars; however, the difficulty in using the Cronbach and Gleser formula was the absence of a measure of SD_y. The problem of estimating SD_y was solved in a classic article by Schmidt, Hunter, and colleagues in a 1979 article (Schmidt, Hunter, McKenzie, & Muldrow, 1979). The development of simplified methods of estimating SD_y led to an explosion of interest and research in utility analysis.

Although many complex models of utility have been proposed, we present a highly simplified model here for illustrative purposes. The simplified equation, adapted from Brogden (1946; 1949) and Cronbach and Gleser (1957) is as shown:

$$\Delta\$ = [(N)(\Delta r_{xy})(SD_y)] - [(A)(C)]$$

where $\Delta\$$ is the average dollar payoff as a result of using the selection procedure. N is the number of employees selected by the procedure. SD_y is the difference in revenue generated by average versus above average employees (commonly estimated at 40% of the average employee salary). Δr_{xy} is an estimate of how much the use of the procedure will improve the quality of hiring. This can be obtained by subtracting the effectiveness of the method currently in use from the method in consideration. For example, Table 8.3 shows validity estimates of selection methods used for hiring salespeople (Vinchur, Schippmann, Switzer, & Roth, 1998). If a company is currently using an unstructured interview and is considering the use of a measure of sales ability, Δr_{xy} would be found by subtracting the r_{xy} for the unstructured interview from the r_{xy} for the sales ability measure (.45 – .20 = .25). C is the estimated cost of the proposed selection method (see Table 8.3), and A is the number of applicants tested. Using this example for a situation in which 100 salespeople will be hired from 500 applicants at $40,000 starting salary would yield:

$$\$375{,}000 = [(100)(.25)(\$16000)] - [(500)(\$50)]$$

Estimates from utility equations have sometimes been staggering and even incredible. It may be that the individual level of analysis typical of most utility studies ignores the system that is the organization and, therefore, exaggerates expected utility. People seeking to convince others of the value of specific programs must be as conservative as possible to make their estimates seem realistic to managers who have seen other projections of potential savings go sour. Utility analysis may be most useful in considering the relative utilities of available options.

TABLE 8.3 Validity and Estimated Cost of Selection Methods for Hiring Salespeople

Selection Method	r_{xy}	*Estimated Cost*
Random	.00	0
Graphology	.00	$75
Unstructured Interview	.20	$50
Biodata	.50	$75
Sales Ability Inventory	.45	$50
Potency (Extraversion)	.28	$75
Achievement Orientation	.25	$75
Cognitive Ability	.30	$50

Utility analysis can help us determine the costs and benefits of implementing an assessment program. It can help identify which of multiple assessment procedures is most economical, and it can help *sell* the benefits of an assessment program to management. In one of the more creative and interesting debates in the selection literature, the argument has been made that presenting utility information decreases, rather that increases, support for tests and selection initiatives (Cronshaw, 1997; Latham & Whyte, 1994; Whyte & Latham, 1997).

The greatest challenge, however, in utility research may be in overcoming the perceived complexity of the method. In a survey of members of regional associations of applied psychologists and HR professionals, Macan and Highhouse (1994) found that managers did not respond well to utility procedures, and that the professionals themselves found the equations complex and difficult to understand. Highhouse (1996) argued that researchers need to focus on the utility estimate as a *communication device,* and that factors such as ambiguity, credibility, and latitude of acceptance deserve attention.

It seems that other disciplines have been far more effective in communicating the utility of such things as an improved diet, increased exercise, and wearing seat belts. Personnel psychologists should follow their lead in determining simpler and clearer methods of communicating utility.

Many consultants have moved to simpler approaches to communicating the value of tests. Some of the approaches currently used are (1) reporting results in terms of turnover rather than job performance, and then converting turnover into dollars; (2) calculating the ratio of number of hires to number interviewed; and (3) reporting the reduction in the hiring of weak or poor performers (which can be seen as a return to the ideas of Taylor and Russell, 1939). The results are then presenting using digital dashboards, which allow decision makers to monitor the contribution of their selection systems to organizational performance.

Concluding Comment

We have tried to argue that judgmental and statistical prediction is not an "either/or" choice (Westen & Weinberger, 2004). Indeed, judgmental prediction can be transformed easily into statistical prediction by transforming impressions into scale

scores and combining these scores with test scores using a judgmentally based formula. It is the *consistency* of the mechanical formula that accounts for its accuracy on average. People have a great resistance to formulas. Why is this so? One reason is that people know that formulas blindly applied are bound to lead to prediction errors. For example, a graduate program that simply selects applicants based on a combination of scores will occasionally overlook those "diamonds in the rough" that would have been successful if given the chance. The graduate program, therefore, is closing its eyes to mistakes that could otherwise be avoided. What people fail to consider, however, is the even greater number of errors that would be made without use of the mechanical formula.

This does not mean that the decision maker should always ignore information that cannot be formalized into an analytical procedure (e.g., the job candidate makes a racist comment at dinner). People can recognize important information that formulas will never consider. The challenge is to recognize when this information is truly job relevant, versus a personal theory with no validity.

Nobel Prize winner Daniel Kahneman (1992; Tversky & Kahneman, 1991) explicitly credits famous social psychologist Kurt Lewin (1951) for recognizing that reducing one's losses is much more powerful than offsetting them by gains when the goal is to unfreeze the current situation and move toward change. Lewin recognized that tension is undesirable, and the optimal way to induce change was to identify and relieve restraining forces (i.e., perceived losses). Consider the potential psychological losses that may accompany a move toward data-based hiring. The employer no longer controls the process and runs the risk of being seen by others as unprofessional and reckless. Promises of gains in validity are not likely to offset these perceived threats to professional dignity. One possible solution to this is to use data-driven screening, which results in two or three candidates who are roughly equally likely to succeed—given what is possible to know at the time of hire (i.e., they are Pareto efficient). These finalists could then be judged individually by the employer. As Kuncel (2008) put it, "Decision makers can exercise their preference for unstructured interviews, firm handshakes, and holistic impressions without gross deviation from top-down decision making" (p. 343). Such a process preserves the benefits of data-based assessment while allowing the employer to feel a sense of control, and others can see that a person—not data—made the decision.

Discussion Topics

1. Why do people resist using formulas to make decisions? What implications does this resistance have for advising managers on making hiring decisions?
2. How can people doing selection be better trained to judge what is job relevant? How can interview impressions be standardized and compared with test scores?
3. Why do people have such a hard time understanding utility analysis? What could be done to simplify it?

References

Boudreau, J. (1996). The motivational impact of utility analysis and HR measurement. *Journal of Human Resource Costing & Accounting, 1*(2), 73–84.

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management, 29,* 79–97.

Brogden H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology, 37,* 65–76.

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology, 37,* 65–76.

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology*, *99*, 332–340.

Cascio, W. F. (1993). Assessing the utility of selection decisions: Theoretical and practical considerations. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 310–340). San Francisco, CA: Jossey Bass.

Cascio, W. F. (2000). *Costing human resources: The financial impact of behavior in organizations* (14th ed.). Boston, MA: Kent.

Cleveland, W. S. (1993). *Visualizing data.* Murray Hill, NJ: Hobart Press.

Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions*. Urbana-Champaign: University of Illinois Press.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana-Champaign: University of Illinois Press.

Cronshaw, S. F. (1997). Lo! The stimulus speaks: The insider's view on Whyte and Latham's "The futility of utility analysis". *Personnel Psychology, 50,* 611–615.

Highhouse, S. (1996). The utility estimate as a communication device: Practical questions and research directions. *Journal of Business and Psychology, 11,* 85–100.

Highhouse, S., & Kostek, J. A. (2013). Holistic assessment for selection and placement. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology* (pp. 565–577). Washington, DC: American Psychological Association.

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory: Manual.* Tulsa, OK: Hogan Assessment Systems.

Johnson, D., & Johnson, R. (1975). *Learning together and alone: Cooperation, competition, and individualization.* Englewood Cliffs, NJ: Prentice Hall.

Kahneman, D. (1992). Reference points, anchors, norms, and mixed feelings. *Organizational Behavior and Human Decision Processes, 51,* 296–312.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology, 3*(3), 185–225.

Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate.* Cambridge, MA: MIT Press.

Kuncel, N. R. (2008). Some new (and old) suggestions for improving personnel selection. *Industrial and Organizational Psychology, 1*(3), 343–346.

Kuncel, N. R., & Rigdon, J. (2013). Communicating research findings. In N. Schmitt & S. Highhouse (Eds.), *Handbook of psychology* (Vol. 12; pp. 43–58). New York, NY: Wiley.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*(1), 65–100.

Latham, G. P., & Whyte, G. (1994). The futility of utility analysis. *Personnel Psychology, 47,* 31–46.

Lewin, K. (1951). *Field theory in social science.* New York, NY: Harper.

Macan, T. H., & Highhouse, S. (1994). Communicating the utility of human resource activities: A survey of I/O and HR professionals. *Journal of Business and Psychology, 8,* 425–436.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361–365.

Meehl, P. E. (1954). *Clinical versus statistical prediction.* Minneapolis: University of Minnesota Press.

Muchinsky, P. M. (2004). When the psychometrics of test development meets organizational realities: A conceptual framework for organizational change, examples, and recommendations. *Personnel Psychology, 57,* 175–209.

Pinker, S. (1990). A theory of graph comprehension. In R. Frele (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Hillsdale, NJ: Erlbaum.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology*, *64*, 609–626.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review, 14,* 47–69.

Slovic, P., Peters, E., Finucane, M. L., & MacGregor, D. G. (2005). Affect, risk, and decision making. *Health Psychology, 24*(4S), S35.

Taylor, H. C., & Russell. J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology, 23,* 565–578.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics, 106*(4), 1039–1061.

Vinchur, A. J., Schippmann, J. S., Switzer, F. S., III, & Roth, P. L. (1998). A meta-analytic review of predictors of job performance for salespeople. *Journal of Applied Psychology, 83,* 586–597.

Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist, 59,* 595–613.

Whyte, G., & Latham, G. (1997). The futility of utility analysis revisited: When even an expert fails. *Personnel Psychology, 50,* 601–610.

9

ANALYZING BIAS AND ENSURING FAIRNESS

Unfair Discrimination, Item and Test Bias, and Reducing Adverse Impact

Fairness and freedom from bias are not the same, but there is some overlap in meaning. Bias is technical, whereas fairness is more a matter of opinion, policy, and power. However, the reality is that in common parlance, terms such as "bias," "discrimination," and "unfairness" take on surplus connotations, most of which tend to be negative. During the civil rights debates of the late 1950s and early 1960s in the United States, and continuing into the civil rights debates of the 2010s, words like "bias," "unfairness," "discrimination," and "prejudice" seemed to be used almost interchangeably, which adds to the confusion when trying to communicate the results of the research literature. Although psychometricians have attempted to develop neutral, scientific, operational definitions, going so far as to coin new terms such as "measurement equivalence," "differential item functioning" and "differential test functioning," definitions and distinctions are still not universal, especially in the popular press and also among attorneys, politicians, and social commentators.

Bias refers to systematic group differences in item responses, test scores, or other assessments for reasons unrelated to the trait being assessed. Bias is more easily alleged than demonstrated; it is easier to imagine the various kinds of third variables that may bias scores than to show their influence. If a test item requires knowledge common in one group but not in another, and if that knowledge is irrelevant to the trait, then the item is biased. It is *culturally biased* if an acceptable response depends on skills or information common in one culture but not in another. Cultural bias can be expected across countries in multinational organizations, but it is less certain for subcultures (e.g., Black and White) within a single national experience where the same media of mass communication (movies, television, print media, school curricula, etc.) give subcultures much in common despite some profound differences.

In employment, fairness refers to the job relevance of a potentially biasing or discriminatory practice. An item in a job knowledge test for chefs may ask about baking a cake, a task likely to result in higher scores for females than males; however, requiring the performance of a task or asking questions related to a topic favoring one gender over another would not be seen as unfair, even though it might be seen as biased, in that such an item can be clearly linked to essential tasks required for acceptable job performance. Answers may distinguish those with prior understanding of the activity from those without it. It discriminates precisely on the basis of knowledge relevant to the job, and that is the intent of the test.

In short, fair discrimination distinguishes those highly likely from those less likely to achieve a performance standard. ***Unfair discrimination*** exists "when persons with equal probabilities of success on the job have unequal probabilities of being hired for the job" (Guion, 1966, p. 26).[1]

Discrimination

Discrimination Based on Group Membership

Under Title VII of the Civil Rights Act of 1964, and the Civil Rights Act of 1991, it is illegal in the United States to discriminate against any person on the basis of race, color, religion, sex, or national origin. Groups of people identified on such bases (different laws identify various bases) are called protected groups.

> The term "protected groups" does not mean that some *people* are not protected under Title VII law, only that some *group characteristics* are not. Various laws also offer protections for other characteristics such as age and disabilities. In the first edition, the book discussed protections for sexual orientation. Since the first edition, there have been judgments in court cases that have argued that sexual orientation is covered under the protections from discrimination offered under the sex category. In particular, the EEOC has held that discrimination against an individual because that person is transgender and stereotyping based on sexual orientation are both examples of discrimination because of sex and, therefore, covered under Title VII of the Civil Rights Act of 1964.

Discrimination need not be intentional to be illegal. A procedure (e.g., tests or interviews) with the effect of unfairly discriminating against people in a protected group is discrimination under the law, even if inadvertent. This form of discrimination is referred to as a "disparate impact" theory of discrimination.

Although less frequent in the current era, procedures with only a "chilling effect," discouraging applications, may also constitute illegal discrimination. Organizational decision makers must be alert to inadvertent or chilling discriminatory practices—even if only to avoid litigation—and be aware that unfair discrimination that is legal is nevertheless unwise.

In the United States, group-based discrimination is so entangled with legal issues that groups defined in other ways are often overlooked. Socioeconomic groups, groups defined by cultural or intellectual habits, and other kinds of groups without legal or political protection may be discriminated against with no threat of litigation. Such discrimination is nevertheless poor management; it can rob the organization of people with excellent qualifications. Many kinds of people are, perhaps routinely, discriminated against on the grounds that for one reason or another they do not fit with the image, culture, or climate of the organization. This may include people with unusually long or short hair, people who are unusually tall or short, people who are not well-dressed or are too well-dressed—in short, people with characteristics that displease the decision makers. Focusing on valid, job-related assessment can reduce such instances of bias in decisions.

Distributional Differences

Statistical analysis of bias and discrimination is necessarily group oriented. Analyses can examine group differences in score distributions, in validity, or in predictions. Unfortunately, the only commonly considered distributional difference is the difference in mean scores. This is not enough; differences in variance, skewness, other distributional characteristics, and psychometric differences that influence the distributions should also be considered in analyses of bias.

Group Mean Differences. A lower mean test score in one group compared with another is not by itself evidence of bias, nor is use of test scores with group mean differences evidence by itself of discrimination. Nevertheless, too often a mean difference is the only basis for allegations of discrimination. Markedly different mean scores can occur for many possible reasons other than bias. Consider just four of them:

1. The two groups are biased samples of their respective populations. One group is among the best in its population, the other from those in the lower tail of its population distribution.
2. The two groups are representative samples of populations that actually differ on the trait being measured.
3. Many of the test items require background experiences not common in the lower scoring group.
4. Conditions of test administration differed in the two groups.

The first of these is plausible if the higher scoring group was subjected to stringent screening and the lower scoring group came from extensive, uncritical recruiting. If the second is plausible, different means may not indicate bias at all. The experiences in the third may be job related. The fourth may describe an error in administering a test in one of the groups. The many reasons for mean differences are extremely difficult to evaluate. A conference on civil rights reached agreement on this if on little else: "Average group differences in test scores do not necessarily reflect bias arising from test construction or use. . . . Average group differences in test scores may remain in tests even if all bias is removed" (United States Commission on Civil Rights, 1993, p. 7). Referring to mean differences as bias, without even thinking about nondiscriminatory potential causes, is simplistic and misleading; citing mean differences as bias and denying a genuine possibility of true differences is dishonest.

Many textbooks sidestep the issue of group mean differences in test scores, such as the rather substantial differences in average cognitive ability test scores between Whites and Blacks. The reason for this is unclear, but it could be that people fear raising an issue that is believed to be inherently racist. In other words, people may fear that group differences in cognitive ability are due to race and not some third variable. Certainly, we know that cognitive ability does have a heritability component. It does *not* follow, however, that the group differences themselves are based on heredity. A trait can have a heritability component *and* show group differences having nothing to do with heritability.

Differences in Other Distributional Characteristics. Distributions may differ in variance. Protected groups may include people from disadvantaged, even dysfunctional, backgrounds—and also people with more education and higher socioeconomic heritage. A plausible hypothesis is that minority groups have higher variance on tests of occupational skills and information influenced by personal background experiences, as illustrated in Table 9.1. The group means are different, but the difference in variability is greater. If it were possible to hire all people with scores of 16 or more, with top-down selection, 50% of Group A would be hired but only 22.5% of Group B. However, at a smaller selection ratio, the effect may disappear or even reverse because of the differences in variance; if only the top-scoring 10 of the 120 candidates are hired (those with scores of 18 or more), then the proportions hired are 7.5% in Group A and 10% in Group B. If Group A has a higher mean, less variance, and less skewness than Group B, is the test biased against either group? Only if these differences stem from causes unrelated to the trait being measured. Nothing in the distributional statistics, however, speaks clearly to that point; the only clarity is that the relative proportions receiving favorable decisions is affected by a combination of these statistics and the selection ratio.

Discrimination as Systematic Measurement Error. Distributional differences may stem from true differences or from systematic sources of measurement error

TABLE 9.1 Hypothetical Distributions for Two Groups Differing in Test Score Means and Standard Deviations

Raw Score	*Group A*		*Group B*	
	f	*cum f*	*f*	*cum f*
20	1	80	1	40
19	2	79	2	39
18	3	77	1	37
17	9	74	1	36
16	21	65	2	35
15	23	44	2	33
14	10	21	5	31
13	6	11	6	26
12	3	5	7	20
11	1	2	6	13
10	1	1	4	7
9			1	3
8			1	2
7			0	2
6			0	2
5			1	1
M	15.3		12.6	
SD	1.7		3.6	

related to group membership. The latter can happen when groups are defined or influenced by unmeasured third variables such as test-taking habits. If the influence of a third variable is greater in one group than in others, it can be a source of unintentional, unknown, and unfair discrimination—even if not illegal. One test user who was too cheap and unethical to buy her tests used poor quality photocopies instead. It was easy to show that this user not only violated copyright laws but also reduced the validity of her intended inferences; visual acuity was a strong influence on the scores. Scores were biased against people with even mild visual disability; that constituted unfair discrimination. The incident occurred long before ADA, so using the scores for decisions was unfair and unwise, but not yet illegal. Unfair discrimination denies jobs to qualified people and denies the services of qualified people to organizations. Unfair discrimination caused by unknown and unmeasured third variables may reduce both psychometric validity and job relatedness of a test.

Analysis of Adverse Impact and Bias in Test Use

Test bias is a psychometric term referring to distortion from different unwanted sources of variance in scores from different groups. *Adverse impact* is a social, political, or legal term referring to an effect of test use (Arthur, Doverspike, Barrett, & Miguel, 2013). Table 9.1 illustrates adverse impact if the test is used to select a lot of people, but no adverse impact for filling only a few positions. The group means are different, but the difference in variability is greater.

Adverse Impact

In Chapter 4, we briefly discussed adverse impact and the disparate impact theory of discrimination. In this section, we provide a more in-depth discussion of the analysis of adverse impact.

Under most circumstances, test scores will have adverse impact against some protected group (Arthur et al., 2013). Adverse impact is a legal term, not a statistical or psychometric one, referring to whether there are practical or significant differences in the selection ratio when comparing different groups. An adverse impact analysis can involve any type of personnel or selection decision. Thus, if I am using a structured interview in managerial selection, and the test results in a significantly higher selection ratio for Whites than Hispanics, I would interpret this result as indicative of race-based adverse impact. The *Uniform Guidelines* define adverse impact as

> The use of any selection procedure which has an adverse impact on the hiring, promotion, or other employment or membership opportunities of members of any race, sex, or ethnic group will be considered to be discriminatory and inconsistent with these guidelines, unless the procedure has been validated in accordance with these guidelines, or the provisions of section 6 of this part are satisfied.
>
> *(Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, 1978, 1607.3)*

Although adverse impact ratios are often cited, along with mean differences, as if they provided evidence of bias, they do not. They may be confused because adverse impact is a term with an attitude problem—a negative attitude forcing adversarial roles. It is "fraught with inferences and implications that there is some kind of inherent biasing characteristic of tests that accounts for different selection ratios among candidate subgroups" and "Instead of selecting a neutral term (e.g., 'pass–fail' ratio), the agencies chose 'impact,' which carries the clear connotation that tests intrinsically have an impelling or compelling effect on candidates from one subgroup" (Lawshe, 1987, p. 493).

Adverse impact can occur for several reasons, of which bias is but one. Other reasons include (a) chance, (b) measurement problems inherent in the test, (c) the nature of test use, (d) differences in distribution sizes, (e) reliable subgroup differences in general approaches to test taking, or (f) true population differences in distributions of the trait being measured. Adverse impact may be said to be due to bias *only* if one or more of the first five of these is shown (except the first, which is not systematic) and if the sixth one can be rejected.

> *Adverse impact* is due to bias only if the groups are truly the same on the trait being measured. If there are true differences between the groups, then adverse impact reflects real differences on test scores and can be defended as a business necessity. For a comprehensive discussion of the controversies involved in adverse impact, as well as the range of expert opinion concerning appropriate analysis, see Cohen, D.B., Aamodt, M.G., and Dunleavy, E.M. (2010). *Technical advisory committee report on best practices in adverse impact analyses*. Washington, DC: Center for Corporate Equality.

A professional analysis of adverse impact involves the following steps:

1. A specific event, decision, or assessment must be identified.
2. A group of applicants must be identified.
3. Identifying the protected group status of the individuals must be possible.
4. Based on the decision-making event, some individuals from the group of applicants will have received a favorable decision and some individuals will have received an unfavorable decision.
5. There must be an appropriate hypothesis or question.
6. An appropriate statistical test must be applied; the most frequently used being the adverse impact ratio, a z test, and the Fisher Exact test.
7. Alternative explanations must be explored (Doverspike, 2014).

The analysis of adverse impact is usually based on the application of one of three possible quantitative indicators. Perhaps the simplest is the 4/5ths or 80% test, which is often referred to as a *rule of thumb*. As described in the *Uniform Guidelines:*

> A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence

> of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.
>
> *(Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, 1978, 4D)*

Thus, in order to calculate the adverse impact ratio, one divides the selection ratio for the protected class by the selection ratio for the nonprotected class. If the resulting ratio is less than 80%, or four-fifths, then adverse impact may be present; we say "may" because adverse impact analysis is never as simple as it may seem. The problems with the four-fifths test, and its interpretation, are well known (Cohen et al., 2010), but one of the issues is that one can avoid adverse impact by simply having a high enough selection rate for both groups. Because of the various issues with the adverse impact ratio, experts often calculate either the z test or the Fisher Exact test.

The z test is simple to calculate and is reported in terms of how many standard deviations of difference there are between two proportions or percentages; it is important to note that this is not the same as a standardized mean difference. The result, expressed in standard deviations, can then be easily compared with the famous standard of 2 standard deviations of difference offered by the United States Supreme Court in *Hazelwood School District. v. United States* (1977). The z test is fairly easy to calculate by hand using a calculator, it is easy to program in Microsoft Excel, and a large number of online calculators are available. All you need are the percentages and the number of people in each group. Thus, the results of applying a z test might be that we determine that there are 3.5 standard deviations of difference between a selection ratio of 15% for minorities and a selection ratio of 30% for the majority group. Because the obtained result of 3.5 standard deviations is greater than the commonly applied value of 2 standard deviations, an expert would conclude that there was evidence of adverse impact. From a practical perspective, the z test has the positive attribute of reporting results in easily interpretable units, as an obtained z of more than 2 standard deviations is viewed as statistically significant, in that it has less than a 5% change of occurring by chance alone. As with any statistical test, the z test is a function of sample size; with large sample sizes, the test is likely to be significant while with extremely small sample sizes it is more difficult to obtain a significant result.

The results from selection decisions can usually be organized into a two-by-two table, where the columns are the outcomes of the decision, for example selected/not selected, and the rows are protected group status, minority or majority group. The resulting contingency table can then be analyzed using a Chi-Square test, but among experts there is a preference for the use of the Fisher Exact test in adverse impact analysis. Standard statistical software can be used to calculate the Fisher Exact test. The results for a Fisher Exact test are not reported as a test statistic, but are instead reported as a probability. If the obtained probability is less than .05, then

TABLE 9.2 Selection Ratios and Adverse Impact Ratios for a Hypothetical Case

Basis for Decision	*Proportion Selected*[a]		*Adverse Impact Ratio*
	Group 1	*Group 2*	
True Ability	.72	.62	.86
Method A	.76	.58	.76[b]
Method B	.67	.67	1.00

Note: From Ironson, G. H., Guion, R. M., & Ostrander, M. (1982). Adverse impact from a psychometric perspective. *Journal of Applied Psychology, 67,* 419–432. Copyright by the American Psychological Association. Reprinted with permission.

[a]Assume all "qualified" candidates are selected.

[b]Adverse impact under the 80% rule.

there is a significant relationship between the selection decision and protected group status, and one could reach the conclusion that the selection decision had resulted in adverse impact for the protected group. Again, as with any statistical test, as sample sizes increase you are more likely to obtain a significant result from the Fisher Exact test.

The calculation of adverse impact, whether using the adverse impact ratio or statistical tests, depends on changing candidate sample characteristics and is, therefore, unstable (Arthur et al., 2013; Lawshe, 1979, 1987). In addition, the analysis of adverse impact ratios does not consider true population differences. Consider Table 9.2, which illustrates the problem. If we knew true ability levels, we would know that Group 1 has a higher proportion of qualified candidates than does Group 2, that is, that selection ratios based on true abilities are truly different, although the impact ratio would be greater than 80%. What we have, however, is two different methods of measuring the ability that give fallible results. Use of Method A results in adverse impact under the 80% rule; use of Method B does not. But is Method B truly superior? Observed selection ratios under either method differ only trivially, yet only Method A implies adverse impact. In fact, it can be argued that Method B adversely affects employment opportunities in Group 1 because it fails to recognize Group 1's greater likelihood of having truly qualified members.

Test Bias as Differential Psychometric Validity

Test bias produces scores with systematically different meanings for people who are alike on the characteristic being measured. To define bias more precisely, the interpretation of test scores is biased for or against members of a group if groups of people matched on the trait measured have different scores because of one or more sources of variance related to group membership. Several features of this

definition merit attention. First, it is the meaning inferred from scores that may or may not be biased, not the test per se, although intrinsic test characteristics may contribute to biased inferences. Second, it is group related. The score of an individual test-taker may be invalid, but bias is only one possible source of invalidity. A score can lead to a wrong inference if the person misunderstands the instructions; bias exists only if the instructions are presented so that many people in the group have a common misunderstanding. Third, the definition requires reason to believe that the groups of people being compared are equal with respect to the trait being measured. A measure of bias that does not disentangle itself from genuine group differences is not interpretable. Finally, the definition places the emphasis on sources of group variances, not on group means. Sources of variance are potentially identifiable. Variance is supposed to be due to the same source—the characteristic being measured—in all groups. Bias exists when other sources of variance influence scores in one group but not in another.

An example offered by Steele and Aronson (1995) is *stereotype threat*—for example, the degree to which Blacks or females are vulnerable to general stereotypes about their abilities and to which that vulnerability affects scores in testing where consequences are important. Recent attempts to find evidence for stereotype threat in the field have met with varied success, suggesting that one should be cautious in linking this phenomenon to subgroup test score differences in employment settings (Cullen, Hardison, & Sackett, 2004; Nguyen & Ryan, 2008; Sackett, Hardison, & Cullen, 2004; Stricker & Ward, 2004).

A definition of bias offered by Cole and Moss (1989) treats it as "differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers" (p. 205). This definition of bias, and its accompanying call for discriminant and convergent validity evidence "within a hypothesis-generating orientation that requires the examination of plausible rival hypotheses" (Cole & Moss, 1989, p. 205), might be called *differential psychometric validity*. Investigation of bias from this perspective includes much more than merely comparing correlation coefficients. Test developers and users must think carefully through the maze of complexities, contradictions, and ambiguities possible in any evaluation of psychometric validity. These requirements are exacerbated when one subgroup is to be compared with all others to decide whether the construct inferences from scores in that group differ from the inferences from scores in the others.

A given research setting may involve several groups, but it is easier to think of bias analysis as comparing them two at a time. The convention in bias analyses refers to a *focal group,* potentially a victim of bias, and a *reference group* (often all others) used for comparisons by any of a variety of statistical tools simultaneously is more directly relevant. Analysis of variance methods have been in use for many years. IRT models, particularly at the item level, are widely recommended and appropriate for the differential validity context. Currently, most psychometricians rely upon one of three methods to investigate internal bias or bias due to psychometric validity. Those methods are (1) simultaneous factor analysis comparing

factor structures in different groups, which can be accomplished using a variety of programs that allow for Structural Equation Modeling; (2) the application of IRT models, the ICC and associated item parameters should be equivalent for the two groups; (3) the application of Mantel–Haenszel tests, which test whether there is an association between protected group status and the probability of a correct response to an item after controlling for the latent trait.

Some litigation has centered on a concept of bias in individual items. Psychometricians prefer the term *differential item functioning* to *item bias* (Holland & Wainer, 1993). Traditional item statistics, such as the proportion of the sample giving a correct answer, are inappropriate for studies of differential item functioning because of their dependence on the trait distribution in the sample studied. Because they reflect group differences in that trait, they cannot disentangle genuine differences from bias. Some litigants, however, have called tests biased merely because of group differences in item pass rates. Drasgow (1987) described out-of-court settlements of two court cases on the basis of this simple-minded item difficulty statistic. In one case, *Golden Rule Insurance Company v. Washburn* (1984), the settlement stipulated that, on future tests, group item difficulties should differ by no more than .15. The second, *Allen v. Alabama State Board of Education* (1985) was more restrictive, specifying a maximum difference of .05.

Item difficulty differences on a widely used test exceeded the .15 maximum of the Golden Rule agreement on 90% of the items when responses of Black males and Black females were compared to those of White males—one of those apparently interesting facts that is without meaning, because genuine group differences in these statistics are confounded with bias. An item response theory method, however, identified fewer than half of the items as biased—and inconsistent in direction of effect; the numbers of items harder for minorities nearly equal the numbers of items easier for them. In short, the canceling effect of these differing directions made the cumulative effect on total test scores very low. With similar findings for other subgroups, Drasgow concluded that no measurement bias existed in total test scores in the six groups studied. This is not an unusual research finding.

Criterion Bias

In criterion-related validation, the criterion should be reliable, valid, and free from third-variable biases. It is amazing how easy that sentence is to write and how difficult it is to accomplish. Reliability is often exceedingly difficult to ascertain for criterion measures; sometimes nothing short of a generalizability analysis will do

it, and often such analyses are not feasible in working organizations. A serious attempt to assess criterion validity may in itself be a way of assessing criterion bias. Evidence of valid measurement of the intended criterion construct is the sort of evidence most appropriate; a major question in psychometric validation is whether extraneous sources of variance influence the measures. If so, and if the numbers of cases allow, it should be possible to determine if the extraneous sources are related to subgroup composition.

Acting on the Findings—Can We Reduce Adverse Impact?

Despite the problems, researchers attempt to analyze for bias or fairness, especially if there is adverse impact. The *Guidelines* continue to call for evidence of differential validity, and professional judgments have to be made, even with flawed data. What should be done when it is reasonable to suppose that test scores are biased against a group? Before anything else, clarify needs. Is the top priority to maximize criterion performance or to avoid even the appearance of discriminatory practice? Is either of these the only priority, or is a balancing trade-off needed? The answer is neither universal nor self-evident; it depends on many things, including the costs of error in the situation at hand. What should be done when adverse impact is present (we discuss some options in the box under Corrective Action Under the *Uniform Guidelines*)? Can we reduce adverse impact? What, if anything, are our options?

Corrective Action Under the *Uniform Guidelines*

***Guidelines* Provisions.** The *Guidelines* recognize adverse impact as prima facie evidence of discrimination, and a discriminatory procedure is treated as biased. Four options are available under the guidelines.

1. *See if the procedure can be justified by law, such as the business necessity argument*. A large body of case law has developed over the years, modifying some aspects of the *Guidelines* and supporting others.
2. *Abandon the procedure*. This eliminates one possible source of discrimination (or of litigation) but it begs the question of how to choose among candidates. Ideally, choices are based on valid assessment with no adverse impact. The ideal is hard to find.
3. *Modify the procedure to reduce adverse impact*. One modification uses compensating procedures so that the bottom line is absence of adverse impact. Another adjusts scores to eliminate adverse impact. These options are no longer available.
4. *Offer convincing evidence of job relatedness*. Valid testing is not discriminatory under the *Guidelines,* although different levels of validity are balanced against different levels of adverse impact.

Personnel decisions often require resolving conflicting values and predictions. Employers and the larger society may have competing objectives. Employing organizations want their personnel decisions to improve, or even maximize, performance and productivity. The larger society wants personnel decisions to increase employment of people who, historically, were excluded systematically from consideration. Inconsistent objectives should be faced frankly and the competing values balanced openly, according to policies and procedures clearly acknowledging required trade-offs. It is, in fact, silly to consider these objectives as competing. Dropping old policies of exclusion and competently assessing the qualifications of all candidates, including those formerly excluded, can yield genuine benefit, both in jobs for those otherwise not considered and in enlarged pools of well-qualified job candidates.

Moreover, hiring people formerly excluded can contribute to overall utility in ways not usually included in the criteria for test validation. Consider, for example, a metropolitan police force in which community political leaders have decreed that police will spend much of their time walking a beat. The objective is to reduce crime and improve the quality of life in the neighborhoods—in part by catching and arresting criminals, in part by a watchful presence, and in part by knowing the people on the beat. Knowing the people implies more than knowing their faces or even their names. It implies knowing the common values and experiences of people in the neighborhood and, beyond that, knowing those of neighborhood leaders. Neighborhoods in a city are diverse, creating a real, not just ideological, requirement for police force diversity. Hiring policies might require hiring to fill gaps in the kinds of community insights currently available. One police class might need overrepresentation of low-income ethnic neighborhoods; another might need recruits who know and understand those with affluence. Hiring should not follow rigidly the traditional top-down policy required in most civil service jurisdictions, even though it would maximize criterion utility.

In short, three concepts are too often confused in arguments related to EEO and personnel decisions: (1) psychometric concepts, which include the reliability and validity of scores; (2) statistical concepts, which include the predictive utility of scores as well as the predictions themselves, and (3) social policy concepts, such as affirmative action. If organizational and social goals are to be met, these concepts must be kept as distinct, well-defined, unconfused, and balanced as possible.

Substantially different score distributions for different groups of candidates, a low selection ratio, and a psychometrically sound predictor with good criterion-related validity, can combine in effect to shut out members of a smaller group with a lower distribution. To whatever extent policy rejects shutting out groups of people, alternative or adjunct procedures may be necessary. Policies should be explicit so they can be debated and their implications thoroughly understood. The alternative or adjunct procedures should be evaluated in terms of their effectiveness in balancing differing policy values. We distrust ideological declarations that a

favored procedure virtually will ensure high validity and very nearly get rid of adverse impact, bias, or general unfairness. Understanding the effect of a procedure on each objective should precede advocacy.

Subgroup Difference Reduction Techniques

There are a number of approaches that have been proposed for reducing the impact of assessments on protected groups. The available options can be divided up into procedures that attempt to reduce subgroup mean differences between groups and techniques that attempt to minimize adverse impact (Arthur & Doverspike, 2005; Ployhart & Holtz, 2008; Sackett, Schmitt, Ellingson, & Kabin, 2001). In this section, we discuss methods for reducing the mean difference on test scores between the minority and majority group (minority and majority are used here in a general sense, and include differences between male and females).

Techniques aimed at reducing subgroup differences are based on the concept that if bias is present in a test score, then the removal of bias will result in higher scores for the minority group. Subgroup reduction methods can be thought of as aimed at improving psychometric validity and, therefore, really involving the improved development and refinement of assessment devices through the removal of variance caused by irrelevant constructs. Commonly used techniques include (1) identifying and removing differential item functioning, (2) increasing test-taker motivation, (3) improving the perception of and reactions to tests by applicants, (4) coaching candidates before the test, (5) using noncognitive predictors such as personality, and (6) changing the medium by which the test is presented or the format of the test (Arthur & Doverpike, 2005). The last option to reducing subgroup differences is also referred to as the "method-change" approach (Aguinis, Cascio, Goldstein, Outtz, & Zedeck, 2009; Ployhart & Holtz, 2008); for example, using the video presentation of situational judgment test items instead of a written format. The difficulty with the method-change approach is in finding a way to change the format of the test without also altering the construct being assessed (Arthur & Villado, 2008).

Adverse Impact Reduction Techniques

Methods that can be categorized as adverse impact reduction typically involve some attempt to alter the nature of the selection decisions either through the weights applied to test items or components, the cut score, or even by adding some type of constant to the test score. Frequently, and for maximum effectiveness, such score adjustments are applied after the test has been given and the scores have been inspected. Adjusting scores to give preference to a group of candidates is not a new idea. Civil service laws have long provided for adding points to the test scores of veterans and many jurisdictions use a type of banding where they select among the top 3 or top 10 scorers. Although such techniques will reduce adverse impact,

Section 106 of the Civil Rights Act of 1991 places severe limits on the use of such adjustments.

Race-Norming in the United States Employment Service (USES). The prohibition is the direct result of controversy over "race-norming" in the use of the GATB in the USES. The GATB was developed as a job counseling and job referral tool. The controversial procedure (as described by Hartigan & Wigdor, 1989) was based on factor analysis and validity generalization (Hunter, 1983). It reduced scores on 10 individual subtests to three aptitude scores: cognitive, perceptual, and psychomotor composites. These scores were added with weights derived independently (based on regression analysis) for each of five job families. The composite scores were expressed as percentile ranks in the applicant's population group: Black, Hispanic, or other. A given score in a group with a lower distribution of scores would have a higher percentile rank than it would if based on all three groups combined. In effect, the separate norms added points to scores of minorities.

Employment Quotas. The USES procedure was denounced in many quarters, including Congress and the executive branch, as a quota system. It reduced adverse impact, and its effect was that of a quota. In an area where the labor market is 20% Black and 10% Hispanic, and nearly all the rest White, a true proportional quota would call for hiring two Black applicants and one Hispanic applicant for each seven White applicants hired. Such a quota need not be filled at random. Many employers have used within-group, top-down selection to avoid adverse impact; applicants are listed in rank order within groups according to their scores, and those hired are the most qualified in their respective groups.

Quotas have long been anathema in U.S. society, where the prevailing view has been that each individual should be considered for opportunities on the basis of his or her own merit. Those who used group norms to fill quotas did so less for ideological reasons than to avoid litigation. It was considered the surest way to reduce adverse impact; moreover, it is the ultimate group parity fairness method. Is its effect on mean performance level detrimental to the hiring organization? There is no strong evidence that it is, although finding people who would admit to having the requisite data may not be easy.

Cutoffs. Selection decisions may be reached by proceeding down a list in rank order fashion—the person with the highest score received the first offer of a job. However, as mentioned in previous chapters, in some cases we may use a cutoff to make decisions that dichotomize the applicant group into minimally qualified versus unqualified, or passing versus failure. As might be expected, adjustments to the point at which the cut is made will have an impact on the degree of adverse impact. In the most extreme situation, if we select everyone, there will be no adverse impact. The problem with adjusting the cutoff score is that it is most likely to result in adverse impact if the cutoff is adjusted after the range of scores has been

subject to inspection and analysis. Such an approach may be seen as resulting in a violation of the *Civil Rights Act of 1991.*

Weight. When we have multiple tests, or even more than one subtest, adverse impact can be reduced by adjusting the weights assigned to the components. Unfortunately, as with cutoffs, the juggling of weights is most effective when applied after data is made available and subject to statistical analysis. Such approaches have been used when approved through a previous court decision or consent decree, but are likely to be seen as a violation of *Section 106 of the Civil Rights Act of 1991.*

Banding. A popular procedure for the reduction of adverse impact has been the use of score bands. Banding involves the identification of ***score bands,*** intervals within which score differences are in some sense trivial, or *ranges of indifference.* Within a band, selection decisions may consider diversity. The procedure is often considered a means of compensating for adverse impact. We have introduced it in that context, but it can serve much broader purposes.

The practice of banding has been needlessly controversial. Arguments favoring or opposing banding are based partly on psychometric grounds (e.g., assessment reliability) and partly on statistical grounds (e.g., statistical significance levels). Framing arguments in these terms is a distraction. The purpose is to transform a raw measurement scale into one that groups unit raw score intervals into larger ones where raw score differences do not matter; within such an interval choices can be based on other, perhaps competing, considerations. In a sense, all score use involves banding. The raw measurement scale has bands, albeit of unit width; for example, statistical analyses consider a raw score of 10 to be the midpoint of the 9.50–10.49 score interval, or band. Use of a cut score creates two bands: one passing, the other not. Most banding is between these extremes. Generally, banding sets score intervals greater than raw scores but smaller than the entire region above or below a cut score. Whatever the size of the interval, the same two problems must be faced: how to define band width, and how to choose within a band where candidates outnumber openings.

Scores contain error. Two people with the same raw scores may differ in ability, and two people of the same ability may have different scores. When faced with two candidates of somewhat (but not dramatically) different scores, it is reasonable to ask whether the scores differ significantly, statistically. The standard error of the difference, S_d, between two scores may be defined as $S_d = S_m\sqrt{2}$, where S_m is the standard error of measurement. A band interval might be defined as 1.96 S_d, within which a score difference is not (at the 5% level) significantly different from 0. A broader interval might be defined as $2 \cdot S_m$. Or one could use the standard error of estimate, S_e, to define an interval of scores within which differences in predicted criterion values might be considered trivial.

The basis does not have to be statistical; a score distribution might be divided arbitrarily into four or five or a dozen intervals with nearly equal frequencies, and these might be treated as bands. Statistical definitions of band limits provide an appearance of scientific objectivity, but appearance masks the arbitrariness involved. Even with statistical definitions, arbitrary choices determine whether band width is based on predictor unreliability or on unreliability of predicted criterion performance and the level of confidence used.

The definition might be based on managerial judgments of how much loss in utility (a band width greater than 1.0 *does* involve some loss) can be sacrificed to other considerations. Managers might agree in conference on a band width they consider about right. They may decide that band width is not constant, making bands narrower (or broader) in the middle of the distribution than at the extremes. All of these are arbitrary, but no more arbitrary than the choice of statistical definitions.

Within a band, decisions can be based on other information. They might be based on information not routinely available, on assessments of traits not part of a general or common predictive hypothesis, or on affirmative action or diversity goals. Contextual criteria, not used in test validation or implicit in the predictive hypothesis, could be considered within bands. Choices could be based on additional assessment; one may have a very desirable selection procedure that is not cost effective if used for all applicants, but is if used only with those applicants within a band. Choice could, of course, be random, but we do not recommend it.

The role of judgment in decision making has already been emphasized, and the quality of judgment is not a unique consideration. Use, or possible abuse, of the judgment opportunity is to be evaluated where it exists. The question of banding is whether the exercise of managerial judgment is a good idea and how well the manager engages in the weighting of information and the process of reaching a final decision. Providing managers with the freedom to make decisions is often necessary, but may also lead to bias and reduced validity.

Final Thought

Discrimination means making distinctions. It is not always pejorative; to call someone a "discriminating person" has a favorable connotation. Assessment procedures are supposed to help their users make distinctions—to discriminate between those with much of a trait and those with less (or those with even more), or to discriminate between those who can do the job acceptably and those who cannot (or who can do it better). The word has an unfavorable connotation when distinctions are based on prejudice, stereotypes, procedures, or policies unrelated to the trait or to the performance it predicts. Such discriminatory (not discriminating) practices, whether the result of a formal test, an interview, or a manager making subjective decisions, are poor organizational policies, and many are illegal.

Discussion Topics

1. Discuss different methods of reducing adverse impact and identify the pros and cons of each approach for reducing adverse impact.
2. Compare and contrast the concepts of bias, fairness, discrimination, and prejudice.
3. In this chapter, we mentioned recent efforts to expand the Civil Rights Act to cover sexual orientation. Laws exist defining other protected classes including the disabled and those over 40. Most recently, arguments have been presented for expanding protections against discrimination to include the long-term unemployed, people with poor credit, those with arrest records, and even those with criminal convictions. What is your opinion of the extension of protections against discrimination in employment to other protected classes? Are there any groups you believe deserve special protections?

Note

1 The definition is somewhat ambiguous and was later cleaned up by Einhorn and Bass (1971). In terms of defining fairness, many selection and assessment professionals prefer the regression definition of fairness offered by Cleary (1968).

References

Aguinis, H., Cascio, W., Goldstein, I., Outtz, J., & Zedeck, S. (2009). In The Supreme Court of the United States: *Ricci v. DeStefano*: Brief of Industrial-Organizational Psychologists as Amici Curiae in support of respondents.

Allen v. Alabama State Board of Education, No. 81–697-N (consent decree filed with United States District Court for the Middle District of Alabama Northern Division, 1985).

Arthur W., Jr., & Doverspike, D. (2005). Achieving diversity and reducing discrimination in the workplace through human resource management practices: Implications of research and theory for staffing, training, and rewarding performance. In R. L. Dipboye & A. Colella (Eds.), *Discrimination at work: The psychological and organizational bases* (pp. 305–327). Mahwah, NJ: Lawrence Earlbaum Associates.

Arthur W., Jr., Doverspike, D., Barrett, G. V., & Miguel, R. (2013). Chasing the Title VII holy grail: The pitfalls of guaranteeing adverse impact elimination. *Journal of Business and Psychology, 28,* 473–485.

Arthur W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93,* 435–442.

Civil Rights Act of 1964, 42 U.S.C. Section 2000e (1964).

Civil Rights Act of 1991, 42 U.S.C. Section 1981A (1991).

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cohen, D. B., Aamodt, M. G., and Dunleavy, E. M. (2010). *Technical advisory committee report on best practices in adverse impact analyses.* Washington, DC: Center for Corporate Equality.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (pp. 201–219). New York, NY: American Council on Education/Macmillan.

Cullen, M. J., Hardison, C. H., & Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89,* 220–230.

Doverspike, D. (2014, October 1). Thoughts on Adverse Impact: Part 1. *Assessment Services Review Blog.* Retrieved from http://asr.ipma-hr.org/2014/10/thoughts-on-adverse-impact-part-1/

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Einhorn, H. J., & Bass, A. R. (1971). Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin, 75,* 261–269.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290–38315.

Golden Rule Insurance Company et al. v. Washburn et al., No. 419–76 (stipulation for dismissal and order dismissing cause, Circuit Court of Seventh Judicial Circuit, Sangamon County, IL, 1984).

Guion, R. M. (1966). Employment tests and discriminatory hiring. *Industrial Relations, 5,* 20–37.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Hazelwood School District v. United States, 433, U.S. 299 (1977).

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hunter, J. E. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery (GATB)* (Test Research Report No. 45). Washington, DC: United States Employment Service, United States Department of Labor.

Ironson, G. H., Guion, R. M., & Ostrander, M. (1982). Adverse impact from a psychometric perspective. *Journal of Applied Psychology, 67,* 419–432.

Lawshe, C. H. (1979). Shrinking the cosmos: A practitioner's thoughts on alternative selection procedures. In P. Griffin (Ed.), *The search for alternative selection procedures: Developing a professional stand* (pp. 1–26). Los Angeles, CA: Personnel Testing Council of Southern California.

Lawshe, C. H. (1987). Adverse impact: Is it a viable concept? *Professional Psychology: Research and Practice, 18,* 492–497.

Nguyen, H.H.D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93,* 1314–1334.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61,* 153–172.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American–White differences on cognitive tests. *American Psychologist, 59*(1), 7–13.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56,* 302–318.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797–811.

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34,* 665–693.

United States Commission on Civil Rights. (1993). *The validity of testing in education and employment.* Washington, DC: U.S. Commission on Civil Rights.

This page intentionally left blank

PART III

Choosing the Right Method

This page intentionally left blank

10

ASSESSING VIA TRADITIONAL TESTS

Traditional Employment Tests, Work Samples, Situational Judgment Tests, Technology, and Global Testing

A ***test*** is an objective and standardized procedure for measuring a psychological construct using a sample of behavior.[1] A test is objective in that responses can be evaluated against external standards of truth or of quality—correct or incorrect, or better or poorer than a standard. Measuring implies quantification. Tests are scored quantitatively, with measurable precision, on numerical scales representing levels of a construct to be inferred from the scores. A *construct*, as we use the term, is a fairly well-developed idea of a trait; most constructs in testing are abilities, skills, or areas of knowledge. Tests use a standardized procedure with the same stimulus component for all test takers.[2]

Standardization refers primarily to controlling the conditions and procedures of test administration, keeping them constant or unvarying. If scores from different people are to be comparable, they must be obtained under comparable circumstances. If people tested in one room have 30 minutes in which to complete a test, and those in another have only 20 minutes, neither the circumstances nor the scores are comparable. Any circumstances of test administration potentially influencing scores should be standardized. More than anything else, it is attention to standard procedure that distinguishes testing from other forms of assessment. The distinction is fuzzy. In this chapter, we describe a variety of procedures for assessing KSAs, ranging from highly standardized tests to assessments with little or no standardization, with no clear line distinguishing tests from other assessment procedures.

Defining a test as a sample of behavior means that the examinee is not passive but does something. In other kinds of testing (e.g., blood tests) the object of measurement sits passively while something is done to it. In psychological tests, the examinee responds to test stimuli by writing answers to questions, choosing among options, recognizing or matching stimuli, performing tasks, ordering objects or ideas, or producing ideas to fit requirements—and this is not an exhaustive list.

When this chapter was written for the first edition, paper-and-pencil tests were still quite common. Since then, technology has had a major impact on testing. If you were to pick up a catalog from a test publisher, you would find that many of their tests are now available only in a computerized form and that paper-and-pencil versions are no longer available. In this chapter, therefore, we will also discuss the interplay between developments in technology and selection. Furthermore, for major corporations, testing is now a global enterprise, demanding that HR professionals consider the impact of language and culture on selection decisions.

Traditional Cognitive Tests

Cognitive tests allow a person to show what he or she knows, perceives, remembers, understands, or can work with mentally. They include problem identification, problem-solving tasks, perceptual (not sensory) skills, the development or evaluation of ideas, and remembering what one has learned through general experience or specific training. For better or worse, the label "cognitive tests" is applied to a variety of instruments including intelligence tests, multiaptitude batteries, a wide range of abilities, achievement tests, and even tests of very specific job knowledge.

At one time such "cognitive tests" were also referred to by critics, often derisively, as "paper-and-pencil" or "multiple-choice" tests. Of course, many different types of tests used a paper-and-pencil or multiple-choice format, or both. Further, as already noted, as a result of technological advances, most tests have migrated to computer platforms, which also allow for a much wider range of options in terms of item formats. Of course, materials do not define traditional tests. Commercial tests of cognitive ability are commonly used and, within 12 minutes or so, can provide reliable scores that predict as well as measures that take hours to administer. One such test, the *Wonderlic Personnel Test,* has been used to screen NFL recruits since 1970.

In Chapter 9, we mentioned a common multiaptitude test battery, the GATB. The abilities or aptitudes measured by the GATB include the following:

- General Learning Ability
- Verbal Aptitude
- Numerical Aptitude
- Spatial Aptitude
- Form Perception
- Clerical Perception
- Motor Coordination
- Finger Dexterity
- Manual Dexterity

As long as a list of common abilities or aptitudes would be, a list of specific job knowledge would be extremely lengthy, if not unlimited. Job knowledge tests do

exist for common jobs or occupations, including police officers, firefighters, informational technology professionals, clerical workers, and manufacturing employees. Standardized content knowledge tests are part of the licensing process for many professions including physicians, lawyers, and psychologists. However, for many types of job knowledge, companies rely upon so-called locally developed or validated tests.

An example item on a test of general cognitive ability:

Which number in the following group of numbers represents the smallest amount?

a. 11 b. 1 c. .111 d. .011

It is almost always cheaper to buy a test than to develop one, but a commercial test may have less face validity than a locally developed test that refers explicitly to specific jobs or sets of jobs within the organization. Job-specific local tests developed by people well-trained in psychometrics can be as reliable and valid as commercially available ones. One study paired three subtests of the *Differential Aptitude Test Battery (DAT)* with related job-specific tests (Hattrup, Schmitt, & Landis, 1992). For example, the *DAT Verbal Reasoning Test,* a measure of the verbal comprehension factor, was paired with a technical reading test based on manuals used on the job. Confirmatory factor analysis showed that the same constructs were measured in each of the three pairs of tests. Hattrup et al. (1992) concluded that test users do not gain much, psychometrically, by building homemade, job-specific tests, even good ones, but that they do not lose anything, either, and may gain considerably in testing program acceptance. No matter how much a test developer tries to make particular tests highly specific to particular uses, general cognitive constructs still account for most of the variance. Those who think they are doing things that are new or highly specific may only be fooling themselves (Murphy, 2009).

Work Samples and Performance Tests

Performance testing in the workplace means assessing proficiency in some aspect of job performance. Performance tests may be cognitive or noncognitive, paper-and-pencil or "hands-on," and anywhere from the most to the least constrained kinds of responses. They may be criteria or predictors intended to predict no further than the immediate future. An applicant who does well on a welding test may be expected do good welding the first day at work; situational variables like equipment, materials, supervision, coworkers, or personal traits like motivational level, may determine whether this good performance continues on the job.

Although prediction is always implied, performance tests are used mainly to assess proficiency, skill, or knowledge at the time of testing—here and now, not at some future time. Unlike low-aptitude candidates, those lacking knowledge or skill may acquire it through special training and reapply when ready. The most common "hands-on" performance tests may be work samples. They are well-established as predictors. Their criterion-related validity is consistently shown in reviews (e.g., Schmidt & Hunter, 1998).

A work sample test is a standard sample of a job content domain taken under standard conditions. Aspects of the work process, the outcome, or both may be observed and scored. In a flight test for a pilot's license, the focus is on process; a check pilot has a checklist of required maneuvers and evaluates how well each is performed. A candidate for an office job may be given a typed manuscript with many scribbled changes on it, be seated at a computer, and told to prepare final hard copy; perhaps only the result is observed and scored. In either case, the work sample is a *standardized abstraction* of work actually done on a job. There are degrees of abstraction. A work sample might be faithful reproductions of actual assignments, sanitized simulations of critical components, or the extreme abstraction—measures of isolated skills used on the job.

Simulations imitate actual work but omit its trivial, time-consuming, dangerous, or expensive aspects. They may imitate a task almost exactly, as in some simulations of aircraft cockpit tasks. They may imitate only the general flavor of reality, as in assessment center management exercises.

Other possibilities carry abstraction still further. Performance tests might use *talk-through* interviews (Hedge, Teachout, & Laue, 1990) to describe the steps, tools used, and decisions made in doing the job. A work diary might be used. A collection of product examples (a "portfolio") may be evaluated. Even a multiple-choice test may abstract from overall performance the knowledge and understanding of processes, tools, and choices that make up performance on the job. Simulations that are not highly abstracted are known as *high-fidelity* simulations; the greater abstractions may be *low-fidelity* simulations (e.g., Chan & Schmitt, 2002).

Work sample development begins with job analysis, although not everything the analysis identifies is included. A complete job analysis identifies a job content *universe*. The part of the universe to be assessed is a job content *domain*. Related assessment possibilities (including scoring methods) make up a test content universe, and the choices among them define the intended test content domain.

Proficiency is the construct measured by a work sample, but it takes many forms. For a criterion, it should identify all tasks critical for overall performance. For selection, it omits critical tasks learned on the job. Ordinarily, tasks defining proficiency should be those that many, but not necessarily all, workers are likely to perform well. Most work samples use only frequent tasks; rarely performed tasks might be in the domain to identify those who can handle unusual job situations.

Equipment or material used should match that actually used on the job. Tolerances and procedures for monitoring equipment should be established; if holes into

which things are inserted get larger over repeated testing, monitoring hole size may be an important aspect of standardization. As always, pilot studies should evaluate the clarity of instructions, scoring procedures, and characteristics of test components (e.g., items) as well as overall reliability and validity of scores.

Scoring Work Samples

Scores are usually ratings. An overall rating of process, product, or component part can be dichotomous (e.g., satisfactory or unsatisfactory) or a scale point. A work sample product might be matched to one of a set of samples previously scaled from *very poor* to *excellent*; the score being the scale value of the sample it most closely matches. More objective measures can be used. A score on machine set up might be the time required to do it. The score can be the pounds of pressure required to break a weld. A computer might count the number of corrections made in a sample word processing task. Ratings predominate, however, and their associated problems (see Chapter 12) can be helped with procedures like these:

1. Job experts should choose work sample content, specify desired performance, and provide at least a preliminary scoring key or protocol.
2. Scorers should be trained to use the protocol: what to look for and how to evaluate specific events or product components.
3. The same performance or product should (if possible) be evaluated by two or more independent observers; impermissible differences in ratings should be defined and the procedures for reconciling differences prescribed.
4. All possible procedural safeguards of reliability should be built into the scoring system.

Situational Judgment Tests

An important challenge when selecting candidates for a position is predicting how a potential employee will respond to important tasks and problems he or she may encounter in the workplace. Some of these problems are difficult, although not impossible, to re-create in a work sample. Situational judgments are low-fidelity simulations of important work tasks, presented in a multiple-choice format. Typically, the situational dilemmas are related to core job competencies, such as responding to irate customers for service-oriented jobs (McDaniel & Nguyen, 2001). Candidates are presented with a series of job-relevant scenarios and a set of possible responses to each situation. They are then asked to indicate which of the responses

they would be likely to employ if confronted with the situation. In this sense, situational judgments are very similar to situational interview questions discussed later.

An example item from the *situational judgment* portion of the FBI Special Agent Selection Process:

You are shopping when you notice a man robbing the store. What would you do?

a. Leave the store as quickly as possible and call the police.
b. Try to apprehend the robber yourself.
c. Follow the man and call the police as soon as he appears settled somewhere.
d. Nothing, as you do not wish to get involved in the matter.

An alternative to the "what would you do?" item stem is one that asks respondents to choose the "best response" to the situation presented. McDaniel and Nguyen (2001) argued that this approach is less susceptible to faking. Ployhart and Ehrhart (2003), however, found that the "would do" approach showed more favorable item characteristics than the "should do" approach (see also Arthur, Glaze, et al., 2014). This is consistent with the idea that intentions concerning what you would do are more predictive of behavior than is knowledge about what you should do (Ajzen, 1991).

It appears that situational judgment tests can be developed that correlate with performance over and above job experience, cognitive ability, and personality (Chan & Schmitt, 2002). Like any other testing method, what is measured by situational judgment tests is dependent upon their construction. Situational judgment tests that are, in essence, proxies for general cognitive ability can be constructed, or tests emphasizing creative problem solving or personality variables can be developed, depending on the content of the scenarios presented and which of the response options is determined to be correct. Therefore, distinguishing between situational judgments as item-development *method* and situational judgment as a *construct* is important. If the goal is to measure good judgment, than this approach holds much promise (Brooks & Highhouse, 2006). But researchers need to focus more on defining what good judgment is and specifying the nomological network of relations to other constructs (M. C. Campion, Ployhart, & MacKenzie, 2014).

Noncognitive Performance

Physical Abilities and Fitness Testing. Measuring strength, muscular flexibility, stamina, and related abilities usually requires equipment and individual testing. Equipment needs described by Fleishman and Reilly (1992) are often simple.

Assessing stamina may use an electronically monitored treadmill with an accompanying electrocardiograph, but a simple step-climbing test can also assess stamina, although with less precision. Physical tests can involve different combinations and methods for assessment but may generally share similar attributes. The two main types are physical ability tests and physical fitness tests. Physical ability tests involve a series of physical activities performed independently such as scaling a wall, running a set distance, or dragging a heavy object, all within a time limit for each activity or for all activities combined (Maher, 1984). Physical ability testing may also include components of other types of tests including physical fitness tests, stamina tests, or wellness tests (Hoover, 1992). A particular type of physical ability test is the physical ability work sample or simulation. Here, job simulation exercises closely simulate actual behaviors required on the job.

Physical fitness tests are defined as using various set exercises to measure different areas of physical strength, agility, and stamina. These tests still present the validation problem of construct validity because these are very indirect measures of on-the-job behaviors based on the general construct of physical condition. In addition, such physical fitness tests usually have significant adverse impact on female applicants (Hoover, 1992).

Physical tests can potentially have adverse impact on protected groups, including women, older applicants, and some ethnic group members, compared with White men. Therefore, evidence of test validity must be present and in compliance with the *Uniform Guidelines* (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978; Hogan, 1991a; Maher, 1984). Gebhardt and associates (Gebhardt & Baker, 2010) have published extensively on the validation of physical ability measures.

Research has supported the use of physical tests because of their validity in predicting future job performance. For example, physical tests have been found to be highly correlated with later job performance ratings for firefighting jobs (Henderson, 2010; Henderson, Berry, & Matic, 2007). Similarly, Campion's (1983) review of physical tests found that maximum performance measures of physical ability were related strongly to maximum performance job samples; however, this relationship was not as clear with on-the-job performance. Psychological research has found support for the use of physical performance tests by using methods such as job analysis to determine that these types of assessments are, in fact, job related (Campion, 1983). Hogan's (1991b) analysis of the dimensionality of physical performance in various occupations and predictability of physical ability tests of job performance revealed three main components of physical abilities: strength, endurance, and movement quality. Classifying the physical performance requirements of a job in terms of these three dimensions can be useful for conducting a job analysis with incumbents and generalize to multiple jobs and fields in determining appropriate performance requirements.

In addition to establishing the predictive validity of tests, studies have looked at applicant perceptions of physical tests. One study found that a sample of police

officer applicants rated physical tests more positively than psychological tests, and interviews were perceived to be the most job-relevant procedure (Carless, 2006, see also Ryan, Greguras, & Ployhart, 1996).

Several suggestions can be made for physical test developers from these results. In order to maximize perceptions of job relatedness, favorability can be increased by orienting applicants to the testing process (through video or in-person sessions), as well as being consistent in the administration of tests and effectively communicating that process to applicants (Ryan et al., 1996). Others would argue that offering physical conditioning or defense training through remedial preparation sessions can bypass adverse impact by helping applicants (especially women) prepare for and pass the physical tests (Maher, 1984). Moreover, physical training programs also have been viewed positively by applicants (Ryan et al., 1996) and may invoke perceptions of fairness and an organization's conscious effort to bring all applicants onto a level playing field.

Sensory and Psychomotor Proficiencies. Work combines cognitive, muscular, sensory, and attitudinal components; a useful work sample might focus on the sensory component. Requisite here-and-now job performance may include sensory proficiency such as correct identification or distinctions of distant shapes, colors, musical pitch, or unseen but touched objects. Except for some classic studies (e.g., occupational vision; see Guion, 1965; McCormick & Ilgen, 1980), little research has addressed the assessment of sensory skill for personnel decisions. Fleishman and Reilly (1992) identified assessment methods for a few sensory abilities; more important, perhaps, they identified some important skills (e.g., night vision) for which no existing measures were identified; these, too, are ripe areas for research.

Psychomotor skills, especially dexterity and coordination, are more widely tested. Especially common is the use of dexterity tests, often requiring examinees to insert pegs or pins in holes. Scores can be the number of pins (or assemblies) inserted within a time period or the amount of time required to fill the board. Examples of tests for other psychomotor skills are provided by Fleishman and Reilly (1992). Commercial psychomotor tests are available, but sometimes manipulations imitating those required on a job should form the test. Job analysis can identify the recurring stimulus patterns and the kinds of coordinated responses required.

High skill levels in some sensory or psychomotor areas may compensate for deficiencies in others, in work as in more general life skills. The compensatory development of unusual auditory skills among the legally blind is one example; the extraordinary skin and muscle sensitivity of the deaf and blind Helen Keller is legendary. Examples need not be so dramatic to have implications for personnel management. Rehabilitation counselors tell about people lacking certain sensory (or motor) skills performing well on jobs many employers would have denied them. Hope for finding compensatory skills is based more on anecdotes than on research. Evidence does not yet lead to general propositions about genuinely compensatory patterns.

Technology and Testing

Technology offers new ways to do conventional testing, as well as the potential for entirely new approaches to unconventional testing. Technology and selection has become a major topic in and of itself. SIOP (Below, 2014) recently reported that the Number 4 Workplace Trend for 2015 was "Increasing Implications of Technology for How Work Is Performed." The Number 2 trend was "Continued Use of HR Analytics and Big Data." Rounding out the list, at Number 1, the top trend was "Mobile Assessments." However, we will start by going back in time a little, to when paper-and-pencil tests were still common and the Internet was in its infancy, and discuss the implications of the computerization of assessments, including the potential offered by computer adaptive testing.

Computerization of Tests

During the initial transition from paper-and-pencil tests to computerized tests, the tendency was to simply adapt a conventional test for computer administration. Thus, tests looks very much like they had since the 1940s or even the 1920s, except they were presented on a computer screen. Testing companies would distribute special disks or keys that would control the administration and scoring of the test on a desktop computer. Gradually, assessments were offered over the Web or Internet, although initially the tests were still very conventional.

Of course, other advantages were soon discovered. Computers allowed for the visual projection of items, visually presented episodes, and even elaborate simulations. Video tests were developed assessing situational judgment in customer service jobs, among others, with gratifying validity coefficients (e.g., Chan & Schmitt, 1997; Dalessio, 1994).

A major advantage of computerized testing was that item banks could be created and calibrated according to stable item characteristics (either those of classical test theory or IRT). Computers could draw items according to specifications to make up unique test forms for each examinee, permitting a large number of psychometrically equivalent forms to be generated from the bank. Item banking, therefore, offers a potential advantage for both test security and the common problem of retesting. Two different candidates might see *some* common items, but item differences would be substantial enough to reduce the test security problems associated, for example, with item memorization.

Computer Adaptive Tests. Conventional testing is also known as *linear testing*; all items are presented one after another to all examinees. A high-ability person flies through the easy items; only hard items show just how able that person is. Linear testing is, therefore, an inefficient use of testing time. The combination of IRT, the massive data storage ability of computers, the computational power of computers, and the flexibility possible in the presentation of items allowed for the use of

computer adaptive tests. Adaptive testing has the advantage of using a branching algorithm and, therefore, fewer items. It begins with one item of moderate difficulty; the next one chosen depends on the response given to the first one—and so on until a predetermined criterion for stopping the test has been reached. If the first item is answered correctly, the next one may be more difficult. If the next one is answered incorrectly, the third item may be between the first and second in difficulty. Adaptive testing long has been used in individually administered ability tests, but it required the combination of modern computers and the development of IRT to bring it to its current level of sophistication.

Adaptive testing can maximize the precision of ability estimation at any point on the ability scale. In personnel decisions, however, precision is important mainly at that part of the scale where most decisions are made. If about 20% of those who apply for a job actually will be hired, and most of those offered a job will accept, precise measurement would not be very important below the 75th or above the 90th percentile. With good item parameter estimates, a brief conventional test can be developed that distinguishes well within that narrow region, but not in the low or very high scores where such differentiation amounts to little more than a nice psychometric exercise. Today, many tests are offered based on various approaches to computer adaptive testing, and the examinee may be unaware of all the computations going into delivering an individualized test to the candidate.

Unproctored Testing

The administration of examination in an unproctored environment is not new. We would wager a guess that teachers have been allowing students the option of take-home tests for as long as schools have existed. The widespread availability of computers and easy access to the Internet, however, has led to a new era in terms of unproctored testing. One can argue that there is a continuum that runs from strict, tight proctoring to totally unproctored testing. Proctoring can be defined as involving the control of the testing environment by a trained, trusted individual so as to (a) limit distractions, (b) verify user identity, (c) monitor time, (d) prevent participation by unauthorized persons, and (e) prevent cheating by the test taker.

The growth of unproctored testing led to concerns over the effect of distractions and the extent of cheating, as well as possible ethical issues. Although we would not recommend the use of unproctored assessments in high-stakes situations, the available evidence suggests that cheating is not widespread with unproctored tests. Although distractions may be greater, and the reliabilities and validities slightly lower, unproctored tests lead to favorable applicant reactions and greatly expand the potential pool of candidates at very low cost (Tippins, 2009).

Not only is unproctored testing here to stay, but also many job candidates are now completing their applications and taking tests on smartphones (Arthur, Doverspike, Muñoz, Taylor, & Carr, 2014; Tippins, 2009). The widespread use of mobile devices has led to a whole new area of research related to technology. A

preliminary conclusion would be that mobile devices exacerbate any disadvantages of unproctored assessment. For example, given the fundamental characteristic of being mobile, smartphones should increase the potential number of distractions and increase the variance in possible environmental conditions.

Simulations, Games, and Gamification

Simulations and games are not new. They have been around since IO psychology began and long have been an integral part of assessment centers. What is new is that technology has improved the options for and quality of simulations and games (Tippins, 2009). Where once high-fidelity simulations and games were quite expensive, computers allow the average individual to have elaborate simulators, for example driving or flight simulators, right on their desktop or mobile device.

The literature commonly distinguishes between low- and high-fidelity simulations. Fidelity can also be evaluated in terms of physical or psychological fidelity. The existing research suggests that even low-fidelity simulations can be valid predictors of job performance.

Both simulations and games can generate tremendous amounts of data. One of the issues is how to deal with the tremendous amounts of information potentially available from extended simulation or game play. This has led to attempts to integrate the simulation, game, and big data literatures (for more on big data see Chapter 11). In the next few years, we would expect to see much greater use of highly complex simulations and games.

The popularity of games has led to the introduction of game-type elements into traditional tests, which is referred to as *gamification*. For example, many consultants now add game elements to other tests such as situational judgment instruments or cognitive assessments. Or a personality inventory may be administered through a series of e-mails delivered to the job candidate as part of an in-basket exercise. Gamification is also used as a way of emphasizing and reinforcing the organization's brand.

Global Testing

Increasingly, organizations are multinational and compete in a global marketplace. Assessment professionals have to expand their skill set in order to become adept at global assessment, which involves the standardization of selection systems across countries.

Legal Issues

In Chapters 4 and 9, we discussed legal issues that confront selection specialists working in the United States. Working globally requires knowledge of legal concerns in other countries. The definitions of protected classes or minority groups

will vary substantially across countries. Many countries, especially in Europe, have much more restrictive privacy laws. There is substantial variance across countries in what is viewed as ethical behavior. The "rights" of workers, the role of labor organizations, and the standards for guilt and innocence will all vary. As complex as complying with the United States laws and regulatory agencies may be, the difficulties are magnified when practicing globally.

Translation and Equivalence Issues

The demands of translating a test into multiple languages is one of the challenges faced by cross-cultural psychologists. Over the years, various approaches to handling translation issues have been developed and refined. Mere translation is not the simple matter it would appear. Literal translations, even if possible, may not have the same psychological meaning in two languages; score equivalence is unattainable with literal translation. Not only may constructs have different meanings across cultures, words for some constructs may not even exist in certain languages.

Translation by "centering" (getting the gist of the meaning) and acceptable back-translation into the original language seems to give equivalent meaning, but that does not ensure equivalence in inferences from scores; centering may change psychometric properties dramatically, including constructs measured. Cultural differences can influence scores and their interpretation at least as much as language differences. Cross-cultural testing faces at least three kinds of problems: differences in approaches to tests, problems of test administration, and score equivalence.

Two psychometric considerations should govern test translations. First, test item parameters must match in the original and translated versions. Item matching is best done by IRT. Perhaps not every item would be translated to achieve precisely the same parameters in a 3-parameter model, but the distributions of item parameters could be kept comparable. Second, the two versions should be pretty much equally valid measures of the same constructs. Do various antecedent and subsequent correlates behave similarly? Do both versions escape the same contaminating sources of variance? Positive answers say that the tests are measuring the same constructs.

Instead of translating a test developed in one country into the language of another, Schmit, Kihm, and Robie (2000) described the development of a "global" personality measure. The idea behind their approach was to develop the measure globally, beginning with item writing, through to item translation and data analysis. Alternatively, multinational companies can treat operations in each country as independent and develop locally valid assessment procedures. With this option, the entire test development process can take place within the culture, cultural factors influence construct definition, item writing, instruction development, and all of the developmental research. This option makes sense only if "home country" and local personnel are not competing for the same opportunities, such as promotion to a specified position. Whether a construct important to performance in one culture is also important in another is a problem cross-cultural staff needs to address.

Tests and Controversy

Testing, and personnel assessment generally, is and has been controversial. There are controversies among psychometrically trained experts, among people trained in different test-using disciplines, between psychometric professionals and people outside of these professions, and in society generally. In the face of all the fuss, it is strange that testing remains an important basis for so many kinds of decisions. Few people would want to get rid of various kinds of licensing exams, despite their sometimes serious deficiencies. The cry for educational proficiency exams has been translated into law in many states. Government civil service procedures using merit examination concepts grew out of disenchantment with less objective bases for selection.

In the face of controversy, it is useful to remember that tests have compiled a good track record. They have successfully predicted performance on jobs and other kinds of criteria as well. Put together in a battery of tests measuring different things, groups of tests have even better records.

Tests are good, tests are useful, but tests are imperfect. Perfection cannot be reasonably expected; too many other things influence criteria for test scores to predict them perfectly. Even so, there is room for improvement. Many things we do well with tests can be done better and with greater understanding. Things we do not do so well with tests provide still greater challenges. The search for new and better ways to measure candidate qualifications, and for new and better definitions of the nature of the qualifying traits, should go forward. However, a lot of bright new ideas, once thought promising, have been tried and have withered. Psychometric history is strewn with the remnants of once-grand new ideas. Many tests that were supposed to measure more important constructs than those traditionally measured have gone out of print with only negative findings resulting from their use. Item types once hailed as panaceas have left the scene in ignominious defeat. Enthusiasm for new ways, commendable as it is, is no substitute for data.

Discussion Topics

1. How do you feel about the use of traditional cognitive tests in the selection of candidates for jobs? Should intelligence tests be used to select physicians? Salespeople? Janitors? Football players? Why or why not?
2. For what type of jobs would a work sample be appropriate? Where would it not be appropriate?
3. What do you think of the gamification of tests? Have you taken a test disguised as a game? If you know a game is a test, is it still fun or does it change the way you view the test?
4. How do you feel about organizations that build brand or advertising messages into their employment tests?

Notes

1 This is the more narrow, and probably more common, definition of the term "test." However, as noted elsewhere, under the *Uniform Guidelines* any assessment procedure, in fact, any personnel decision no matter how subjective and unstructured, can be judged using the standards for a good and valid test.

2 "Stimulus component" may, but does not necessarily, mean "item." Computer adaptive testing, and some other test procedures do not require the same set of items for all takers.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior & Human Decision Processes, 50,* 179–211.

Arthur, W. A., Jr., Doverspike, D., Muñoz, G. J., Taylor, J. E., & Carr, A. E. (2014). The use of mobile devices in high-stakes remotely delivered assessments and testing. *International Journal of Selection and Assessment, 22,* 113–123.

Arthur, W. A., Jr., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology, 99,* 335–345.

Below, S. (2014). New year, new workplace! SIOP announces top 10 workplace trends for 2015. *Society for Industrial and Organizational Psychology, Inc.* Retrieved from http://www.siop.org/article_view.aspx?article=1343

Brooks, M. E., & Highhouse, S. (2006). Can good judgment be measured? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 39–55). SIOP Frontier Series. San Francisco, CA: Jossey Bass.

Campion, M. A. (1983). Personnel selection for physically demanding jobs: Review and recommendations. *Personnel Psychology, 36,* 527–550.

Campion, M. C., Ployhart, R. E., & MacKenzie, W. I., Jr. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance, 27,* 283–310.

Carless, S. A. (2006). Applicant reactions to multiple selection procedures for the police force. *Applied Psychology: An International Review, 55,* 145–167.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143–159.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15,* 233–254.

Dalessio, A. T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology, 9,* 23–32.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290–38315.

Fleishman, E. A., & Reilly, M. E. (1992). *Handbook of human abilities: Definitions, measurements and job task requirements*. Palo Alto, CA: Consulting Psychologists Press.

Gebhardt, D. L., & Baker, T. A. (2010). Physical performance. In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment: Evidence-based practices for selecting and developing organizational talent* (pp. 165–196). San Francisco, CA: Jossey Bass.

Guion, R. M. (1965). *Personnel testing.* New York, NY: McGraw-Hill.

Hattrup, K., Schmitt, N., & Landis, R. S. (1992). Equivalence of constructs measured by job-specific and commercially available aptitude tests. *Journal of Applied Psychology, 77,* 298–308.

Hedge, J. W., Teachout, M. S., & Laue, F. J. (1990). *Interview testing as a work sample measure of job proficiency*. AFHRL-TP-89–60. Brooks Air Force Base, TX: Air Force Systems Command.

Henderson, N. D. (2010). Predicting long-term firefighter performance from cognitive and physical ability measures. *Personnel Psychology, 63,* 999–1039.

Henderson, N. D., Berry, M. W., & Matic, T. (2007). Field measures of strength and fitness predict firefighter performance on physically demanding tasks. *Personnel Psychology, 60,* 431–473.

Hogan, J. (1991a). Physical abilities. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 753–831). Palo Alto, CA: Consulting Psychologists Press.

Hogan, J. (1991b). Structure of physical performance in occupational tasks. *Journal of Applied Psychology, 76,* 495–507.

Hoover, L. T. (1992). Trends in police physical ability selection testing. *Public Personnel Management, 21,* 29–40.

Maher, P. T. (1984). Police physical ability tests: Can they ever be valid? *Public Personnel Management Journal, 13,* 173–183.

McCormick, E. J., & Ilgen, D. R. (1980). *Industrial psychology* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103–113.

Murphy, K. R. (2009). Content validation is useful for many things, but validity isn't one of them. *Industrial and Organizational Psychology, 4,* 453–464.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11,* 1–16.

Ryan, A. M., Greguras, G. J., & Ployhart, R. E. (1996). Perceived job relatedness of physical ability testing for firefighters: Exploring variations in reactions. *Human Performance, 9,* 219–240.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schmit, M. J., Kihm, J., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology*, *53*, 153–193.

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology, 2,* 2–10.

11

ASSESSING VIA INVENTORIES AND INTERVIEWS

Self-Report Personality and Other Inventories, Biodata, and Unstructured and Structured Interviews

Testing and scaling are two basic psychometric procedures; other kinds of assessment procedures are derived from one or both of these approaches. Some methods are derivatives of both, developed like tests and using rating scales in scoring. Others evolved from the two psychometric foundations and also from forms of assessment that developed outside of the psychometric tradition. Commonly used approaches to assessment, derived both from testing and scaling traditions, include inventories and interviews.

Inventories

Inventories are usually self-report measures of interests, motivation, personality, and values; personality tests and similar inventories can be viewed as self-ratings on some attribute. Most inventories are developed using test construction principles and, like tests, are scored by summing scores for item responses. Unlike tests, responses are based on opinions, judgments, or attitudes, not on externally verifiable information. Responses may be dichotomous (e.g., agree or disagree), multiple choice, forced choice, constructed response (as in sentence completion tests), or on rating scales with three or more levels (e.g., agree, uncertain, disagree).

Varieties of Inventories

Checklists. Lists of words or phrases can be assembled, and people can be asked to check those that describe them and leave blank those that do not. Items might be chosen to fit a theory. Alternatively, panels of experts may judge whether an item fits a designated trait, and a decision rule (e.g., 80% agreement or more) may be set for retaining items.

Scaled Response Inventories. Choosing from three or more categories in an ordinal sequence is a response. The *Minnesota Multiphasic Personality Inventory* may be

the oldest of these still in use; its response options are "true," "false," and "cannot say." Such a scale amounts to little more than a dichotomy with an escape clause. Many commonly used scales have more categories, such as a 5-point scale ranging from low to high in appropriateness as a self-description of the respondent.

Multiple-Choice or Forced-Choice Instruments. Many inventories are multidimensional; items may have multiple response options each reflecting a different construct (e.g., *Sixteen Personality Factor Questionnaire*). Options may be responses to a

Figure 11.1 contains several sentence stems from a sentence completion form. For fun, you may want to finish or complete the sentence. Many projectives, including sentence completion items, are scored based on the extent to which they contain reference to various needs. Three of the more common needs are Need for Affiliation, Need for Achievement, and Need for Power. After you complete the sentences in Figure 11.1, you may go back and try to determine whether the content of your sentence corresponds to a positive or negative need for affiliation, achievement, or power.

Instructions: After reading the stem, complete each sentence. Do it as quickly as you can based on your initial feeling after reading the stem.	
Stem	Rest of Sentence – Write or type in your response.
Example: My brother:	Is my best friend in the whole world.
Today, after class I will:	
When I retire:	
If I won the lottery:	
My favorite clothes are:	
My mother:	
My boss:	
Money:	
My hometown:	
My friends:	
Studying:	
In my free time I:	

FIGURE 11.1 A sample sentence completion inventory.

question or simply sets of words or phrases arranged in sets of three or four from which respondents choose one that is the most (or least) descriptive.

Alternatives to Inventories. Common alternatives for personality assessment were (and are) *projective* techniques. These consist of ambiguous stimuli ranging from inkblots and vague pictures to cartoons and picture arrangement tests to sentence completion forms.

Most projective devices do not measure specific traits, making psychometric validation difficult. They are based on the idea that a person will "project" his or her own personality characteristics on an ambiguous stimulus. The usefulness of the tests for assessing personality continues to receive lively debate (see Lilenfeld, Wood, & Garb, 2000; Viglione & Hilsenroth, 2001), but there is meager evidence for their usefulness in making inferences about occupational success (Highhouse, 2002).

Validity of Personality Inventories

In a recent survey of members of the Society for Human Resources Management (2012), only 18% of HR professionals reported using personality inventories in any capacity. There continues to be a lack of agreement among even psychologists about the predictive efficacy of personality tests for employment decisions (e.g., Morgeson et al., 2007a, 2007b). Nevertheless, well-developed personality measures have been shown to predict a wide range of work-relevant outcomes (Hough & Oswald, 2008; Judge, Klinger, Simon, & Yang, 2008; Ones, Dilchert, Viswesvaran, & Judge, 2007). In addition, they are a particularly valuable source of incremental validity (Schmidt & Hunter, 1998). This is because personality inventories are generally uncorrelated with cognitive ability measures.

Personality inventories have been found to be useful especially for early identification of leadership (Bentz, 1967; Sparks, 1990), and validities for service and sales jobs are particularly strong (Sitser, van der Linden, & Born, 2013; Vinchur, Schippmann, Switzer, & Roth, 1998). Research suggests that phrasing personality items in a work context (e.g., "I strive for excellence in my work") may enhance their usefulness (Bing, Davison, & Smothers, 2014; Shaffer & Postlethwaite, 2012). One concern with this approach, however, is that it may encourage socially desirable responding. The issue of response distortion is discussed next in more detail.

Distorting Responses

When applying for a job, people like to make a good impression. Sometimes they are not very truthful in describing themselves, deliberately faking to make a favorable impression. Sometimes people are not truthful because they lack real insight into their own behavior.

Faking. A *response set* (also called *response style* or *response bias*) is a tendency to follow a particular habit in responding to such stimuli as inventory items. A very common example is a ***social desirability*** response set, the tendency to say things one thinks others want to hear, or the tendency to try to look good to other people. Candidates for a job usually want the job; they are motivated to present themselves favorably during interviews, when taking tests, or when completing inventories. A social desirability set can slip into a deliberate attempt to look good known as *faking*. Faking has been a particular concern in employment offices. It would not be remarkable if an applicant for a position requiring much alertness were to respond "no" to the question, "Do you daydream frequently?" An applicant for a sales position is unlikely to say "yes" to the question "Do you dislike talking to other people?"

Some inventories have special scales to try to detect faking. The *Minnesota Multiphasic Personality Inventory* has a "Lie Scale." Hough and Tippins (1994) had a similar scale they called "Unlikely Virtues." With high faking scores, and maybe very low ones as well, one loses confidence in inferences drawn from personality scores. Often scores on faking are used for score adjustments on the trait scales, but the adjustments rarely enhance prediction of job performance (Goffin & Christiansen, 2003). If a candidate understands the demands of a job well enough to fake appropriately, it is quite possible that on-the-job behavior will be appropriate, regardless of the person's behavior away from work.

James (1998; James, McIntyre, Glisson, Bowler, & Mitchell, 2004) presented an innovative approach to personality assessment based on *conditional reasoning*. The notion is to reduce faking by indirectly measuring unconscious cognitive biases that people rely on to justify or rationalize their behavior. Individual differences in these biases are assumed to relate to different motives or traits. For example, an item measuring aggression might present respondents with a list of reasons for why Americans prefer foreign cars. A more aggressive respondent might prefer a reason that describes American car makers as greedy and unconcerned with quality. A less aggressive personality might choose a more innocuous reason. Put simply, respondents with different motives are assumed to pick different solutions to the reasoning problems. Although this approach has attracted a lot of attention among personality researchers, it is still too early to say if it has promise for assessment in employment settings.

Another attempt to reduce faking uses *forced-choice* items, those in which choices must be made between equally desirable or favorable options. Bernardin (1987) presented a forced-choice method for measuring job-related discomfort. The logic of the method is that everyone has things that they dislike, but some people's dislikes are more job related than others. Thus, Bernardin and his colleagues (Bernardin, 1987; Villanova, Bernardin, Johnson, & Dahmus, 1994) developed items that put job-related discomforts (e.g., sitting for long hours) up against everyday discomforts unrelated to the job (e.g., standing in long lines). The logic is that applicants who repeatedly choose the job-related discomforts as *most* uncomfortable are

less likely to fit with the job in question. This is a very promising approach that remains underresearched.

Several personality inventories have used forced-choice principles. Some offer a choice between equally attractive alternatives, each assessing a different trait. Some offer a choice between equally attractive alternatives for the same trait but differing in item discrimination indices, somewhat like forced-choice performance ratings. The logic is that social desirability contributes no variance to the trait scores. This logic has not worked as well as anticipated in practice. The most effective way to discourage faking of personality inventories simply may be to warn test takers against it (Dwight & Donovan, 2003).

Acquiescence. The tendency to accept or agree with an item regardless of what it says, the *acquiescent response set,* has been well documented (e.g., Jackson & Messick, 1958). Suppose a set of positively stated inventory items were rewritten in a second form as negatively worded statements. Responses to the positive and negative forms should logically be negatively correlated. Agreement with any positively worded item should ordinarily predict disagreement when the item is reversed and worded negatively, for example, when the positively worded item is "I like my job" and its negatively worded counterpart is "I do not like my job." A person who agrees with the first statement is generally expected to disagree with the second. However, for many inventories, such reversals of item content often result in the same responses for both, and over several items and several people, the correlation of scores is positive, not negative. That is, no matter whether the item is worded in one direction or its opposite, people tend to respond in the same way—to acquiesce, however it is worded.

Applicant Reactions

An applicant for a cashier position is unlikely to object to an employment process using an arithmetic test including items on addition and subtraction. The same applicant may be offended bitterly, however, if the process includes an inventory intended to measure trustworthiness or asks questions about one's religious background. The example shows one of the kinds of reactions that concern people using inventories for employee selection: the "what business is this of yours?" reactions. Civil liberties and civil rights groups are wont to support offended applicants, contending that many interest and personality inventories contain material that is prurient, illegal, and an invasion of privacy.

People may feel offended by personality inventories for other reasons. Certain words, some more than others, may be offensive to some people and especially to some groups of people. Some research suggests that the problem may not be large. There was much concern in the 1980s about integrity tests (which are better described as inventories) and the reactions of those who take them. However, Ryan and Sackett (1987) found that participants in an experimental trial generally (with a few exceptions) considered integrity testing an appropriate management tool. In

general, there appears to be no differences in reactions to selection procedures on the basis of sex, age, or ethnic background (Hausknecht, Day, & Thomas, 2004).

Applicant reactions to tests can be a problem when candidates see no relevance of the inventory items to the job sought. Some very personal questions may be relevant to some jobs, and the relevance may have been verified by competent research. Perhaps candidates should be told in advance that some questions may seem irrelevant to the job but have been shown to differentiate between those who succeed and those who fail. It seems likely that a candidate who wants a job, and is given the courtesy of an explanation of an inventory's relevance to that job, will be less likely to take offense at individual items. If so, such courtesies may further safeguard validity.

Although applicant reactions remain a very popular area of research, their influence is likely very minimal when compared with all of the other factors that go into choosing a place to work. It is also doubtful that reactions to a selection procedure are likely to have long-term effects on things such as job satisfaction, organizational commitment, or job performance (McCarthy et al., 2013).

Employer Reactions

Researchers are just beginning to turn their attention toward employer reactions to selection devices (e.g., Chapman, & Zweig, 2005; Furnham & Jackson, 2011; König, Klehe, Berchtold, & Kleinmann, 2010; Lievens, Highhouse, & De Corte, 2005; Nowicki & Rosse, 2002; van der Zee, Bakker, & Bakker, 2002). Lodato, Highhouse, and Brooks (2011) found, for instance, that the majority of professional members of the Society for Human Resources Management believed that one can learn more from an informal discussion with job candidates than from scores on standardized measures. Dipboye (1997) noted that resistance toward the adoption of personnel assessment technologies may be caused by basic desires for autonomy or power (e.g., Nolan & Highhouse, 2014). Other factors may include perceptions about a technology's diffusion in the field or its potential for negative applicant reactions (König et al., 2010). Our experience suggests that people also may not believe that data-based assessment practices work for predicting work performance (at least not in their own "unique" organization). Understanding employer resistance to change is an important frontier for selection research.

Personal History Assessment

The best predictor of future behavior is past behavior—a cliché, to be sure, but generally true. Students who skip a lot of lectures in one semester are much more likely to skip lectures the next semester. A candidate who performed well on a job in the past is likely to perform well on a similar job in the future. The assessment problem is to learn about and evaluate past behavior of candidates. An internal candidate might be known by others in the organization. In an earlier, less litigious era, one

could learn about an outsider's past behavior from reference checks; such queries now produce little more than verification of dates of enrollment or employment, if that. Candidates can be asked about their own past behavior, performance, or experience. Whatever the source, the first problem is to get information that is neither distorted nor unreliable. The second problem is to turn the information into a useful assessment. Information can come from answers to questions of limited scope, whether asked of candidates (the usual way) or others who have known them. It can become an assessment method by treating answers like inventory responses.

Weighted Application Blanks

In many organizations, scoring keys were developed for what became known as *weighted application blanks*. Several of these were developed and described in publications in the 1950s and 1960s; a variety of methods for assigning weights to responses was described in Guion (1965).

The use of weighted application blanks has waned, but they remain a useful method for trying to understand and reduce turnover for lower level jobs. However, today many applications are filled out online. This allows for the use of complicated scoring algorithms, some of which rely upon big data techniques. Another option is the use of text analysis software to scan résumés or applications for key terms or phrases. For example, the software may have been trained to look for applicants whose résumés or applications contain words that suggest that they have been active video game players.

Biodata

Biodata includes items about prior events or behaviors, but is it a biodata item or a personality inventory item if it asks about prior feelings or attitudes? An item such as "How did you feel when . . . ?" may be found in either type of inventory. There is a substantial overlap in the kinds of constructs measured with biodata and those measured by personality inventories, but there are differences, too. Both reflect personality attributes, but biodata is the larger domain, reflecting interests, attitude, skills, and abilities in a single set of questions. If a biodata form includes all of these, the meaning of its scores is obscure. Too often, users do not worry about understanding scores; a good validity coefficient satisfies them. The meaning of scores matters when trying to explain or understand the validity coefficients. What makes biodata predictive? What constructs does it measure? Answers to such questions are especially elusive for biodata, so defining the boundaries of biodata content may be useful.

> Biodata instruments are often referred to as *Biographical Information Blanks*, or by the abbreviation BIBs.

A guide to the boundaries of biodata was provided by Mael (1991). According to Mael's taxonomy (1991), biodata items:

- Must be *historical,* with the items referring to events or experiences that have taken place in the past (and in some cases, are continuing to occur). Intentions, or presumed behavior given hypothetical circumstances, are not biographical and, therefore, are outside the boundary. An example item would be "How old were you when you first learned to drive?"
- Are *external* actions. They may involve others. They may be observable by others. They do not involve events solely within one's own head. This restriction seems not to be followed widely; many forms identified as biodata forms ask questions of the "How did you feel?" variety. However, an example item would be "How many times have you served as the president of a group or organization?"
- Are *objective* in the sense that there is a factual, not interpretative, response. It follows that it should be firsthand information, not attributions to others. An item like "I think my parents were disappointed in me" lies outside the domain on both counts. It attributes to others (the parents) attitudes they may or may not have held, and it probably does so because of subjective interpretations of words, facial expressions, or actions—or false memories. An example of an objective item would be "When was the last time you received some type of communication, for example a phone call or an e-mail, from an aunt, uncle, or cousin?"
- Are *discrete* actions or events that have beginnings and endings; a driver's license was in fact obtained (or not) within a time period. By asking for discrete information, the recollection task of the respondent is simplified. There is also the possibility, even if remote, that someone might know or can find out whether the answer given is correct. *Verifiable* answers, even if no one is likely to take the trouble to verify them, seem less likely to be faked. Example items might include "Did you take Introduction to Psychology in College?" and "What grade did you receive in Introduction to Psychology?"
- Must not ask people about things over which they had no *control.* Past experiences that have shaped and influenced present or future behavior are within the boundaries; even if the experiences themselves are beyond the person's control, reactions to the experiences are controllable. Items with specific historical inequalities in accessibility seem inherently discriminatory, such as experience opportunities that traditionally have been closed to females or to certain ethnic minorities. For example, the item "What position did you play on your high school football team?" could be considered to be discriminatory toward females, because it is unlikely they played high school football.
- Should be seen as *relevant to the job* sought, or the nature of their relevance should be clearly explained; they should have face validity. Items appearing

irrelevant to the job are not likely to be very effective even if within biodata boundaries. Again, this seems to be a rule that is frequently broken in the construction of biodata instruments. For an insurance salesperson, a job-relevant item would be "In the last five years, how many times have you met or exceeded your sales quota?"

- Should be *noninvasive*. As a matter of ethics, empathy, and good sense, the boundaries should draw the line excluding background actions or events people are likely to consider none of an employer's business. Some topics are more acceptable than others in a biodata questionnaire, and some topics are more acceptable for some purposes than for others. An invasive item would be "In the past five years, how many serious illnesses have you had?"

Developing Biodata Forms

Biodata items, like others, can be found by plundering forms used by others. Imagination will add a few more, the whole set can be given an empirical trial, and those with "good" item statistics can form the "new" questionnaire. This unpleasant procedure is fairly typical, but the result can be pleasant; Reilly and Chao (1982) found biodata validity coefficients on par with those of standardized tests. Nevertheless, such biodata forms are criticized as excessively empirical, with no clear understanding of what is measured or why it might be working. The alternative is to specify a construct (or several) to be assessed, to develop its theory or rationale, and to generate systematically the kinds of items believed to tap it (e.g., Breaugh & Dossett, 1989).

Efforts to enhance both prediction and understanding begin by clarifying the measurement purpose. For selection, transfer, or promotion, this begins with job analysis. For training and development purposes, it may begin with a diagnostic analysis of problems. Dean, Russell, and Muchinsky (1999) offered further points of departure based on personality and vocational choice theories and suggested procedures for generating items for the constructs identified as likely to be predictive. In theorizing about what trait may account for the predictive validity of biodata measures, the authors dusted off the old term "moxie" to describe a kind of personal resiliency in the face of negative life events. Mael (1993) described a procedure in which biodata items were mapped on to common personality traits. This approach, dubbed "rainforest empiricism" enables the employer to use biographical items instead of personality items to measure personality traits. More recently, Taylor, Pajo, Cheung, and Stringfield (2004) described the development of reference check items that map on to Big Five personality traits. This approach avoids self-report distortions by having referees, such as former employers, assess the candidates' personalities. Where biographical data is concerned, a combination of data and thought surely is superior to either thoughtless empiricism or naïve theorizing.

Big Data

The big data movement in organizations involves taking large volumes of data and creating narratives. Such narratives include the discovery, reported by Orbitz CEO Barney Harford, that Mac users book more expensive hotels than PC users (Mattioli, 2012). Orbitz uses information like this to decide which hotels to recommend to users searching on Mac versus a PC. Credit card companies purportedly have found that customers who buy antiscuff pads for their furniture are more likely to make timely payments on their monthly balances (Shaw, 2014). Big data findings like these are curious and provocative and whet the organizational appetite for more narratives that will allow more accurate prediction of what consumers will buy and who is a greater credit risk.

One of the hot trends in personnel assessment and selection is the use of big data and predictive analytics. Although what exactly constitutes "big data" is somewhat arbitrary, what is clear is that much of what is labeled "big data" is extrinsic data on individuals. That is, in some respect it is a case of old wine—biodata and personal history information—in the guise of new bottles—big data. For instance, in 1922, Grace Manson discovered that existing job application data on thousands of insurance agents could be used to distinguish those agents who later succeeded from those agents who later failed. In 1935, Albert Kurtz found, studying over 10,000 insurance agents, that personal history items found on applications blanks (e.g., occupation, number of dependents, living expenses) discriminated with a high degree of accuracy the more successful agents from the less successful ones. Baier and Dugan (1957) found that the amount of life insurance owned by State Farm Insurance agents at the time of hire was the best predictor of their later success in sales. Although these big data examples may seem like small potatoes today, they certainly fit the criteria for big data prior to the electronic calculator! This is not intended to diminish the potential importance of big data and predictive analytics; we will have to see what happens in the near future, but it is to serve as a caution that at least when used for selection purposes, big data represents a selection test, and, therefore, must be demonstrated to meet the criteria for a good test.

Interviews

Judgments are made during interviews, whether formally recorded as ratings, and judgments include assessments, predictions, and decisions. These judgments are often intuitive and haphazard. Assessment may be no more than "sizing up" an interviewee. Prediction may be no more than a vague hunch that the person sized up will, if hired (or retained, promoted, or whatever), be great, not be bad, or just not work out. Assessments are often secondary to decision; some interviewers want only to reach a decision and then get on with other matters. The *Watson v. Fort Worth Bank & Trust* case (1988; see Chapter 4) affirmed that interviews intended for personnel

decisions *are* psychometric devices, are based on assessments, and should be evaluated by rules applied to other psychometric devices. Moreover, decision making with no concern for quality of assessment and prediction simply is irresponsible.

Researchers often refer to "the" interview as if all interviews were alike. Just as there are many different tests, there are many different interviewers, looking for many different things, and using many different methods. Some are entirely unplanned; others are as tightly structured as any test. Assessment is the avowed purpose of some; it is a hidden purpose in others. Some are short; some are long. Some use one interviewer; others use panels. Some are done by highly skilled interviewers; others are done by people who do not have a clue to useful procedures. Interview content consists partly of the questions or tasks posed and partly of the medium, the individual interviewer. Interviewers are not as standardized as questions; the same questions can be asked in different ways by different interviewers. Stimulus content consists partly of the attitudes interviewers present or the interviewee perceives.

Interview Research Reviews

Interviews have been considered too unreliable to be valid since Hollingworth (1923) reported rank orders assigned to 57 candidates by each of 12 sales managers—with virtually no agreement. A 20-year series of narrative reviews consistently identified unreliability as a major problem. Not until Schmitt (1976) was much said about the lumping together of data from interviewers varying in skill. Early reviewers also tentatively proposed that **structured interviews**, those with preplanned procedures and sets of questions to be asked, would be better. The idea was later supported in reviews (e.g., Huffcut & Culbertson, 2011; Judge, Higgins, & Cable, 2000; Posthuma, Morgeson, & Campion, 2002). We know a lot more about assessment by interviewing, and how to make valid, interview-based decisions, than we have communicated to the world at large—where poor interviews remain the rule.

A series of meta-analyses have augmented the narrative reviews and provided explicit generalizations about the validity of (generally) aggregated interviews as predictors of job performance and other criteria. Mean validity coefficients reported in early studies were low but positive; in later analyses, mean coefficients were substantially higher as the literature grew and, perhaps, reported research with better interviews. A reasonable figure is a corrected coefficient (for criterion unreliability and range restriction) of about .36 or .37 (Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994); Conway, Jako, and Goodman (1996) used upper limits of about .56 for moderately structured interviews and as high as .67 for those that are highly structured—and .34 for poorly structured ones. Interview validity may not be as bad as once believed.

Meta-analytic conclusions evaluate interview validity more favorably than did the narrative reviews. That may be an artifact of the demands of meta-analytic research; a correlation coefficient serving as a data point implies some degree of

structure. If validity coefficients for the casual conversations called interviews could be computed, they would probably be lower on average than those with correlation coefficients computed but still called unstructured (cf., Schmidt & Zimmerman, 2004). Interviews, if well structured, can be quite valid predictors, but too often are neither structured nor valid.

Varieties of Structured Interviews

It is not easy to define what is meant by structured. Structured versus unstructured is a rhetorical, not a realistic, dichotomy; there are big differences in the degree and the rigidity of structure. In fact, the descriptive term of choice has changed over the years. Wagner (1949) did not call for *structured* interviews; he called for *standardized* interviews. By the time meta-analyses were examining moderators of interview validity, Wagner's term had almost disappeared, although some authors used both terms interchangeably. They are not synonyms; structure does not necessarily mean standardization. Every time an interviewer decides before an interview what questions will be asked, what judgments will be made, and how they will be recorded, some degree of structure exists. If such structure is developed uniquely for every interview, it is certainly not standardized, it is structured only to fit an individual candidate. It is preparation for the interview, usually done after examining a candidate's credentials—application form, résumé, letters of recommendation, and so forth—and noting some concerns worth exploring.

However, the term "structured" more typically refers to interviews tailored to fit a job, not an individual candidate. Structuring in this sense begins with the job description, pay classification, promotion patterns, and related data. From such information, traits relevant to performance may be inferred and appropriate questions (to be asked of all candidates) identified. This form of structuring implies at least some standardization.

Different people have different ideas of how interviews should be structured. Four general procedures are described here. The first uses minimal structure, guiding rather than dictating an interviewer's progress through an interview. The second is more tightly structured yet relatively flexible, permitting different candidates to be asked different questions. The other two are more firmly structured, allowing little deviation.

Patterned Interviews. McMurry (1947) developed patterned interviews, a precursor to many lightly structured procedures. It required stating clear, acceptable bases for selection such as desired traits, background experiences, or training. An interviewer's guide provided kinds of questions that might be asked for each of these, and training was supposed to ensure understanding of its questions and the selection standards. Appropriate rating scales were provided for recording summary evaluations.

Behavior Description Interviewing. A more complex modification was called the *Patterned Behavior Description Interview* (see Janz, Hellervik, & Gilmore, 1986). Janz et al. (1986) gave examples of the interview patterns of questions for 16 jobs. The method is based on the aphorism that the best predictor of future behavior is past behavior; all questions in a pattern ask about past behavior, making it an oral personal history inventory. Question development begins from critical incidents classified into dimensions of behavior. Questions (initial and follow-up) are written for each dimension unless that dimension can be assessed better by an alternative to an interview (e.g., tests, biodata, credentials). The correspondence of question to dimension need not be one-to-one; the same initial question can, with appropriate follow-up probes, provide information for more than one job dimension. For example, a critical incident for an employment test specialist might have been "Developed a valid hands-on performance test to measure problem-solving skills when informed under court order that written tests would not be permitted." The initial question might be, "Tell me about a time when you solved a measurement problem that precluded conventional testing procedures." Follow-up questions might include "What was unusual about your solution?" and "How did you get your solution accepted by others?" If the job dimensions included creative problem solving and persuasiveness, this question and its probes can tap both. After the interview, the candidate is rated on each job dimension on a simple 5-point graphic rating scale. The sum of the dimension ratings provides a total score.

Situational Interviews. Situational interviews are based on goal-setting theory that states that behavior depends in large part on goals or intentions. Theoretically, if people are asked to say how they would respond to critical situations others have faced on a job, their answers reveal their behavioral intentions. Responses can be scored systematically using a scale anchored by behavioral responses.

DEVELOPING A SITUATIONAL INTERVIEW

Latham (1989) outlined the steps in developing a situational interview:

1. Conduct a job analysis using the critical incident technique.
2. Develop an appraisal instrument such as behavioral observation scales (Latham & Wexley, 1977, 1981) based on the job analysis.
3. Select one or more incidents that formed the basis for the development of performance criteria (e.g., cost consciousness) which constitutes the appraisal instrument.
4. Turn each critical incident into a "what would you do if . . . " question.
5. Develop a scoring guide to facilitate agreement among interviewers on what constitutes a *good* (5), *acceptable* (3), or an *unacceptable* (1) response to each question. If 2 and 4 anchors can also be developed, do so.

6. Review the questions for comprehensiveness in terms of covering the material identified in the job analysis and summarized on the appraisal instrument.
7. Conduct a pilot study to eliminate questions where applicant/interviewees give the same answers, or where interviewers cannot agree on the scoring.

Like behavior description patterns, situational interviews begin with critical incidents but use them differently. Situational interviews emphasize the future, not the past: "What would you do if . . .?" rather than "What did you do when . . .?" Situational interviews usually use panels of two or more interviewers. According to Latham, the typical panel has two managers from the job area and one HR staff member. One person reads the questions, but all record and evaluate the answers. An example of a question and scoring guide is shown in Figure 11.2.

Instructions: Read the question, then listen as the applicants provide an answer and record the answer in the space provided. Then assign a score using the provided rating scale:	
Question:	You are selling cars on commission. Due to bad weather the night before, all the cars are covered with dirt. The sales manager tells you to go out and clean the cars. All the other salespeople are standing by the door waiting for customers to arrive. What would you do?
Response:	
Scoring Key:	(5) Would do the best job possible of cleaning the cars as I know that customers want to see the cars and having clean cars sells more cars, so everyone benefits. I would look for possible interested customers as they approach the store. (3) Would complain to the supervisor and ask why the others do not have to do it, but would go out and do the job if required. You need to follow orders from your supervisor. However, would do a quick, basic job so I could get back into the showroom as soon as possible. (1) Would refuse to do it. I make money selling cars. Would tell the supervisor to get other staff or another newer salesperson to do it.

FIGURE 11.2 An example of a question and a scoring guide for a situational interview question.

Attitude researchers have learned that asking about *intentions* is more effective for predicting behavior than asking about attitudes (Ajzen, 1991). The *situational interview* is based on the notion that asking about what you *would* do (i.e., your intention) is better for predicting what you *will* do, than is asking about the best thing to do (i.e., your attitude).

Comprehensive Structured Interviews. The term *comprehensive structured interview* is borrowed from Harris (1989) to distinguish the specific procedures described by Campion, Pursell, and Brown (1988) from the generic term *structured interview*. Campion et al. (1988) described their procedure as "more highly structured" than most other approaches. The procedure begins with job analysis to identify KSAs from which interview questions can be developed. Acceptable questions might include those used in behavior description or situational interviews, job knowledge questions, simulations or walk-throughs, and "willingness" questions presenting aspects of realistic job previews. If job requirements differ in importance, the difference is supposed to be reflected by the relative number of questions related to the different ones. The form of the questions is simpler than the previous two methods, more like those in a printed test; all candidates are asked precisely the same questions, and no prompting or follow-up questions are permitted (although a question may be repeated if necessary). Moreover, scores of all candidates should be available before the decision is made; this is an explicitly norm-referenced procedure. If feasible, 3-member panels are used; the same panel and the same process is to be used for every candidate. The same panel member is to conduct all interviews and ask all questions; all panel members are to take extensive notes. Questions, answers, and candidates are not to be discussed between interviews, but, after all candidates have been interviewed, large discrepancies in ratings may be discussed and changes made if appropriate. Candidates may not ask questions during the interview, although the procedure calls for a later nonevaluative interview with a personnel representative in which questions are permitted.

Comparison of the Examples. These examples have been presented to show variety, not as prototypes to be matched. All have shown reasonable reliabilities and validity coefficients, statistically significant and competitive with other predictors. All have been defended as practical.

There are, of course, unanswered questions. How much structure is necessary? In comparing the four examples, one should keep in mind the diminishing returns of structure as identified by meta-analysis (Huffcutt & Arthur, 1994). In doing so, however, other questions surface. The most highly structured interview guides are essentially oral tests with constructed responses. Is test-like standardization an

essential feature of interview structure? The same questions could be asked and answered in written form, the responses scored by readers. Would oral and written versions be alike in reliability and validity? Would one form or the other be more susceptible to contaminating sources of variance? Would examinee reaction be the same? Are we ignoring things that should be assessed (e.g., interpersonal communication skills)?

Interview Validity

Interview validity usually is described only with criterion-related validity coefficients; they are apparently higher than formerly supposed. Pooling data across interviewers who differ in individual validity, who make different systematic errors, and whose judgments are not independent, may have seriously underestimated validity coefficients. Very little attention has been given to the psychometric validities of interviewers' ratings. What inferences, if any, can be drawn validly about interviewees from interviewers' judgments? General answers are unavailable, so no general principles can be offered for improving the meaningfulness of interviews as assessments. Although interviewer ratings are made in a context different from many other ratings, they are, after all, subject to the problems of other ratings. We will not understand clearly what interviewers can assess until the research enterprise starts to develop theoretical statements of constructs appropriate for interview assessment, train interviewers in their meanings and manifestations, structure interviews appropriately, collect data, and conduct the confirmatory and disconfirmatory research needed to determine whether interviewers' ratings on these constructs lead to valid inferences about them.

Interview guides, rating scales, and general structure of interviews are often content related, relying on job analysis in their development. Lawshe's *content validity ratio* (CVR; Lawshe, 1975) was computed for items in each of three structured interview guides developed by Carrier, Dalessio, and Brown (1990). One of the guides was for use with experienced applicants, the other two for inexperienced ones. For experienced candidates, the approach worked quite well; the highest CVR items combined to form the best criterion-related validity. Not so for the inexperienced ones. Is content sampling, then, a useful approach to structuring interviews only for experienced people? We can't say. The finding is interesting but needs replication.

Interview questions and ratings can be informed by the job analysis or derived from it as content samples. The former is like the choice of predictors in a predictive hypothesis and may lead to more appropriate questions for inexperienced applicants. The latter may distinguish truly experienced candidates from those who merely claim the experience. Inexperienced applicants need to be assessed for aptitudes for the work they have yet to learn; aptitude is surely assessed better by tests than by interviewers' ratings.

Interviewer Characteristics

Research has shown that there are individual differences in the way interviewers use information to reach overall judgments and in the criterion-related validity of those judgments, and the studies have shown that treating different interviewers as mere replications of each other (i.e., pooling data across interviewers) is unwise. In a unique study by Dougherty, Ebert, and Callender (1986), three interviewers audiotaped interviews used in initial screening for entry clerical and technical jobs. Each interviewer saw some applicants and rated them on eight job-related dimensions and on an overall rating scale. All three interviewers rated all applicants from the tapes. Those hired were subsequently rated by their supervisors on 10 dimensions, including overall performance. Validity coefficients are shown in Table 11.1. ("Live" judgments are those of the actual interviewer at the

TABLE 11.1 Validity Coefficients for "Live" Overall Judgments, Mean of Overall Judgments, and Individual Interviewer Judgments

Criterion Dimension	*Live*[a] *Judgments (n = 57)*	*Mean of*[b] *Judgments (n = 57)*	*Interviewer*		
			1 Judgment (n = 56)	*2 Judgments (n = 54)*	*3 Judgments (n = 56)*
Learning Tasks	.10	.17	.09	.07	.24*
Minimal Supervision	.05	.32**	.19	.09	.41**
Organizing	.09	.18	.13	–.05	.26*
Judgment	–.05	.24*	.23*	.07	.26*
Job Knowledge	–.09	.12	.07	–.11	.23*
Cooperation	–.04	.09	.13	–.01	.08
Productivity	.03	.19	.12	–.05	.32**
Accuracy	.18	.28*	.25*	.19	.27*
Involvement	.06	.28*	.27*	.04	.34**
Overall Performance					
Actual	.06	.21	.15	.02	.26*
Predicted[c]			.23*	.19	.26*

[a]Overall judgments made by interviewers in the actual, live interviews; all other columns are correlations based on judgments from the tape recordings.
[b]Mean of the judgments based on tapes by the three interviewers.
[c]Using judgments predicted from the interviewer's own policy equation.
*$p < .05$;
**$p < .01$

Note: Adapted from Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71,* 9–15. Copyright by the American Psychological Association. Reprinted with permission.

time of the interview; all other columns refer to judgments based on the tapes.) Again, aggregated interviewer overall judgments were not significantly correlated with supervisory ratings of overall job performance; neither were ratings from two of the interviewers. The third, however, significantly and substantially predicted all supervisory ratings but one. The study went beyond demonstrating individual differences in interviewer validity; it also showed that interviewers can be trained to use more effective policies.

Interviewer Experience and Habit. Most managers prefer people with lots of experience, but sometimes we learn things from experience that are bad habits or misinformation. Gehrlein, Dipboye, and Shahani (1993) demonstrated that experience is not necessarily helpful to interviewers. Admissions officers (experienced interviewers) interviewed college applicants; other applicants were interviewed by alumni, faculty, and others termed inexperienced. Validity coefficients of interviewer ratings against GPA were nonsignificant for all of the individual experienced interviewers; surprisingly, inexperienced interviewers did much better. The authors suggested that experience tends to breed confidence even if it is unwarranted. Perhaps the less experienced people compensated for less confidence by planning their interview strategies—in effect, by developing a personal structure for their interviews.

> Judgment research has generally shown that *experience* leads to greater confidence, but *not* to greater ability to predict. Studies using experts such as livestock judges, physicians, psychotherapists, parole officers, and court judges have found experienced judgments to be just as susceptible to error as novice judgments.

Some interviewers habitually talk too much. Daniels and Otis (1950) found that interviewers generally do most of the talking, sometimes two or three times as much as the interviewees. Moreover, it has been shown that interviewers talk more with applicants they accept (Anderson, 1960). That finding is hard to interpret. Do interviewers talk more to applicants who show signs of success early in the conversation? Or do they simply feel good about themselves when they talk more, thereby feeling kindly toward the listening applicant?

If the interviewer is seen as an instrument for assessing candidate characteristics through conversation, it seems logical that the interviewer's contributions to the conversation would be relatively brief, encouraging the candidate to speak freely. When the purpose of the interview is to persuade the candidate to accept an offer, perhaps the interviewer should, in fact, talk more. But in nearly all other purposes, for example, where public relations is to be enhanced, the interviewer is likely to make a better impression on the interviewee by listening than by talking.

Apparently, the amount and kind of talking done by interviewers depends in large part on prior impressions of the candidates. In a decision-making interview, an interviewer often gets prepared by checking out application materials. If this preparation produces a favorable impression, the interviewer is likely to talk more and listen less; there are other first impression effects that bring the validity of interviews into question.

Stereotypes, Prototypes, and Biases. The notion of an ideal applicant need not be stereotypic. Prototypes of ideal candidates can be developed by deliberation, perhaps from job descriptions or with the help of supervisors and senior employees.

> How do different interviewers develop and use prototypes of desired candidates? . . . I distinguish between a stereotype (which develops willy-nilly, is widely accepted, and seems implicitly to apply to all members of a group) and a prototype, by which I mean something like a car designer's prototype, a carefully and systematically developed ideal to be achieved; for selection, the prototype should be defined by a set of attributes that not only describe the desired candidates but distinguish them from those less desired. . . . I suspect that work on the idea of a prototype as a planned ideal will be more fruitful than work on more or less generally accepted stereotypes of what is.
>
> *(Guion, 1987, p. 202)*

Whereas "Similar-to-me" is a bias, "Similar-to-ideal candidate" seems a useful match to an ideal prototype; if the prototype is valid, matching it should imply valid assessment as well.

Interviewers' biases potentially include demographic variables like sex, race, ethnicity, or age. Research generally reports little or nonsignificant differences in interviewers' ratings of men and women, but differences have been observed for racial and ethnic groups (Huffcut & Roth, 1998). The pessimistic view of these group differences attributes them to interviewer biases against minority group members. The optimistic view is that the group differences in average interview ratings are much smaller than group differences on cognitive ability measures. Research is needed that controls for differences in factors other than race and ethnicity.

A more general "similar-to-me" bias could inflate tendencies toward bias. In one study, racial similarity effects were stronger in conventional than in structured interviews, although mixed-race panels of interviewers avoided the effect (Lin, Dobbins, & Farh, 1992); similarity effects were not found for age. Another study of panels of interviewers showed a similar racial effect, giving higher ratings to candidates of the same racial identity as the majority of the panel (Prewett-Livingston, Feild, Veres, & Lewis, 1996).

Similarity biases are natural, but they are an example of a fundamental flaw in human intuition called judgment by representativeness (Tversky & Kahneman, 1982). ***Judgment by representativeness*** is the tendency to assume that things that look like each other *are* like each other. For example, an effective executive who made it to the top without a college education may think that a candidate with only "street smarts" will be similarly effective. That is, if the candidate has the same history as me, he must be as competent as me. According to Gilovich (1991):

> People assume that "like goes with like": Things that go together should look as though they go together. We expect instances to look like the categories of which they are members; thus, we expect someone who is a librarian to resemble the prototypical librarian. We expect effects to look like their causes; thus, we are more likely to attribute a case of heartburn to spicy rather than bland food, and we are more inclined to see jagged handwriting as a sign of a tense rather than a relaxed personality.
>
> *(p. 18)*

The problem with judgment by representativeness is that it often leads to predictable errors. Consider the aforementioned executive. What this successful executive fails to consider is (a) How many *effective* executives have no college education? (b) How many *ineffective* executives had no college education? and (c) How many effective executives *have* college educations? Odds are that sets b and c are much larger than set a. The point is not that the executive should necessarily hire the candidate with a college education, but that the executive should not let similarity on this one attribute interfere with his judgment about the candidates' other strengths and weaknesses.

Interviewee Characteristics

Obviously, characteristics of the person interviewed should influence decisions; they include the characteristics sought. Two special cases, however, merit concern as potential sources of error.

Memory. Interviews generally consist of questions requiring the interviewee to respond with a remembered event, state, or behavior. Personal recall may not be accurate. People may have implicit theories of their own personalities that emphasize stability (e.g., This is how I think now, so I must have thought similarly then). Other people, or the same people for other questions, have implicit theories that lead them to exaggerate changes that have occurred. That is, people make the implicit assumption that behaviors match attitudes. If a person recalls behavior (e.g., leaving a job) associated with an attitude, and if the attitude has changed, the response may describe behavior more in line with the present attitude than with the earlier reality.

Impression Management. Candidates try to make good impressions, and some are better at it than others.

Impression management is the attempt to influence the impression made on others. There are surely individual differences in self-presentation skills, but there is little information about kinds of job performance these skills may predict or the kinds of assessments they may contaminate. Interview research needs to study the effect of impression management. Does behavior successfully creating the desired impressions with one interviewer work equally well with another? Can interviewers learn to detect the deceptions the term "impression management" implies? If so, can they successfully ignore it in making job-relevant assessments or decisions? In a widely cited article, Kinicki, Lockwood, Hom, and Griffeth (1990) found that two factors described interviewer ratings on six dimensions. One they labeled "interview impression," the other was called "relevant qualifications." The terms are adequately descriptive; only the relevant qualifications factor validly predicted independent job performance ratings.

A study by Ellis, West, Ryan, and DeShon (2002) found that type of interview question influenced the type of impression management engaged in by the interviewee. Specifically, interviewees used more ingratiation tactics when answering situational questions, but they used more self-promotion tactics when they answered experience-based questions. We might expect, therefore, that behavioral description interviews would elicit more self-promotion (i.e., bragging), and that situational interviews would be met with more ingratiation (i.e., kissing up). Impression management tactics also appears to be related to personality: One study found that extraverted interviewees engaged in more self-promotion, whereas agreeable interviewees engaged in more ingratiation (Kristof-Brown, Barrick, & Franke, 2002).

In General

A large body of research on interviewing has given too little practical information about how to structure an interview, how to conduct it, and how to use it as an assessment device. Research suggests that (a) interviews can be valid, (b) for validity they require structuring and standardization, (c) that structure, like many other things, can be carried too far, (d) that without carefully planned structure (and maybe even with it) interviewers talk too much, and (e) that the interviews made routinely in nearly every organization could be improved vastly if interviewers were aware of and used these conclusions. There is more to be learned and applied in this domain.

Discussion Topics

1. What kinds of biodata items might be developed to distinguish between high-GPA and low-GPA students? Can you take these biodata items and turn

them into behavioral interview questions? What might you do to reduce faking in the responses?

2. Have you ever taken a highly structured interview, one where the interviewer has very little interaction with you other than asking the questions and waiting for answers? What did you think of it? Are there any negative consequences that you see that might be associated with using highly structured interviews?
3. The use of big data may lead to the conclusion that personal data serves as a useful predictor of job performance or turnover. Do you think organizations should be able to use personal data, such as credit scores, number of previous jobs, where you were born, or size of your high school, in order to make personnel decisions? If so, what should the limits be on their use of publically available information?

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior & Human Decision Processes, 50,* 179–211.

Anderson, C. W. (1960). The relation between speaking times and decision in the employment interview. *Journal of Applied Psychology, 44,* 267–268.

Baier, D. E., & Dugan, R. D. (1957). Factors in sales success. *Journal of Applied Psychology, 41,* 37–40.

Bentz V. J. (1967). The Sears experience in the investigation, description, and prediction of executive behavior. In F. R. Wickert & D. E. McFarland (Eds.), *Measuring executive effectiveness* (pp. 147–205). New York, NY: Appleton-Century-Crofts.

Bernadin, H. J. (1987). Development and validation of a forced choice scale to measure job-related discomfort among customer service representatives. *Academy of Management Journal, 30,* 162–173.

Bing, M. N., Davison, H. K., & Smothers, J. (2014). Item-level frame-of-reference effects in personality testing: An investigation of incremental validity in an organizational setting. *International Journal of Selection and Assessment, 22*(2), 165–178.

Breaugh, J. A., & Dossett, D. L. (1989). Rethinking the use of personal history information: The value of theory-based biodata for predicting turnover. *Journal of Business & Psychology, 3,* 371–385.

Campion, M. A., Pursell, E. D., & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41,* 25–42.

Carrier, M. R., Dalessio, A. T., & Brown, S. H. (1990). Correspondence between estimates of content and criterion-related validity values. *Personnel Psychology, 43,* 85–100.

Chapman, D. S., & Zweig, D. I. (2005). Developing a nomological network for interview structure: Antecedents and consequences of the structured selection interview. *Personnel Psychology, 58,* 673–702.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1996). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80,* 565–579.

Daniels, H. W., & Otis, J. L. (1950). A method for analyzing employment interviews. *Personnel Psychology, 3,* 425–444.

Dean, M. A., Russell, C. J., & Muchinsky, P. M. (1999). Life experiences and performance prediction: Toward a theory of biodata. *Research in Human Resources Management, 17,* 245–281.

Dipboye, R. L. (1997). Structured selection interviews: Why do they work? Why are they underutilized? In N. Anderson & P. Herriot (Eds.), *International handbook of selection and assessment* (pp. 455–473). New York, NY: Wiley.

Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71,* 9–15.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16,* 1–23.

Ellis, A. P. J., West, B. J., Ryan, A. M., & DeShon, R. P. (2002). The use of impression management tactics in structured interviews: A function of question type? *Journal of Applied Psychology, 87,* 1200–1208.

Furnham, A., & Jackson, C. J. (2011). Practitioner reactions to work-related psychological tests. *Journal of Managerial Psychology, 26*(7), 549–565.

Gehrlein, T. M., Dipboye, R. L., & Shahani, C. (1993). Nontraditional validity calculations and differential interviewer experience: implications for selection interviewers. *Educational and Psychological Measurement, 52,* 457–469.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life.* New York, NY: The Free Press.

Goffin, R. D., & Christiansen, N. D. (2003). Correcting personality tests for faking: A review of popular personality tests and an initial survey of researchers. *International Journal of Selection and Assessment, 11,* 340–344.

Guion, R. M. (1965). *Personnel testing.* New York, NY: McGraw-Hill.

Guion, R. M. (1987). Changing views for personnel selection. *Personnel Psychology, 40,* 199–213.

Harris, M. M. (1989). Reconsidering the employment interview: A review of recent literature and suggestions for future research. *Personnel Psychology, 42,* 691–726.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57,* 639–683.

Highhouse, S. (2002). Assessing the candidate as a whole: A historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology, 55,* 363–396.

Hollingworth, H. L. (1923). *Judging human character.* New York, NY: Appleton.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and I-O psychology: Reflections, progress and prospects. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 272–290.

Hough, L., & Tippins, N. (1994, April). New designs for selection and placement systems: The Universal Test Battery. In N. Schmitt (Chair), *Cutting edge developments in selection.* Symposium at meeting of the Society for Industrial and Organizational Psychology, Nashville, TN.

Huffcutt, A. I., & Arthur, W., Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79,* 184–190.

Huffcutt, A. I., & Culbertson, S. S. (2011). Interviews. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology* (pp. 185–204). Washington, DC: American Psychological Association.

Huffcut, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83,* 179–189.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55,* 243–252.

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods, 1,* 131–163.

James, L. R., McIntyre, M. D., Glisson, C. A., Bowler, J. L., & Mitchell, T. R. (2004). The conditional reasoning measurement system for aggression: An overview. *Human Performance, 17,* 271–295.

Janz, T., Hellervik, L., & Gilmore, D. C. (1986). *Behavior description interviewing.* Boston, MA: Allyn & Bacon.

Judge, T. A., Higgins, C. A., & Cable, D. M. (2000). The employment interview: A review of recent research and recommendations for future research. *Human Resource Management Review, 10,* 383–406.

Judge, T. A., Klinger, R., Simon, L. S., & Yang, I. W. F. (2008). The contributions of personality to organizational behavior and psychology: Findings, criticisms, and future research directions. *Social and Personality Psychology Compass, 2,* 1982–2000.

Kinicki, A. J., Lockwood, C. A., Hom, P. W., & Griffeth, R. W. (1990). Interviewer predictions of applicant qualifications and interviewer validity: Aggregate and individual analyses. *Journal of Applied Psychology, 75,* 477–486.

König, C. J., Klehe, U. C., Berchtold, M., & Kleinmann, M. (2010). Reasons for being selective when choosing personnel selection procedures. *International Journal of Selection and Assessment*, *18*(1), 17–27.

Kristof-Brown, A., Barrick, M. R., & Franke, M. (2002). Applicant impression management: Dispositional influences and consequences for recruiter perceptions of fit and similarity. *Journal of Management, 28,* 27–46.

Latham, G. P. (1989). The reliability, validity, and practicality of the situational interview. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 169–182). Newbury Park, CA: Sage.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal. *Personnel Psychology, 30,* 255–268.

Latham, G. P., & Wexley, K. N. (1981). *Increasing productivity through performance appraisal.* Reading, MA: Addison-Wesley.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28,* 563–575.

Lievens, F., Highhouse, S., & De Corte, W. (2005). The importance of traits and abilities in supervisors' hirability decisions as a function of method of assessment. *Journal of Occupational and Organizational Psychology, 78,* 453–470.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*(2), 27–66.

Lin, T. R., Dobbins, G. H., & Farh, J. (1992). A field study of race and age similarity effects on interview ratings in conventional and situational interviews. *Journal of Applied Psychology, 77,* 367–371.

Lodato, M. A., Highhouse, S., & Brooks, M. E. (2011). Predicting professional preferences for intuition-based hiring. *Journal of Managerial Psychology, 26*(5), 352–365.

Mael, F. A. (1991). A conceptual rationale for the domain and attribute of biodata items. *Personnel Psychology, 44,* 763–792.

Mael F. A. (1993). Rainforest empiricism and quasi-rationality: Two approaches to objective biodata. *Personnel Psychology, 46,* 719–738.

Mattioli, D. (2012, August 23). *On Orbitz, Mac users steered to pricier hotels.* Retrieved from http://online.wsj.com/news/articles/

McCarthy, J. M., Van Iddekinge, C. H., Lievens, F., Kung, M. C., Sinar, E. F., & Campion, M. A. (2013). Do candidate reactions relate to job performance or affect

criterion-related validity? A multistudy investigation of relations among reactions, selection test scores, and job performance. *Journal of Applied Psychology, 98,* 701–719.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.

McMurry, R. N. (1947). Validating the patterned interview. *Personnel, 23,* 263–272.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60,* 683–729.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60,* 1029–1049.

Nolan, K. P., & Highhouse, S. (2014). Need for autonomy and resistance to standardized employee selection practices. *Human Performance, 27*(4), 328–346.

Nowicki, M. D., & Rosse, J. G. (2002). Managers' views of how to hire: Building bridges between science and practice. *Journal of Business and Psychology, 17,* 157–170.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60,* 995–1027.

Posthuma, R. A., Morgeson, F. P., & Campion, M. A. (2002). Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *Personnel Psychology, 55,* 1–81.

Prewett-Livingston, A. J., Feild, H. S., Veres, J. G., III, & Lewis, P. M. (1996). Effects of race on interview ratings in a situational panel interview. *Journal of Applied Psychology, 81,* 178–186.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35,* 1–62.

Ryan, A. M., & Sackett, P. R. (1987). Pre-employment honesty testing: Fakability, reactions of test takers, and company image. *Journal of Business and Psychology, 1,* 248–256.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schmidt, F. L. & Zimmerman R. D. (2004). A counterintuitive hypothesis about employment interview validity and some supporting evidence. *Journal of Applied Psychology, 89,* 553–561.

Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology, 29,* 79–101.

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65,* 445–494.

Shaw, J. (2014). Why "Big Data" is a big deal. *Harvard Magazine, 3,* 30–35.

Sitser, T., van der Linden, D., & Born, M. P. (2013). Predicting sales performance criteria with personality measures: The use of the general factor of personality, the big five and narrow traits. *Human Performance, 26*(2), 126–149.

Society of Human Resource Management. (2012, January 4). *SHRM Poll: Most Employers Don't Use Personality Tests*. Retrieved from http://www.shrm.org/hrdisciplines/staffingmanagement/

Sparks, C. P. (1990). Testing for management potential. In K. E. Clark & M. B. Clark (Eds.), *Measures of leadership* (pp. 103–111). West Orange, NJ: Library of America.

Taylor, P. J., Pajo, K., Cheung, G. W., & Stringfield, P. (2004). Dimensionality and validity of a structured telephone reference check procedure. *Personnel Psychology, 57,* 745–772.

Tversky, A., & Kahneman, D. (1982). Judgment of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). Cambridge, England: Cambridge University Press.

van der Zee, K. I., Bakker, A. B., & Bakker, P. (2002). Why are structured interviews so rarely used in personnel selection? *Journal of Applied Psychology, 87,* 176–184.

Viglione, D. J., & Hilsenroth, M. J. (2001). The Rorschach: Facts, fictions, and future. *Psychological Assessment, 13,* 452–471.

Villanova, P., Bernardin, H., Johnson, D. L., & Dahmus, S. A. (1994). The validity of a measure of job compatibility in the prediction of job performance and turnover of motion picture theater personnel. *Personnel Psychology, 47,* 73–90.

Vinchur, A. J., Schippmann, J. S., Switzer, F. S., III, & Roth, P. L. (1998). A meta-analytic review of predictors of job performance for salespeople. *Journal of Applied Psychology, 83,* 586–597.

Wagner, R. (1949). The employment interview: A critical summary. *Personnel Psychology, 2,* 17–46.

Watson v. Fort Worth Bank & Trust, 108 S. Ct. 2777 (1988).

12

ASSESSING VIA RATINGS

Rating Formats, Research on Rating, and Errors and Rater Bias

Ratings are ubiquitous. Ratings of job performance are common; they are also used in many other assessment methods. Raters may be peers, superiors, or subordinates; they may be outsiders used for special purposes or used because of their special expertise. One person or several, working independently or as a panel, may do the rating. Ratings may be criteria or predictors. More research has been done on ratings of job performance than on ratings for other purposes, but it is relevant to other purposes and settings. This chapter emphasizes performance ratings. The focus is on *ratings as assessment methods or as a criterion for evaluating the validity of assessment,* not on their use in performance management.

Performance rating predates scientific psychology. Robert Owen, an early 19th century English industrialist and Utopian, developed a "silent monitor," a tapered wooden object about 4 inches long painted and numbered on the four sides. Each day, the supervisor would turn one side forward for each employee to indicate conduct the day before. Conduct consisted of hard work, being on time, producing well, and so forth. The black side, numbered 4, was shown for "bad" conduct; "indifferent" was blue, numbered 3; "good" was yellow and numbered 2, and "excellent" was white with a 1. A rating could be appealed to Mr. Owen; after time for appeal elapsed, the rating was recorded in a "book of character" (Cole, 1953, p. 56).

Rating requires at least three things: (1) a *source* of information, preferably observation or records, (2) *organizing* and *remembering* that information in preparation for rating, and (3) quantitatively *evaluating* what was remembered according to some rule. Remembering observations is central. In rating a product, the time from observation to evaluation is a few minutes; for annual job performance ratings, it might be a full year.

Whatever the use, ratings are psychometric measurements, even if not very precise. Ratings are often held in low esteem as measurements. They are victim to

countless forms of error, both random and systematic. Kane (1987) claimed the field of personnel psychology was stagnant because it cannot adequately measure its major dependent variable, work performance. However, Campbell, McCloy, Oppler, and Sager (1992) said, "Although ratings generally have bad press, the overall picture is not as bleak as might be expected" (p. 55), and they claimed that ratings are more likely to be explained by actual ratee performance than by contaminants. Yet they agree that there are problems. Ratings need all the help they can get, and most of the attempts to help have come mainly in three forms: (1) to improve rating formats, (2) to train raters, and (3) to influence the evaluation process.

A number of classic articles appeared in the late 1970s – early 1980s, including Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72–107. This article changed the way many in the field thought about the rating process. Two excellent books summarizing the complexity of performance appraisal issues are Murphy, K., & Cleveland, J. (1991). *Performance appraisal: An organizational perspective*. Boston, MA: Allyn & Bacon, and Murphy, K., & Cleveland, J. (1995). *Understanding performance appraisal: Social, organizational and goal-oriented perspectives*. Newbury Park, CA: Sage.

Rating Methods

Ratings can be based on scales, comparisons, or checklists. They can be used for overall assessment or for assessment of more specific dimensions. Sometimes diagnostic ratings of relative strengths and weaknesses are made. Some predictive hypotheses specify that a predictor should be related more to some aspects of work rather than others. Some call for a global, overall rating. Methods and formats should fit needs.

Graphic Rating Scales

Graphic rating scales are the most common of all rating methods. They can be used for overall ratings, but they are used more often to rate different aspects or dimensions of overall performance. Variants of graphic rating scales are shown in Figure 12.1. The basic form is *a,* with *b* showing how ratings become numbers. Some users prefer to give more structure to the scale by using verbal phrases instead of numbers, as in *c.* Numbers or words anchor the scale points.

The number of scale divisions varies widely; it is usually an odd number with "average" occupying a central position in the scale. More discrimination may be

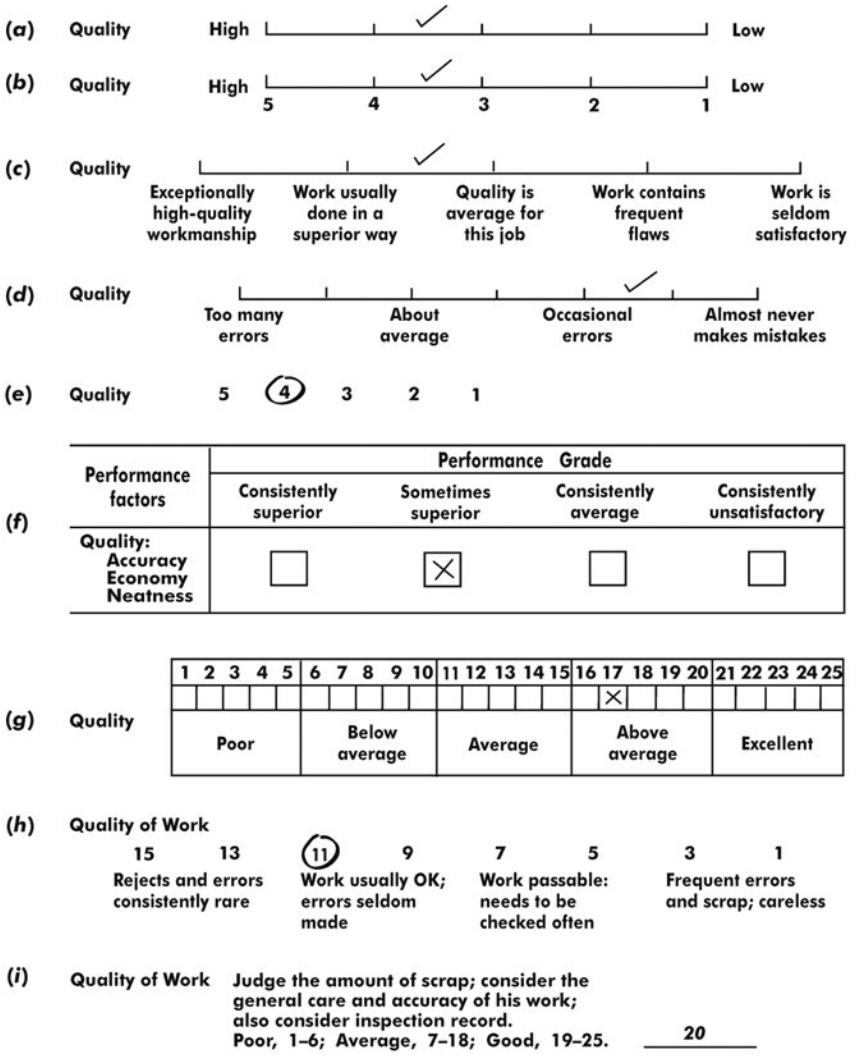

FIGURE 12.1 Some variations of a graphic rating scale; each line represents one way in which a judgment of the quality of a person's work may be recorded.

From Guion (1965).

needed at the "above average" levels, so scales like *d* can put average somewhat off center. Eliminating the basic line, as in *e,* eliminates problems in knowing where a rater means to put a sometimes hasty check mark, as does scale *f,* which includes verbal anchors and more definition of the performance trait being rated. The numerical and verbal anchors are combined in *g,* which also uses more and finer

gradations from the low to the high end of the scale. How many response categories is an optimal number? Little discrimination is possible with only two or three (although this may be enough when several ratings are added for an overall rating). It is probably absurd to ask raters to make distinctions along a 25-point range (although scale *g* simplifies the task by asking, in effect, for sequential judgments identifying first a group of five units). The 5-point scale is used so widely that it seems as if it had been ordained on tablets of stone. Some writers put the limit at nine scale points, but it is an arbitrary decision; there is little evidence that the number of scale units matters much, and the choice comes down to the researchers' preferences.

Scale *h* also combines verbal and numerical anchoring for eight possible responses. Numerical values of the responses can be changed according to the relative importance of the dimension being rated. In the example, "quality" has been prejudged to be worth a maximum of 15 points; other dimensions might have a maximum value of 8 or 10 or 30 or more points in a differential weighting scheme. If, for example, "cooperation" is deemed worth 25 points, the scale would have different numerical values, but still placed in eight response positions. Scale *i* entirely abandons the visual scale; it does not aid the rater by dividing the scale visually into five broader categories. It does, however, further structure the rating task by defining more clearly what is to be rated.

These variations show that the rater's task can be changed by changing (a) the nature and clarity of the anchors that define the values at points along the scale, (b) the nature of the required response, and (c) the clarity of the definition of the dimension to be rated. The developer of a graphic rating scale should try to avoid ambiguity; beyond that, the research literature gives little help in choosing one format over another.

Employee Comparisons

Another well-established practice compares the ratee to others, either on overall performance or on multiple dimensions. The usual result is a ranking of ratees, achieved in different ways by different methods.

Method of Rank Order. Ratees might be listed on a sheet of paper, and raters may be asked to put the number 1 by the name of the best of the lot, a 2 by the next best, and so on through the list. Names might be placed on cards to be arranged. A more systematic procedure, using cards presented in random order is "alternation ranking." When the dimension to be rated (e.g., conscientiousness, or overall job performance) is understood, the rater first identifies the best of the lot on that dimension and then the poorest. Cards with these names are pulled and the sorting has begun. Of the remaining names, the rater again selects the best and the poorest and places those cards accordingly. The process of alternating from best to worst continues until all have been ranked. The task becomes progressively harder;

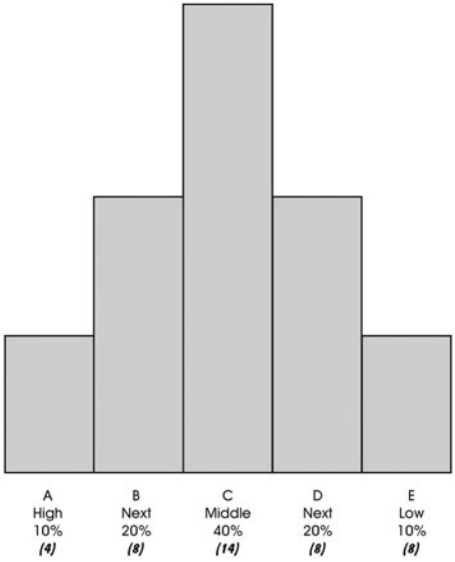

FIGURE 12.2 A sample form for forced distribution ratings; numbers in parentheses show how a rater with 42 people to rate should distribute them.

From Guion (1965).

extreme judgments are easy, but differences near the center of the distribution are harder to identify.

Method of Forced Distribution. When many people are to be rated and fine distinctions are not needed, gross ranking can be done with a *forced distribution*. This is a variant of graphic rating scales in that each person is assigned to a category in frequencies that mimic the assumption of a normal distribution. A number of categories is chosen (typically five, rarely more than nine), and proportions of distributions (translated into frequencies) to be placed in each category are specified. A 5-category example is shown in Figure 12.2. A rater with 38 names to rank writes the names of the four top people in column A, the names of the next best eight people in column B, and so on.

Method of Paired Comparisons. Each ratee can be compared with each of the others in a set. For each pair of names, the rater indicates the better one on some specified dimension; the top of the rank order is the one chosen most frequently. The same name should not appear in two consecutive pairs; each person should be listed first and second equally often. There might be a lot of pairs; if five people are to be compared, there are 20 pairs of names. Ten people require 45 pairs, 190

pairs for 20 people. These numbers assume that each pair is compared only once; the number of pairs is n(n—1)/2, where *n* is the number of people to be ranked. Every pair can be listed twice using both orders of presentation, but this requires twice as many pairs.

Lawshe and Balma (1966) provided tables for setting up such pairs. The number of times a given name is preferred can be transformed into a standard score scale, often with a mean of 50 and standard deviation of 10. With a long list of people to be compared, the amount of time required can get out of hand. Reasonable people disagree about how long is too long. Guilford (1954) put the limit at about 15 people, but Lawshe, Kephart, and McCormick (1949) reported that a list of 24 names (276 pairs) was rated reliably in 30 minutes—not an excessively wearying task, and one that could be shortened using computers.

Behavioral Descriptions

It seems reasonable to assume that raters can offer better assessments if they avoid glittering generalities or ambiguities and describe specific on-the-job behavior or outcomes.

Behaviorally Anchored Rating Scales (BARS). Smith and Kendall (1963) described a logic of rating and a procedure for developing a rating system. Many rating scales have been said to follow the Smith and Kendall approach, but they do only if using a full system of supervisory observation, recording, and rating of behavior. It was the form, and its use of scaled behavioral anchors, which attracted attention and resulted in the generic term *behaviorally anchored rating scales* or BARS. The many rating methods called BARS, and some criticisms of BARS not relevant to the procedures recommended by Smith and Kendall, called forth a clarification by Bernardin and Smith (1981). They pointed out that the Smith–Kendall approach was a sequence beginning with observation followed in order by inference, scaling, recording, and summary rating. Some say that the method has evolved, and that evolution accounts for the variety. It can be said more accurately that it has been distorted by treating it merely as another rating format, without treating the form as part of a system. We provide only a rudimentary summary here.

Table 12.1 outlines the general steps in developing a BARS. First, the behavioral anchors were not intended to describe behavior a rater had actually observed; they were descriptions of behavioral *expectations* at different levels of performance on specified dimensions—examples that might be anticipated or "expected" of a ratee at any of these levels, even if they did not actually occur. They were "expectations" in the sense of "That's just the sort of thing you come to expect from Joe." Expectations, in the Smith–Kendall sense, are anticipations of reality, not idealistic dreams of job demands or obligations.

TABLE 12.1 General Steps in Developing a BARS

Step #	*Description*
1	Convene one or more groups of potential raters.
2	Develop a list of the performance dimensions that should be evaluated.
3	Develop definitions of high, low, and acceptable performance for each dimension.
4	Develop lists of behavioral examples of high, low, and acceptable performance.
5	Give the lists of behavioral expectations and the dimension definitions to one or more new groups of potential raters not included in the first groups.
6	Designate judges among potential raters to identify the behavioral examples, within each dimension, that describe a worker whose performance is outstanding and another whose performance is unsatisfactory.
7	Give statements that survived the preceding steps to judges from another group of potential raters for scaling by the method of equal appearing intervals.
8	Develop and distribute a final rating form to raters before ratings are due.

The second key provision is *retranslation*—a procedure to ensure that behavioral statements originally written for a certain dimension are seen by others as illustrations of that dimension. The procedure is analogous to that in translating a passage from one language into another. A first group of judges writes behavioral expectations to fit each dimension. A second, independent, group of judges (a) reads statements for all dimensions, mixed together in random order, (b) discusses definitions of the dimensions for a common understanding, and then (c) independently allocates each statement to a dimension. A "good" item is allocated by most judges to the dimension for which it was developed. If there is no modal agreement about where it belongs, the statement is dropped.

The third key provision minimizes ambiguity of scale value by having judges sort statements on a range from extremely unfavorable to extremely favorable. The variance of judgments is a measure of the ambiguity of the statement; high variance statements are eliminated.

A fourth feature of the Smith–Kendall procedure is usually ignored. It permits raters to give at least one example of ratee behavior actually observed for each dimension rated. It could be inserted at that place on the scale that appropriately identifies its position relative to the defining anchors.

The term BARS has come to mean a kind of rating scale format that uses only some of the Smith–Kendall procedures. Figure 12.3 illustrates a BARS format; it used the Smith–Kendall procedures for scale definitions and for generating, retranslating, and scaling behavioral expectations. It does not illustrate a procedure for getting continuous observations and recording them as part of the rating process. Note, however, a difference in the form from traditional graphic scales: The scale separation marks are not the scale points that are anchored. True, there are very general descriptions apparently anchoring the top, bottom, and midpoint of the scale (shown on the left). Instead, the scale values anchored are those of the

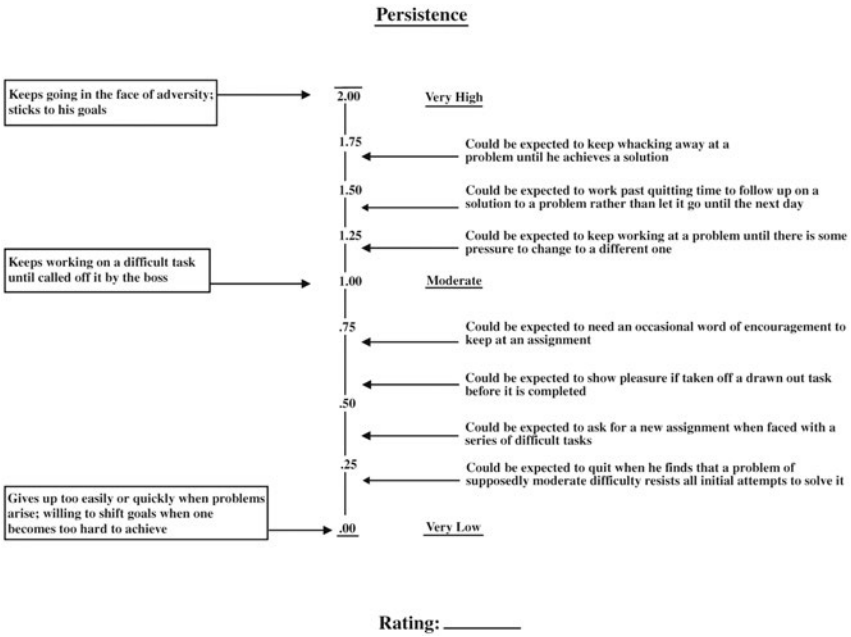

FIGURE 12.3 One example of a behaviorally anchored rating form devised using some key features of the retranslation of expectations; it is a form for measuring an aspect of motivation, work persistence, in a group of engineers.

behavioral examples, shown by the arrow pointing from the statement to the scale. A rater can decide which statements exemplify the kinds of behavior one might expect from the ratee.

Behavioral Observation Scales (BOS). Instead of largely unobserved behavioral expectations, Latham and Wexley (1981) rated behaviors actually required on the job, grouping them for specific job dimensions. Their scales are called *Behavioral Observation Scales (BOS)*. The response scale is *frequency* of observation, a 5-point scale ranging from 1 *(almost never)* to 5 *(almost always)* as shown in Figure 12.4. The five points are defined in terms of the percent of the time the behavior is observed. Latham and Wexley (1981) suggested percentages of 0%–64% for *(almost never)* through 65%–74%, 75%–84%, 85%–94%, and 95%–100% *(almost always)*; they have also reported using a straightforward 20% increment for each scale point.

A BOS can be developed in less time than BARS because prior item scaling is not needed. If job analysis is well done and well organized, behavioral statements should be prepared with minimal effort and time. Job analysis surveys may be too elemental; if so, job experts may consolidate elementary items into broader, more comprehensive statements. Items are usually considered equally weighted, but differential weights could be assigned by expert judgment. The job relevance of the ratings is obvious.

Instructions: Observe the clerk's performance during five, eight-hour shifts and then indicate the percent of time that the behavior is observed out of the number of times the behavior could be exhibited.					
Frequency of Behavior	0–64% Almost Never	65–74%	75–84%	85–94%	95–100% Almost Always
Points	1	2	3	4	5
Greets customer					
Has a friendly face and smiles					
Fills customer order correctly					
Answers customer questions correctly					
Tries to sell additional product					
Asks customer if there is anything else they need					
Thanks customer and asks them to come again					
Calculate total score by adding points. Total Points =					

FIGURE 12.4 A Behavioral Observations Scale (BOS) for assessing a retail clerk's customer service skill.

An important feature of the *BOS* is that the *evaluation takes place in scale development,* not during the rating process itself. In other words, the developers of the rating instrument determine what behaviors should be engaged in frequently. The rater, therefore, is merely an *observer* and *reporter,* rather than an evaluator.

Although the BOS is aimed at minimizing rater involvement in the evaluation process, by simply asking him or her to report frequency with which the behaviors occur, we suspect that rater evaluation is injected commonly into the frequency reports. Certainly, a supervisor can see which behaviors lead to a positive evaluation and which lead to a negative. Moreover, prototypes of effective employees are likely to influence frequency reports as much or more than the actual behavior of the ratee. An advantage of the method, however, is that it communicates clearly to the ratee which behaviors should be engaged in frequently, and which behaviors should be avoided.

Points	Level of Performance	Summary of Behavioral Anchors
5	Easily Exceeds Expectations and Performs at Expert Level	Gave memorable presentation. Engaged audience. Projected a sense of control. Spoke very clearly. Was highly organized and used time appropriately. Gave detailed answers to all questions; answers were responsive to questions and very clear; answers to questions exhibited a great deal of expertise.
4	Exceeds Expectations	(Use if performance falls between 3 and 5 on scale).
3	Meets Expectations	Spoke clearly. Maintained eye contact with audience. Maintained the attention of audience. Had some level of organization but lacked clear point, introduction, or conclusion. Tried to answer questions, but lacked detail in some answers or lacked expertise to answer questions. Used most of time allowed, but could have been more efficient in use of time. Became stressed at times but quickly gained composure.
2	Below Expectations – Minimally Acceptable Performance	(Use if performance falls between 1 and 3 on scale)
1	Well Below Expectations	Swore or used inappropriate language. Was difficult to hear or understand. Ran short of time or ran well over time limit. Did not maintain eye contact. Could not or would not answer questions. Became anxious or stressed during presentation.

FIGURE 12.5 An example of a Behavior Summary Scale (BSS) developed to rate performance in an oral presentation exercise.

Note: Could also be used as a method of evaluating teaching performance for a professor in a class.

Behavior Summary Scales (BSS). Assume a rating scale like row *h* in Figure 12.1, where each of a few descriptors covers a range of scale values. Now replace those relatively vague descriptors with a set of behavioral statements, each of which consolidates or summarizes a larger number of highly specific behavioral examples. The result would be what is referred to as a *Behavior Summary Scale (BSS)*. The BSS-type scale in Figure 12.5 could be used for rating performance in an oral interview.

Although a BSS could be seen as similar to a BARS, with simply multiple behaviors per anchor point, Borman (1986) described the differences between a BARS and a BSS especially as it related to the developmental process. As discussed by Borman (1986), in developing a BSS for use with Navy recruiters, experts

generated hundreds of examples of specific behavior in a two-day workshop; others were derived from stories told by recruits about experiences with recruiters. Content analysis resulted in nine performance categories. Eliminating redundancies and retranslating the remainder resulted in a pool of 352 examples. In BARS development, a few of these with the least variance in scale values, with scale values scattered nicely throughout the range, would be used. Instead, the group described by Borman tried to write more general summaries so that the summaries would represent, as much as possible, the content of all 352 examples.

From a user perspective, many practitioners today refer to any rating scale that uses behaviors anchors to define scale points as a BARS, including those that use several behavioral examples or behavioral summaries. Although from a historical perspective, it is important to differentiate between approaches such as the BARS and the BSS, especially with regard to the nature of the rating and developmental processes, as noted previously, the term "BARS" is now used in a generic sense to refer to most types of rating scales that employ behavioral descriptors as anchors.

Two other observations could be made with regard to the BSS-type scale we have provided in Figure 12.5. First, given the emphasis is on the use of ratings in assessment for selection, the example offers a scale that was developed for use in rating performance in an oral presentation exercise. This type of exercise often is used in an assessment center or in panel interviews. However, the scale could also be used to assess the job performance of a teacher or professor in terms of his or her ability to serve as an interesting and engaging lecturer. Second, the resulting rating scale is also very similar to the type of device that might be used in competency studies.

Forced-Choice Scales. Finding at the outbreak of World War II that performance ratings used by the United States Army did not help distinguish officers ready for promotion from others, Sisson (1948) developed a new, more differentiating system known as *forced-choice ratings*. The method used tetrads of four descriptive statements, each with two statements about equally favorable and two equally unfavorable. Prior research determined, for every statement, a preference index (P) for favorability and a discrimination index (D) of how well the statement distinguished between those independently identified as superior and others. Let + indicate high preference or discrimination and – indicate low; every tetrad had statements described as P+D+, P+D–, P–D+, and P–D–. The rater chose one statement as most descriptive and another as least descriptive, without knowing the scoring key (which was limited to discriminating items). The method gave valid ratings, but raters resisted use of a system they could not control.

Multisource Ratings—360s. Performance ratings can be obtained from a variety of sources. However, one specific type of rating method, multisource or 360s, relies upon ratings from multiple sources. Despite a number of psychometric and practical issues, 360s have become very popular as a rating method. The sources may include supervisors, self, peers, subordinates, and customers. Unfortunately, the

correlations between different sources is often quite low, leading to the recommendation that 360s be used for development rather than for administrative purposes. This would appear to limit the usefulness of 360 ratings as a predictor. However, it would not affect the use of such ratings as a criterion variable and it is possible that differences in the relationship between various predictor—source combinations may be an important source of information in building predictive hypotheses (Putka, Hoffman, & Carter, 2014).

Psychometric Research on Ratings

Regardless of purpose or quality, ratings are measures. Questions and issues in the psychometric evaluation of tests and other assessments apply also to ratings, with added ones as well.

Measurement implies individual differences in the trait measured, and they imply variance and the evaluation of possible sources of variance in the resulting measures. Variance in ratings (or "scores") should, of course, be associated mainly with variance in the actual performance of ratees. Variance in the ratings also stems from influences of the measurement procedure, irrelevant worker characteristics, characteristics of the situation in which performance is measured, and characteristics of the raters. In short, common psychometric problems are exacerbated in ratings.

Constructs Assessed

Constructs rated are rarely well defined, so psychometric validation of ratings is difficult. Factor analyses has been used with sets of ratings to identify underlying dimensions, and other forms of correlational analysis have been used to see whether different ratings of presumably the same constructs correlate well with each other but not with ratings presumed to assess dissimilar constructs. One serious problem with this approach is that so-called convergent validity, or of interrater reliability, may be little more than evidence of converging biases.

Agreement, Reliability, and Generalizability

Interrater agreement is often treated as a form of reliability, but agreement and reliability are different. Judges agree if they make the same ratings; they are reliable if they put ratees in roughly the same relative order. The distinction is clear in Table 12.2. Reliability can be high without agreement about the degree to which the characteristic being judged describes the ratees (Case 2). It can be low without necessarily meaning much disagreement among raters (Case 3). Both agreement and reliability are useful information about a set of subjective ratings.

Which statistic do you want? The answer depends on the intended use. If the ratings are used as validation criteria, interrater reliability (or "rate–rerate" reliability of a single rater) is more important because reliability limits validity. If the

TABLE 12.2 Hypothetical Ratings Illustrating Different Levels of Interrater Agreement and Interrater Reliability for Interval-Scaled Data

Ratee	*Case 1: High Interrater Agreement and High Interrater Reliability*			*Case 2: Low Interrater Agreement and High Interrater Reliability*			*Case 3: High Interrater Agreement and Low Interrater Reliability*		
	Rater			*Rater*			*Rater*		
	1	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
A	1	1	1	1	3	5	5	4	4
B	2	2	2	1	3	5	5	4	3
C	3	3	3	2	4	6	5	4	5
D	3	3	3	2	4	6	4	4	5
E	4	4	4	3	5	7	5	4	3
F	5	5	5	3	5	7	5	5	4
G	6	6	6	4	6	8	4	4	5
H	7	7	7	4	6	8	5	5	4
I	8	8	8	5	7	9	4	5	3
J	9	9	9	5	7	9	5	5	5
M	*4.8*	*4.8*	*4.8*	*3.0*	*5.0*	*7.0*	*4.7*	*4.4*	*4.1*
SD	*2.7*	*2.7*	*2.7*	*1.5*	*1.5*	*1.5*	*.5*	*.5*	*.9*

Note: From Tinsley, H.E.A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjunctive judgments. *Journal of Counseling Psychology, 22,* 358–376. Copyright by American Psychological Association. Reprinted with permission.

ratings are to be used for decisions based on level of proficiency, or if they are to aid interpretations of correlated test score levels, agreement is more important.

Estimating Rater Agreement

With two raters, each rating all ratees, and only a few rating categories, an easy index of agreement is the percentage of ratees assigned to the same categories by both raters. An early index was *kappa* (Cohen, 1960), appropriate for the case with two raters and nominal ratings:

$$kappa = (p_a - p_c)/(1 - p_c) \quad (1)$$

where p_a = actual proportion of agreements, and p_c = expected or chance proportion of agreements. For perfect agreement, *kappa* = 1.00. A more sophisticated, and much more complicated, approach to estimating rater agreement is given by generalizability theory, which was discussed in Chapter 5.

As part of the Job Performance Measurement Project in the military services of the United States, Kraiger (1990) studied experimental performance ratings in each of four Air Force specialties. The major source of variance in all four specialties was not ratees but the interaction of ratees with rating sources. Increasing the number of raters, if they are reasonably independent and parallel, will increase reliability; Kraiger (1990) concluded, on the basis of his full data, that the generalizability of these ratings (unlike their classical reliability estimates) can best be improved by averaging ratings from more sources. Raters with different perspectives will see different pieces of information about any given ratee; averaging across these perspectives can give opportunity for more relevant information to influence composite ratings.

Validity of Ratings as Predictors

Performance ratings are often criteria but can be predictors. Ratings are incorporated in interviews, assessment centers, work samples, portfolios of past achievements, auditions, or free-response tests. Note that the predictor in these assessment methods is not the interview, assessment center, or whatever, but rather it is the rating summarizing someone's judgment based on observations. Evaluation of job relatedness of ratings is done just as it is for other predictors.

Psychometrically, validity describes inferred meaning. Ratings are too often accepted uncritically as meaning whatever the rating scale label says, even when raters, if challenged to define the label, would not agree. Competent evidence of psychometric validity is rarely sought; in fact, most often, no psychometric evaluation occurs at all beyond possible checks on interrater agreement or reliability.

Bias as Invalidity

Ratee characteristics not being rated are sources of bias if they influence ratings; they reduce validity. In a widely cited meta-analysis, Kraiger and Ford (1985) found that raters gave higher ratings to ratees of their own race. Later, however, Sackett and DuBois (1991) compared the Kraiger and Ford (1985) findings with those in gigantic USES and Project A data sets. In the USES data, and in the Army technical proficiency and personal discipline ratings, both White and Black raters gave higher mean ratings to Whites. For military bearing, Blacks received higher ratings from both Black and White raters. Why did these big studies differ from that by Kraiger and Ford? Sackett and DuBois (1991) wondered and looked at the studies by Black raters in the Kraiger-Ford meta-analysis. Two were lab studies, four used peer ratings; of the eight supervisory ratings studies, four were done before 1970. They concluded that the finding of higher ratings within same race rater–ratee pairs were premature. The Sackett and DuBois (1991) comparison of pre- and post-1970 studies suggests that research findings, particularly when major social issues are involved, are specific to the times, to the zeitgeist, in which they

are obtained—as Cronbach (1975) warned. Perhaps this presumed interaction is another example where social change—greater acceptance of diversity—has resulted in a change of what is scientifically demonstrable.

The Rater in the Rating Process

The Classical Psychometric Errors

Central Tendency. Some raters cluster all ratings around a central point on the scale, a midpoint or a subjective average, resulting in low variance. Central tendency seems to indicate raters who avoid unpleasant consequences by avoiding extreme ratings.

Leniency or Severity. Some raters are easy, some hard; some lenient, others severe. Early discussions of the leniency error described it as giving higher ratings to people the rater knows; the more general idea of habitual leniency or severity in rating long has been included in the definition and is now dominant. Raters with very high mean ratings are considered systematically lenient; those with low means, systematically severe.

Halo. E. L. Thorndike defined *halo error* as a "marked tendency to think of the person in general as rather good or rather inferior and to color the judgments of the [specific performance dimensions] by this general feeling" (Thorndike, 1920, p. 25, as quoted by Balzer & Sulsky, 1992, p. 975). According to Balzer and Sulsky (1992), his work used contradictory definitions of halo: (a) correlations of ratings on specific scales with overall ratings, and (b) intercorrelations among dimension scales. The one operational definition assumes that a general impression influences ratings on dimensions; the other assumes that raters simply fail to distinguish dimensions. Both assumptions of halo lead to spurious intercorrelations.

Dimensions to be rated are not ordinarily orthogonal, so some observed correlations are not errors. The decades of research have "provided documentation that the phenomenon is ubiquitous. More recently, a great deal of effort has been expended on reducing halo, a modest amount on articulating the sources of halo, and surprisingly little on whether haloed ratings are inaccurate" (Cooper, 1981, p. 219). Intercorrelations may be influenced by reality or by the rater's implicit theory of personality or performance; they may also be due to error.

> A simple example of *halo error* is the widespread perception that tall people are better managers than short people. Research has shown that height is more strongly related to *subjective* ratings of performance than to *objective* performance measures (Judge & Cable, 2004).

Other Psychometric Errors

Prior information about a ratee may have a biasing effect, although the effect seems to diminish over time. *Prior impressions* may be based on knowledge of prior ratings. An experiment by Murphy, Balzer, Lockhart, and Eisenman (1985) found that knowledge of a ratee's previous performance rating influenced ratings of subsequent performance. This well-replicated finding has important implications in assessment centers where ratings are made on several dimensions, then discussed in sequence by the panel of raters. A similar concept (called *escalation bias*) was studied by Schoorman (1988). Raters had prior information about ratees; some raters had participated in decisions to hire ratees; some had agreed, some did not. The bias effect of participation and agreement with the hiring decision accounted for fully 6% of the rating variance.

Individual Differences in Ability to Rate

Rater Qualifications. Some raters are more qualified to rate than others. The main qualification is relevant knowledge, including knowledge of demands on the ratee as well as understanding ratee behavior. It may include knowledge of the work process and of both desirable and flawed product characteristics. Qualifying knowledge comes from observation or experience, not from hearsay, prejudice, or stereotypes. Typically, although not always, immediate supervisors are more qualified to rate job performance than second-level supervisors who are more removed from the person and the work being rated; the relevance of the contact, not merely its frequency, seems to be the key qualification. For some work samples, the most qualified raters may be people who have demonstrated a high level of skill at the work, although highly skilled people may have automatized their skills too thoroughly to observe clearly.

Training. Minimal rater training should include instruction in the meaning of words used on the rating form, the procedure to be followed in making the ratings, and aspects of the judgment process such as avoidance of rating errors. Much can be added.

Borman (1979) suggested that rater training might produce, and be evaluated by, three kinds of outcomes: (1) reduction in classical rating errors, (2) improved psychometric validity, including interrater agreement, and (3) improved accuracy. Of these, we think the most practical efforts are those to increase psychometric validity, but most research emphasis has been placed on the other two. According to Bernardin and Buckley (1981), efforts to replace classical errors have amounted to little more than trading in one kind of response set for another. They advocated training that emphasizes observation of behavior, such as the following:

1. Diary keeping, in a formal system, with support at all higher organizational levels ensuring that supervisors are themselves evaluated on how well they keep diaries.

2. Frame of reference (FOR) training. FOR training involves identifying raters whose ratings are peculiar and helping them develop a common understanding of the dimensions to be rated and of the observations that support different levels of ratings.
3. Training raters how to be critical. Many raters hate to give negative ratings. Training might increase ability to handle encounters resulting from negative appraisals (Waung & Highhouse, 1997). Increased self-efficacy in giving feedback is likely to reduce rating errors such as leniency, yet this topic has not received much research attention.

Different people may observe a worker's performance from different perspectives, or frames of reference. Usually there is a dominant, modal frame of reference in an organization, maybe not deliberately. With a common frame of reference, raters can define levels of performance effectiveness for different performance dimensions with a common language. To see if there is one, raters can be asked to rate the relative effectiveness of each item in a list of critical behaviors and the importance of job dimensions. Raters who do not agree with most other raters are considered idiosyncratic and targeted for FOR training. They are brought together to consider the job description, to discuss the important performance dimensions, and to understand the differences between "correct" (modal) evaluations and various idiosyncratic ones. Such training uses a conference method of group problem-solving techniques to arrive at a consensus about how rating should be done. Day and Sulsky (1995) considered FOR training the most promising of all rater training methods.

Organizational Level. Self, supervisory, and peer performance ratings typically do not correlate well. People at different organizational levels may have different qualifications to rate. Oppler, Peterson, and McCloy (1994) found that peer and supervisory ratings were predicted by different things and were not interchangeable. They attributed the differences to the greater exposure of peers to fellow trainees, especially in Army settings. These results might also be explained by differences in the constructs most salient at the different levels. Research on supervisory ratings may not apply to other rating problems; self or peer ratings may work better (i.e., be more valid) for some purposes. In assessment centers, assessors are not necessarily supervisors, but they do occupy a hierarchical position of authority; peers may be in a better position to rate some kinds of assessment center performance. Peers may be better judges of certain traits (e.g., work motivation). For some purposes, self-ratings may be more valuable, such as self-ratings of confidence. For other purposes, other raters may be better: Customers can rate service; experts can rate work sample results; or professional people can rate readiness for something (such as readiness to return to work after trauma or to profit from specific training).

Rater Motivation. Poor, invalid ratings may be expected from a rater who lacks confidence in the purpose of the ratings, distrusts the researcher, or simply "has

other fish to fry." Understanding and acceptance of purpose is crucial; a supervisor who sees the request for ratings as "still more paperwork" is likely to look on the request more as an infringement on his or her time than as a positive means of achieving personal or organizational goals.

Rater motivation might differ for different rating purposes. A rater might be more highly motivated to rate people where a "deservedness" decision is to be made, where the ratings may determine who gets merit pay or special recognition, or where "designation" decisions are the outcome, such as picking out one ratee among others for promotion or a special training opportunity. We have only begun to scratch the surface in understanding the effects of the social context on performance ratings (Levy & Williams, 2004).

Aids to Observation and Memory

Records. In many settings, daily production records are kept. Review of such records can jar the rater's memory and point out aspects of performance such as level and consistency of production, recorded errors, and related facts. If the task is to assess performance quality, and if such factual information is available, why rate? A part of the answer is that information in the files may be uneven in quality and relevance. A simple thing like the number of widgets produced each day may be tempered by a rater's knowledge of the specific equipment a ratee uses; some pieces of equipment are more prone to breakdown, slower in function, and so on. The best assessment may still be a subjective judgment—but it must be an informed judgment reached by getting and considering an array of factual information.

Incident Files or Diaries. Some appraisal forms list job duties on one side of the page and require the rater to write an anecdote or critical incident illustrating a ratee's performance of each of them. The principle is similar to that in the Smith–Kendall BARS approach of assigning ratee behavior examples to appropriate points in the scale: The rating given is supported with specific behavioral evidence. A problem with this is that the evidence recalled at the time of rating may not be a good summary description of the ratee or ratee behavior. The rater is more likely to remember the dramatic, salient example of a single brilliant achievement or major blunder than more typical incidents (of these the blunder is more likely to be remembered). Recent events are more likely to be recalled than those that happened earlier.

Bernardin and Buckley (1981) recommended diary keeping as a training method, but only if it is systematic and has support from the top of the organization. Top support for diaries implies that supervisors themselves are evaluated on how well they keep diaries. Diaries, however, offer no panacea. In an experiment, Balzer (1986) found that a diary system can slip badly for those who have good impressions of the ratee but do not see the rating task as very important; it will work best for those who have good impressions *and* see the task as central to their jobs. This field-testable hypothesis deserves testing.

Bernardin and Beatty (1984) offered recommendations for training people to maintain such records: (a) Tie training in recording observations to scale familiarization training so that observations are recorded relative to the behavioral dimensions to be rated, (b) record objectively, not evaluatively, (c) record a predesignated minimum number of observations per scale, (d) make the diary-keeping system a formal part of organizational policy and practice, and (e) require the rater's supervisor to monitor the diary keeping.

Comments

The best procedures for one rating purpose may not fit a different one. Rating people on behavior shown only during the course of an audition or interview is different from rating performance over the span of a year; rating aspects of objects, such as work samples or portfolios, is different from rating people or aspects of their behavior. They may differ in time span of observations or of memory, in complexity of dimensions rated, in organization of data, in opportunity to reconsider, and in many other details. Effects of such differences have not been studied.

Although much remains to be learned about cognitive influences on ratings, ratings will not be accurate if the rater is afraid to give feedback or is concerned about the negative impact of poor ratings on ratees' willingness to work hard. It is certainly important for a rater to know how to rate accurately, but it is equally important for this person to see some positive outcomes (and few negative ones) associated with accuracy.

Practical Considerations

Ratings are the most frequently used criterion in validation studies. Thus, in order to evaluate the usefulness of a test, we must have accurate ratings and there must be a range of obtained scores. However, this is all too often the exception in real-world validation studies. Ask any consultant and he or she can tell you a full night's worth of horror stories about obtaining only ratings of 4 on a 1–5 scale. The second author recently had a case where every individual in the validation study had received a 3 on the performance appraisal scale, because supervisors knew they had to enter some rating into the computerized system but did not want to differentiate among their subordinates. Even when there is variance in the reported performance appraisal scores, all too many of the correlations between two sets of raters are extremely low, suggesting that the ratings are unreliable.

For the aforementioned reasons, many testing firms have tried to move away from a reliance on performance appraisal ratings as a criterion, to the use of objective data including turnover. Turnover, in particular, serves as an attractive criterion in that it is objective, measured easily, simple to place a monetary value on, and understood easily by management.

Objective criteria, however, are no panacea. Turnover may be highly subject to the whims of temporary economic and labor force conditions. Or consider another widely used criterion: sales in dollars. Would you rather sell air conditioners in Texas or Alaska? Objective criteria are often influenced by a large number of variables beyond the control of the individual. In addition, for many jobs in the U.S. economy, easily obtainable objective criteria do not exist.

Thus, for validation purposes, we will continue to rely upon performance appraisal ratings. This means that we will have to continue to look for methods for improving the appraisal process.

Of course, ratings of performance, whether on the job or from an interview, may also be used as a predictor or as the source of a hiring or promotion decision by the manager. This may be a formal process or the result of informal consideration. When used to make selection decisions, ratings of performance constitute a *test* and, as a result, should be validated.

Discussion Topics

1. What factors might reduce a rater's motivation to provide accurate ratings? Which rating errors are due to motivation, and which are due to cognitive limitations?
2. How would you provide evidence for the validity of a performance rating system for use in making promotion decisions?
3. For what types of jobs might objective criteria, such as sales, be a better criterion than performance ratings?

References

Balzer, W. K. (1986). Biases in the recording of performance-related information: The effects of initial impression and centrality of the appraisal task. *Organizational Behavior and Human Decision Processes, 87,* 707–721.

Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77,* 975–985.

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisals: Assessing human behavior at work.* Boston, MA: Kent Publishing.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205–212.

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales (BARS). *Journal of Applied Psychology, 66,* 458–463.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64,* 410–421.

Borman, W. C. (1986). Behavior-based rating scales. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 100–120). Baltimore, MD: Johns Hopkins University Press.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1992). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco, CA: Jossey-Bass.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cole, M. (1953). *Robert Owen of New Lanark.* New York, NY: Oxford University Press.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90,* 218–244.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116–127.

Day, D. V., & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80,* 158–167.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York, NY: McGraw-Hill.

Judge, T. A., & Cable, D. M. (2004). The effect of physical height on workplace success and income: Preliminary test of a theoretical model. *Journal of Applied Psychology, 89,* 428–441.

Kane, J. S. (1987, April 22). *Wish I may, wish I might, wish I could do performance appraisal right.* Unpublished manuscript, School of Management, University of Massachusetts, Amherst, MA.

Kraiger, K. (1990, April). *Generalizability of performance measures across four Air Force specialties* (Technical Paper AFHRL-TP-89–60). Brooks AFB, TX: Air Force Systems Command.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology, 70,* 56–65.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72–107.

Latham, G. P., & Wexley, K. N. (1981). *Increasing productivity through performance appraisal.* Reading, MA: Addison-Wesley.

Lawshe, C. H., & Balma, M. J. (1966). *Principles of personnel testing* (2nd ed.). New York, NY: McGraw-Hill.

Lawshe, C. H., Kephart, N. C., & McCormick, E. J. (1949). The paired comparison technique for rating performance of industrial employees. *Journal of Applied Psychology, 33,* 69–77.

Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management, 30,* 881–905.

Murphy, K. R., Balzer, W. K., Lockhart, M. C., & Eisenman, E. J. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology, 70,* 72–84.

Murphy, K., & Cleveland, J. (1991). *Performance appraisal: An organizational perspective.* Boston, MA: Allyn & Bacon.

Murphy, K., & Cleveland, J. (1995). *Understanding performance appraisal: Social, organizational and goal-oriented perspectives.* Newbury Park, CA: Sage.

Oppler, S. H., Peterson, N. G., & McCloy, R. A. (1994, April). *A comparison of peer and supervisory ratings as criteria for the validation of predictors.* Paper presented to the Society for Industrial and Organizational Psychology, Nashville, TN.

Putka, D. J., Hoffman, B. J., & Carter, N. T. (2014). Correcting the correction: When individual raters offer distinct but valid perspectives. *Industrial and Organizational Psychology, 7*(4), 543–548.

Sackett, P. R., & DuBois, C.L.Z. (1991). Rater-ratee race effects on performance evaluation: Challenging meta-analytic conclusions. *Journal of Applied Psychology, 76,* 873–877.

Schoorman, F. D. (1988). Escalation bias in performance evaluations: An unintended consequence of supervisor participation in hiring decisions. *Journal of Applied Psychology, 73,* 58–62.

Sisson, E. D. (1948). Forced choice—The new army rating. *Personnel Psychology, 1,* 365–381.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149–155.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25–29.

Tinsley, H.E.A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology, 22*, 358–376.

Waung, M., & Highhouse, S. (1997). Fear of conflict and empathic buffering: Two explanations for the inflation of performance feedback. *Organizational Behavior and Human Decision Processes, 71,* 37–54.

13

INDIVIDUAL AND GROUP ASSESSMENT

Complex Candidate Judgments, Individual Assessment, and Assessment Centers

A single predictor of performance is rarely as useful as several. Most personnel decisions are based on multiple assessments. For simple jobs, formal assessment of one truly critical trait may be enough, but even that assessment is likely to be augmented by other information; more complex jobs call for more complex assessment programs.

The usual prototype of multiple assessment is a battery of tests combined to predict a single criterion. Scores on the several tests are added (with or without weights) to form a composite score which, by itself, has no particular meaning beyond a predicted criterion level. Prediction, based on linear multiple regression, is enhanced when scores on each test in the battery predict the criterion and have low correlations with each other, that is, are not redundant. The use of additive, compensatory models is well established and not to be abandoned capriciously. However, we need a new concept of compensatory batteries. Essentially, the additive model is described with the word *and*: The decision is based on a composite consisting of test A *and* test B *and* test C, and so on. What is often needed is compensation by alternatives where the operative term is *either or* or *if then.* This seems especially necessary under the accommodation provisions of ADA. It is also necessary when making complex judgments about the relative strengths of candidates. Such decisions occur all the time in real organizations, especially in final selection decisions for managerial and higher level positions. The traditional additive, compensatory model does not always apply neatly and thus we need other, more complex judgments.

Individual assessments and group assessment centers combine multiple assessments judgmentally, not statistically. Of course, what we know of judgmental versus statistical prediction suggests that deviation from an additive, compensatory model can have the serious disadvantage of being less valid. This chapter begins

with a brief discussion of the issues involved with combining multiple predictors. This chapter also considers two special cases of multiple assessment, that is, *individual assessment* and group *assessment centers* that usually include assessments of performance on specially developed exercises.

Individual Assessment

Individual psychological assessment is commonly used for assessing the suitability of candidates for executive positions or for specialized assignments, such as law enforcement agents (Highhouse, 2002). Characteristic of these positions is a situation where job performance is difficult to define and relatively few people occupy the roles. Individual assessment can be quite expensive, ranging anywhere from $1,500 to $10,000 per candidate. Because of the high prices charged by assessors, along with the fact that assessor performance is difficult to judge, the practice attracts many charlatans. Recently, individual assessment has also increased its reliance on unproctored Internet based testing (see Chapter 10).

The high cost of individual assessment also makes it prohibitive for all but the highest level hires. Executive-level assessment requires executive-sounding dimensions of assessment. Table 13.1 shows dimension labels used for assessment centers, which usually are used for supervisors and middle managers, compared with those used for individual assessment. Although it is quite likely that the assessment center dimensions involve the same behavioral criteria as the individual assessment dimensions, the latter sound much more appropriate for a high-priced individual assessment of an executive.

The distinguishing feature of individual assessment, of course, is that it assesses one person at a time. Ryan and Sackett (1987) said that an equally important and defining feature is that one psychologist conducts a final, integrating interview and one psychologist (maybe the same one) writes the assessment report. Very little research on individual assessment has been reported, and most published reports

TABLE 13.1 Comparison of Assessment Center Dimension Labels With Individual Assessment Dimension Labels

Assessment Center Dimension	*Individual Assessment Dimension*
Judgment	Seasoned Judgment
Persuasiveness	Influencing and Negotiating
Analysis	Visionary Thinking
Work Standards	Shaping Strategy
Behavioral Flexibility	Leadership Versatility
Sensitivity	Inspiring Trust
Oral Communication Skill	Building Organization Relationships
Tenacity	Driving Execution

are old (cf., Morris, Kwaske, & Daisley, 2011). Nevertheless, individual assessment is alive and well as an area of professional practice, if not as an area of research (Jeanneret & Silzer, 1998; Silzer & Jeanneret, 2011).

Analytic Versus Holistic Approach

The analytic approach to employee assessment and selection originated with Max Freyd (1923) and Walter Van Dyke Bingham (Bingham & Freyd, 1926), who outlined the steps in the traditional validation model (see Chapter 1). In general, these steps involve using standardized procedures that have a demonstrable relation to performance on the job. They emphasize the need to use data-based insights and to keep intuition from creeping into the judgment process (Freyd, 1925; Highhouse, 2008).

An alternative to the analytical viewpoint is the holistic approach to assessment (Office of Strategic Services, 1948; Viteles, 1925), which is based on the notion that prediction of future success requires taking into account the "whole" person. Measurement is subordinate to expert diagnoses. Expert intuition is used, not only to gather information, but also to properly execute data combination. Viteles, for example, objected to the practice of making decisions about applicants on the basis of scores alone. According to Viteles, the psychologist in industry considers all the data, much like a physician, and makes a diagnosis of the prospective employee. Henry Murray (Office of Strategic Services, 1948) similarly rejected the prevailing wisdom that consistency is key to good measurement. Murray encouraged examiners to vary testing procedures from candidate to candidate and to give special attention to tests they preferred.

The cornerstone of the holistic viewpoint—that one may develop expertise in predicting future performance—has never found much support (Camerer & Johnson, 1991). For example, Walter Dill Scott observed, "As a matter of fact, the skilled employment man probably is no better judge of men than the average foreman or department head" (Scott & Clothier, 1923, p. 26). The first empirical test of the assertion that experts could better integrate information holistically than analytically was conducted by T. R. Sarbin in 1943. The results showed that admission counselors who had access to test data and interview observations did significantly worse in predicting first-year student success than a simple (high school rank + aptitude test score) formula. Paul Meehl's classic 1954 book *Clinical Versus Statistical Prediction* summarized the lack of support for clinical intuition in making predictions. In addition, early organizational studies seemed to support Meehl's conclusions (e.g., Huse, 1962; Meyer, 1956; Miner, 1970). Large-scale testing programs at Exxon and Sears in the 1950s demonstrated that using standardized, data-driven approaches to identifying high-stakes talent can be quite effective (Bentz, 1967; Sparks, 1990).

Previously, in Chapter 8, we outlined a number of lay assumptions about applicant assessment (vs. statistical formulae) that are considered conventional

wisdom—even though none of them hold up against the evidence. Reviews of the organizational literature have shown that statistically integrating test scores almost always outperforms human expertise and results in up to 50% improvement in prediction (Kuncel, Klieger, Connelly, & Ones, 2013). As modern methods of machine learning do even better than traditional statistical combination, one can only expect the gap between the performance of statistical versus intuitive integration to grow larger.

Individual Assessment in Practice

Ryan and Sackett (1987) surveyed members of SIOP and found that those doing individual assessments are likely to be full-time, licensed consultants and to be one of several in the organization who do such assessments. Many who conduct individual assessment, however, are not SIOP members, and were trained in subdisciplines such as clinical, counseling, or educational psychology.

The SIOP respondents reported many purposes of individual assessment; selection, promotion (including planning for succession), and outplacement were the major ones. They also reported that assessment typically required at least a half day; some were shorter, and some required two full days. Assessment tools for individual assessors usually include personal history data, ability tests, personality and interest inventories, interviews, and often projective devices. The general pattern for arriving at conclusions about assessees is strictly judgmental; mechanistic techniques for setting specific composite scores as cutoffs for recommendations tend to be unpopular.

Information about organizations and positions was gathered typically through conversations and interviews, *not* from more systematic organizational and job analyses. Information sought included the usual emphasis on tasks and responsibilities, KSAs, and critical incidents involving prior successes and failures. Individual assessments were thought to need a wider variety of information than is common in job analysis: Interpersonal relationships, supervisory expectations, and broad statements of functions were common, and some respondents mentioned such considerations as organizational climate, opportunities for advancement, subordinate characteristics, and the criteria used in evaluating performance in the position.

Written reports usually were followed by telephone or face-to-face discussions with the client. Reports rarely included actual test scores, but strengths and weaknesses and suggestions for personal development usually were included. Reports did not necessarily include recommendations; about one-third of the respondents reported making ratings on specific traits or expected performance dimensions.

Improving Individual Assessment

People who conduct individual assessments are often extremely confident in their ability to make clinical judgments. When asked to provide validity evidence, the common response is to say that psychometrics does not apply to this kind of selection situation, or that assessors would not be in business long if they were not valid (Hanson & Conrad, 1991). Both statements are false; individual assessment is subject to the same standards as any other method of assessment for selection, and many people pay for selection techniques that are unsound. Individual assessment programs are open to other criticisms, such as the following:

1. Individual assessment rarely is subjected to serious validation efforts. Traditional validation is often not possible, but job-related constructs could be identified and evidence could be acquired to evaluate the validity of inferences drawn. Program evaluation methods could also be used, at least in firms doing a lot of assessment.
2. Assessment conclusions are often unreliable. Different assessors evaluate candidates differently, perhaps because they rely on different information and perhaps because they have no standard basis for consistency. An intractable reliability problem exists insofar as different readers of a report draw different inferences from it. Three assessors described by Ryan and Sackett (1989) did not agree on ratings of suitability for the position.
3. Assessment summaries are influenced too often by one or two parts of the assessment program that could have been used alone. This is not surprising; assessment summaries are judgments of the report writer, and judgment research shows that judgments typically are based on only a few of the available (usually negative or early) cues.
4. Great emphasis is placed on personality assessment without matching evidence of the relevance of the traits assessed. Where assessments are statistically validated against job performance criteria, scores on one or two traditional cognitive tests are usually more valid than scores or clinical judgments based on personality tests.
5. Individual assessments, limited to one person, cannot assess interpersonal skills from actual interpersonal behavior. Most individual assessment is done with candidates for managerial or sales work, work that requires interaction with others. Assessment without such interaction may be deficient; this may be one reason why group assessment center approaches have dominated the assessment literature in recent years.
6. It may be ethically and legally questionable to seek information not explicitly relevant to the work to be done, yet individual assessments typically include intellectual and personality exploration, gathering general and diverse data about a person. Many people think collecting information without direct job relevance is an unwarranted invasion of privacy. On the other hand, some

consider it unfair or unethical to base decisions on one or two traits without a complete picture of the individual.

All of these points can be answered by appropriate design. Validation efforts combining evidence of relevance of traits assessed with evidence of the construct validities of the assessments provide better validity evidence than a single validity coefficient; that is, well-developed predictive hypotheses should dictate and justify assessment content. Greater use of work samples (or of exercises based on them) could provide easy justifications of content. The absence of interpersonal behavior in the assessment process itself is not so serious if personal records of interpersonal achievements can be assessed by achievement records and other biodata, or where interview structure focuses on such history.

Assessment Centers

Instead of assessing characteristics of one person at a time, *assessment centers* assess small groups of people at more or less the same time. Instead of one person being responsible for the final assessments, a group of observers may work together to form a consensus about assessees. Like individual assessments, assessment center programs use multiple methods of assessment to make multiple assessments. The methods may not include literal work samples or simulations, but they nearly always include exercises chosen to reflect a major aspect of job performance.

Whereas *individual assessment* deemphasizes structure and emphasizes the expertise of the assessor, the modern *assessment center* emphasizes structure and deemphasizes the role of assessor expertise.

Assessment Center Purposes

Most assessment centers are organization specific. Consulting firms may provide generic assessment center services, primarily for smaller organizations, but they are more likely to assist organizations in developing their own programs. Most assessment centers, especially for managerial purpose, are built around organizationally specific values and practices. This may be because many of them are used for employee development and feedback, rather than for employee selection. They are not always designed for managers; many are for sales people or public safety jobs.

In this section, we describe the traditional assessment center. Such assessment centers featured multiple assessments, multiple assessors, and multiple assessees, and were usually conducted off-site in a nice hotel or at a nice resort. However, as with every other area of assessment, technology has changed the assessment center. Today, many testing firms offer computerized assessment centers of varied time durations. Some computerized assessment centers are quite simple, while others present very complex, virtual worlds. But all allow the assessment process to be conducted in a shortened time period, at reduced cost, right at the individual's desk, or potentially even on his or her smartphone.

Purposes differ within occupational categories. Thornton and Byham (1982) divided managerial assessments into those for early identification of potential managers, for promotions, or for management development. The different purposes call for differences in program design. Different purposes may call for assessments of different constructs. Some diagnostic purposes may require psychologists or educators as assessors, but other developmental purposes may be better accomplished with managers as assessors—managers similar to those to whom assessees will later report. An overall assessment rating (OAR) may have no importance for diagnostic purposes but may be crucial for personnel decisions like hiring or promotion. The discussion here focuses on selection or promotion.

Assessment Center Components

An organizing principle of assessment center development is that the program should be a *multiattribute* assessment, assessment on several dimensions relevant to the decision to be made. A further principle is that the assessments should not depend on specific methods of assessment—they should be *multimethod* assessments. Any attribute is to be assessed by more than one method. The multimethod aspect of assessment center programs is not merely a matter of numbers for increasing reliability, although it may serve that purpose. The reason is "rather that the process of seeking confirmation from several exercises leads to more validity of measurement of complex dimensions" (Thornton & Byham, 1982, p. 227). It leads to greater validity through more comprehensive domain sampling and through the convergent evidence of validity of dimensional assessments.

Assessment centers have many components. Job or task analysis provides background; special exercises are based on task analyses. Standardized tests are often chosen for important KSAs. Any component should be clearly relevant to the job, provide reliable and valid assessments, and contribute meaningfully to an OAR, if one is used. Some varieties of assessment procedures are briefly discussed here, but the list is not at all exhaustive.

Tests and Inventories. Traditional tests and inventories are included in most assessment centers. Their role in an OAR raises questions. How should they be combined with various ratings? Statistically? In an additive model with nominal weights? If given to the assessors as information to consider with exercise ratings in arriving at the OAR, should they be given as raw scores, or as *z*-scores, percentiles, or other interpretive scores?

Tests and inventories, by themselves, often have validities as high as any overall composite of assessment center components. This may be due to superior reliabilities but, whatever the reason, should they be given credence beyond that of the exercises? This is in part a reprise of the performance test versus traditional test issues of Chapter 10, but it is more than that. When considering the importance of interpersonal skills in many kinds of jobs, do the group exercises provide more important assessments than those obtained with traditional tests and inventories?

Exercises. Most assessment center exercises are performance tests; they are samples or abstractions of aspects of the jobs for which people are assessed. Many are low in fidelity to the job, but high-fidelity simulations would be inappropriate for assessees who do not yet know the job. It is content sampling of sorts, but it is content suggested by (highly abstracted from) the job, not literal job content.

The most frequently used assessment center simulation is an *In-Basket* exercise. In-Basket tests simulate administrative work, usually with a set of reasonably typical memos, clippings, letters, reports, messages, and even junk mail that can accumulate on a person's desk. Instructions generally tell the assessee to play the role of a person new to the job, working when no one else is around, trying to clear the desk; In-Baskets are not group exercises. Materials range from simple to complex, from trivial to urgent, and are often interrelated. Additional documents may or may not be provided as reference material (e.g., a file cabinet containing both relevant and unrelated items of information). The assessee may be interviewed after the exercise to explain reasons for actions taken, with ratings based on the interview. Some In-Basket tests, however, have scoring protocols and require no further information from the assessee.

An equally common exercise, not clearly a simulation, is the *Leaderless Group Discussion.* The group is given a problem to solve, a time limit in which to do so, and perhaps a requirement for a written solution. No one is assigned the role of chair; leadership functions must emerge during the discussion. Specific roles might be assigned to the various group members, often with the competitive requirement of trying to convince others to adopt a particular position. Many variants on the theme have been used.

Interviews. Assessment centers usually use interviews, but they are not like employment interviews. Various examples include stress interviews, interviews as

role-playing simulations, and panel interviews. Ordinary problems of interviews occur in these, but assessment center interviews can be more standardized, without being test like, than other interviews.

> An example of a *role-play* simulation might include having the job candidate play the part of a candy bar sales representative who is charged with making a phone call to a buyer for a chain of gas stations. The job candidate would have to convince the buyer (played by the assessor) to stock the new candy bar in the gas station convenience stores. The buyer would make the sales representative's job more difficult by feigning indifference and asking difficult questions about the product's characteristics.

Assessors

Functions of Assessors. Zedeck (1986) identified three assessor functions. One major function is as an *observer and recorder* of behavior in the exercises. Behavior is commonly recorded in descriptive (and perhaps evaluative) reports written about the observations. Fulfillment of this function requires careful, standard training. Ratings for a given dimension may be made by different assessors in different exercises. Differences in the behavior observed and the dimensional inferences drawn from it are necessarily attributable in part to the differences in exercises, but they should not be attributable to different assessors having different understandings of the nature of the dimension. A related problem is that the observers may also be part of the stimulus, and different observers may stimulate different reactions. Videotaped exercise performance may help with this; the assessors can observe tapes, even with "instant replay" if needed (Ryan et al., 1995).

A second function is as a *role player,* an active participant in an assessment exercise. In many exercises, assessors are interviewers, usually with another assessor in a purely observer role. In such exercises, an assessor serves as a stimulus to which the assessee responds. One problem for this function is lack of standardization. Role players may change their own behavior during the sequence of interviews. In a stress interview, for example, some may become harsher over a sequence of interviews, whereas others may say, in effect, to heck with it—and cause less stress. If different assessors play the same role, standardization is still more unlikely. Trying to be an actor and an observer simultaneously is cognitively difficult, and assessors are unlikely to be good actors. It is probably best if assessors are as unobtrusive as possible.

Zedeck's third function is as a *predictor.* Assessors may make explicit predictions, or prediction may be based on the ratings, whether they are dimensional ratings or OARs.

Assessor Qualifications. Assessors may be psychologists, HR staff, or job experts (e.g., managers in managerial assessment centers). Staff psychologists may be assessors with managers, they might chair assessor panel discussions, or they might simply be resource people. Assessors from whatever source should receive intensive training with frequent refreshers; they should be fully familiar with the exercises and the kinds of behavior they might observe, and they should fully understand the language and concepts related to the ratings they are asked to make.

Numbers of Assessors Needed. Typically, the ratio of assessees to assessors is 2:1. Thornton and Byham (1982) considered this a desirable ratio. Cognitive demands on observers are heavy and can be reduced by adding more assessors, but that can be daunting for the assessees. Using fewer assessors over a longer time period viewing video recordings may be better.

Other questions emerge. Should assessors become specialists? Should one assessor be a specialist in the leaderless group discussion and another a specialist in personal history interviews? Perhaps a specialist for certain dimensions? In group exercises, should each assessor try to observe and rate all candidates in the group or be assigned to observe and rate no more than two at a time? These are questions about ways to use assessors as observers and raters to maximize reliability and validity of the assessments provided. They must be answered locally; no general answers have been found empirically.

Dimensions to Be Assessed

There is disagreement about the dimensions to be rated. The dimensions (constructs) might be personal traits, job-defined competencies, or performance levels on aspects of jobs reflected in simulations. Assessors might be asked to rate only overall performance in an exercise, or perhaps component aspects of exercise performance. Traits rated might be generalized, habitual behaviors. Task performance may be rated in terms of outcomes or processes. A dimension can be defined by behavior exhibited only in particular kinds of situations. All of these constructs, except the last one, should generalize across situations, therefore across exercises. The last one is an idea of a dimension that traditionally has not been espoused in the assessment center literature. An example of typical assessment center dimensions for a supervisory position is shown in Table 13.2.

In early assessment centers, personal traits, largely personality traits that were thoroughly defined and discussed by psychologists like Henry Murray, were rated. More recent ones favor behavioral categories that, unfortunately, are often poorly defined. The literature on assessment center dimensions is not exemplary. More bluntly, much of it is silly. Dimensions are given names, but the names are not defined. Trait constructs are often rejected by some who mistakenly define traits simply as personality variables that "cause" behavior, apparently irrespective of circumstances. Task competencies are often rejected because they are seen as being

TABLE 13.2 Typical Assessment Center Dimensions and Their Definitions

Dimension	*Definition*
Tolerance for Stress	Stability of candidate's performance under pressure, opposition, or both.
Oral Communication Skill	Effective expression in individual or group situations (including gestures and nonverbal communication).
Work Standards	Setting high goals or standards of performance for self, subordinates, others, and the organization.
Persuasiveness/Sales Ability	Utilizing appropriate interpersonal styles and methods of communication to obtain agreement or acceptance of an idea, plan, activity, or product from clients.
Sensitivity	Actions that indicate appropriate consideration for the feelings and needs of others.
Behavioral Flexibility	Modifying behavior to reach a goal when obstructed by the attitudes, beliefs, opinions, or behavior of another person or persons.
Analysis	Relating and comparing data from different sources, identifying issues, securing relevant information, and identifying relationships.
Judgment	Developing alternative courses of action and making decisions that are based on logical assumptions and that reflect factual information.
Organization Sensitivity	Perceiving the impact and implications of decisions on other components of the organization.
Tenacity	Staying with a plan of action until the desired objective is achieved or is no longer reasonably attainable.

concerned only with outcomes, another unwarranted restriction in definition. Thornton and Byham (1982) preferred to refer to "behavioral dimensions" (p. 118)—dimensions inferred from job analysis, defined behaviorally in terms of directly observable behaviors, and free of any inferences about underlying personality traits.

It seems the real issue is not whether the dimensions should be called traits, competencies, or behavioral dimensions; after all, they can all be defined in behavioral terms. The real problem is in the operations defining the dimensions. For most assessment center exercises, ratings are the assessments, so the issue is the typical problem with ratings.

Ratings. Exercises stimulate behavior, the behavior is observed by assessors, and the observations are the foundation for the ratings—which, like scores on tests, must have a meaning to be validated. If ratings of an attribute are valid, and if the same attribute is rated in two or more different exercises, then permissible inferences from ratings of the attribute should be at least somewhat consistent across those exercises; if not, they are not assessing a common construct.

The logic of multiple assessment is that prediction is better, because assessment is more reliable when assessments are replicated. In assessment centers, replication implies measuring a predictor in more than one way. If the predictors are the rated dimensions, then assessments (ratings) of an attribute in one exercise should generalize to (be correlated with) assessments in another. This is not a psychometric statement of parallel or equivalent forms; it is a statement that, if two exercises are designed so that they reveal, for example, skill in oral communication, then the communication effectiveness in one should be similar to the communication effectiveness in the other. As we see later, it rarely works out that way.

Overall Assessments. Most assessment centers call for an overall summary rating, the OAR. Different programs use different procedures for developing the OAR. Is the OAR an operational definition of a definable attribute? It might be, but often it is not; it is likely to be analogous to the composite score computed implicitly in multiple regression, when the composite is simply a complex predictor variable composed of a set of essentially unrelated but valid predictors. Other procedures might call for a mechanical averaging of overall ratings given to individual candidates by the various assessors. More commonly, however, there is a consensus meeting at which candidates are discussed, ratings on attributes are agreed on, independent OARs are made and shared, differences are discussed and resolved, and a consensus achieved.

Although many people consider the consensus meeting a key feature of the assessment center concept, research does not support its effectiveness. The entire group discussion process can take several days to complete, and no mechanical or statistical formulas are used. Nevertheless, Oldfield (1947) noted that "Discussion of the merits of candidates merely amounts to a somewhat clumsy method of averaging the individual judgments of the members" (p. 129). Sackett and Wilson (1982) found, for example, that a simple average of assessment center dimension ratings predicted post-discussion ratings 94% of the time. Arthur, Day, McNelly, and Edens (2003) meta-analyzed assessment center dimension scores and found that the OAR is a less valid predictor than some of the dimension scores *alone* (e.g., organizing and planning; problem solving; influencing). Dilchert and Ones (2009) examined assessment scores for nearly 5,000 managers, along with their scores on a cognitive ability test and a measure of the five-factor model of personality. The researchers found that the typical consensus-based OAR provided no incremental validity over the ability and personality measures in predicting managerial success. Simply adding the dimension scores together, however, resulted in meaningful incremental validity over the test sores.

Assessment Center Problems and Issues

There is no orthodox "one best way" in assessment center design; each program is different, in part to fit its different set of circumstances. The differences also highlight some problems of program design and some issues on which experts may disagree.

Construct Validities of Dimension Assessments

The biggest issue focuses on the dimensions, or constructs, rated by observers. Can internally consistent constructs be assessed validly by substantially different assessment exercises? Converging validity evidence is consistency in assessments of the same construct across exercises in which it is rated. Lack of convergence may not indicate invalid ratings; it may indicate only confusion about what is being assessed.

Dimensional Consistency. If the attribute (construct or dimension) is defined and the exercises developed so that the attribute rated reflects the same construct in two different exercises, then the correlation between the two ratings on the dimension should be, not high, but substantial. Exercises are not designed as parallel forms, so correlations need not approximate reliability coefficients. The multitrait–multimethod logic should apply, however, where different exercises are intended to tap the same constructs. Correlations between ratings of the same dimensions on different exercises should be larger than the correlations between ratings on different dimensions within the same exercise; factor analysis of such a matrix should yield factors consistent with the dimensional constructs.

> One of the great and enduring debates has been over whether assessment centers actually measure the dimensions or constructs they purport to measure, or whether they are really assessing exercise dependent behaviors.

Results of Factor Analyses. Factor analysis results for a police assessment center are summarized in Table 13.3.[1] Clearly, the factors are defined, not by the dimensions, but by the exercises. The factor analysis provides no support for the construct validity of the dimension ratings and, in fact, supports the alternative position that dimension ratings are exercise specific rather than generalizable over exercises (and, by extension, to comparable aspects of performance on the job). This is not an isolated example. Sackett and Dreher (1982) analyzed data from three independent assessment centers and, in all three cases, found factors that were defined by exercises, not attributes. Many others have reported similar results.

Silverman, Dalessio, Woods, and Johnson (1986) looked for convergence experimentally in an assessment center with three exercises, each rated on the same six dimensions. In a within-exercise method, candidates were rated on each relevant dimension immediately on the conclusion of the exercise. A within-dimension method, a modification of the AT&T procedure, made dimension ratings after a staff conference. In the within-exercise method, all factors were exercise factors. Results were less clear for the within-dimension method, however. It also gave three factors, somewhat like the three within-exercise factors, but there were strong secondary loadings. Ratings of leadership, for example, had factor loadings of at

TABLE 13.3 Rotated Factor Pattern of Police Assessment Center Ratings

Exercise and Dimension	*Factor*[a]						h^2	*rxx*[b]
	I	*II*	*III*	*IV*	*V*	*VI*		
Leaderless Group Discussion (NR)								
Assertiveness—1[c]	30			88			88	70
Persuasiveness—1				88			87	66
Oral Communication Skill—1			81			75	58	
Assertiveness—2	32			82			80	75
Persuasiveness—2				60			40	64
Oral Communication Skill—2			82			78	40	
Leaderless Group Discussion (AR)								
Assertiveness—1	87						86	79
Persuasiveness—1	88						86	73
Oral Communication Skill—1	79						70	65
Assertiveness—2	82						78	76
Persuasiveness—2	57						38	68
Oral Communication Skill—2	75						63	52
Competitive Exercise								
Assertiveness—1			83				77	76
Team Spirit—1			83				72	72
Assertiveness—2			83				76	70
Team Spirit—2			86				78	67
Analysis Interview								
Reasoning—1						88	83	89
Reasoning—2						89	84	85
Stress Interview								
Stress Tolerance—1					87		78	83
Stress Tolerance—2					87		77	71
Situations Test								
Reasoning—1		88				79	42	
Written Communication Skill—1		83				69	32	
Reasoning—2		63					37	50

[a]Only factor loadings of .30 or higher are listed; decimal points omitted.
[b]Interrater reliability coefficients. These are spuriously low; acceptable reliability estimates are necessarily higher than communalities.
[c]1 indicates graphic ratings, 2 indicated checklist summated ratings.

least .35 on all three exercises. In short, these were not dimension factors supporting the validity of construct inferences from dimensional ratings, but neither did they form the clear alternative factor pattern. Their results suggest that procedural adjustments can improve the construct validity of the ratings.

Reasons for Inconsistency in Dimension Ratings. Dimensions can be viewed from two extreme points of view. Neither makes much sense but, together, they help focus on the problem posed by the factor analysis results. At one extreme the dimensions are viewed as unalterable traits (not necessarily genetic, but well-established by adulthood) exhibited consistently in behavior in virtually all circumstances. From this extreme view, multiple assessments of a dimension would serve no purpose other than increasing reliability, but they *would* converge. At the other extreme, the dimensions are simply aspects of behavior in a given situation, without generalizability to any other situation. Carrying this position to its logical extreme, prediction of future performance is impossible; even generalization from one exercise to another, where both call for somewhat similar behaviors, is unlikely.

The extremes are obviously false; behavior can have both typical and situationally determined components. Low consistency across exercises may reflect inconsistent behavior; typical behavior may not be elicited in atypical situations or in situations having their own intrinsic behavioral imperatives. A person who is typically judgmental and vocal may be judgmental and vocal in a leaderless group discussion, but inhibit that typical behavior in a simulation where one is to help two conflicting parties negotiate or reconcile their differences. Ratings on forcefulness of oral communication in these two exercises cannot reasonably be expected to correlate highly.

Many reasons might be offered to explain the tendency to get only exercise factors. The specificity of situational demands, described earlier, is one. Another is that rating dimensions with the same names does not necessarily mean the same constructs were rated. Constructs to be rated are rarely defined thoroughly; the usual case settles at best for a brief definitional phrase. Developers of assessment centers, perhaps even more than developers of other assessment methods, need to be extraordinarily precise in presenting their dimension definitions to the raters who must use them.

Solutions. The most common suggestion for solving the problem is the use of behaviorally based ratings or checklists. Reilly, Henry, and Smither (1990) asked assessors to write examples of behavior corresponding to dimensions they had rated previously on 5-point scales. A large pool of items remained after editing, and the Smith–Kendall retranslation procedure was used (Smith & Kendall, 1963). Items that were almost always assigned to the intended dimensions were placed in checklists for the dimensions, and assessors were instructed to use the checklists immediately after an exercise, indicating 0 *(behavior did not occur),* 1 *(behavior occurred once),* or 2 *(behavior occurred more than once).* Ratings on the dimension, for that exercise, were then made on the same 5-point scale previously used. Much better convergent validity was found when the checklist was used.

The use of a different sort of construct might help. Joyce, Thayer, and Pond (1994) classified possible dimensions as either person oriented ("traditional" dimensions) or task oriented (their alternative to traditional dimensions). Examples of task-oriented dimensions included "Structuring and staffing tasks: Allocating manpower and resources to tasks, delegating assignments, and organizing the work of subordinates," and "Establishing effective work group relationships: Recognizing, praising, and encouraging employees and co-workers; maintaining a high level of morale" (Joyce et al., 1994, p. 113). A natural experiment was possible because two essentially parallel assessment centers were run by the same organization. One of these assessed managers as they entered a management training program, the other assessed them again two years later at the completion of the program. The first used traditional dimensions; the second used task-oriented dimensions. Factor analysis of the dimensions, whether personal attributes or job functions, resulted in factors defined by exercises, not by either of the alternatives.

Recent meta-analytic research has suggested that the OAR may just be a combination of a small number of cognitive and personality factors (Arthur et al., 2003; Collins et al., 2003). For example, Arthur and his colleagues (2003) found that the dimensions of problem solving and influencing others had correlations with performance that, by themselves, *exceeded* the OAR-performance correlation. Collins and her associates (2003) similarly found that cognitive ability and extraversion accounted for most of what was being measured by the OAR. It may be that a small number of dimensions would be sufficient for prediction purposes, and may result in the expected dimension factors.

Using meta-analysis to examine whether assessment center methodology had an impact on the construct validity of dimension ratings, Woehr and Arthur (2003) found that better convergent validity across exercises was found when fewer dimensions were rated by the assessors. The researchers also found that dimension ratings showed better convergent and discriminant validity when ratings were made by dimension *across exercises,* rather than *within exercises*.

Criterion-Related Validities

Despite many problems, assessment centers have amassed a good record of criterion-related validities. We do not know the underlying constructs, and we have little evidence to say that ratings of these (usually) poorly defined constructs are valid assessments. Nevertheless, dimension ratings and OARs have been valid predictors of future performance. Meta-analyses have found mean corrected validity coefficient in the area of .36 to .37, with a lower bound of the 95% confidence interval well above 0, indicating generalized validity (Arthur et al., 2003; Gaugler, Rosenthal, Thornton, & Bentson, 1987).

There are, however, differences in validities across studies. This is not surprising; different assessment centers have different exercises, performance is rated on

different (and different kinds of) dimensions, rater training varies widely, purposes differ, and different kinds of criteria are predicted. A review of meta-analytic results suggests the following conclusions:

1. Predictive validity is higher when more kinds of assessment exercises are included. The multiple exercises should be multiple samples of job-related behavior, and multiple assessments of the same constructs.
2. Validities are higher in those assessment centers where peer evaluations are included. This point, and the preceding one, may merely indicate that more thoroughly developed programs are more valid.
3. Assessors' backgrounds and training moderate validity. OARs in assessment centers using psychologists as assessors are more valid than those where managers were the assessors.
4. Although the typical assessment center involves ratings on 10 or more dimensions, approximately 4 dimensions account for the assessment center's ability to predict performance. Demanding fewer dimension ratings from assessors is also related to better convergent validity across exercises.
5. Validities are much higher for ratings of potential for management progress than for predictions of future performance.

Assessor judgments appear to be highly influenced by the candidate's ability to *solve problems* and *influence others*. Exercises used in traditional assessment centers may be best suited to the assessment of these two abilities.

Some skepticism is warranted. Apparent validity may be attributed to common stereotypes. A "good leader" stereotype can influence both assessments and criteria, providing no more than an illusion of validity. A contaminating source of variance destroys validity—unless it contaminates both assessment and criterion, in which case it gives an illusion of increasing validity by increasing the coefficient.

A Point of View

Both educational and managerial assessment people have been uncomfortable with traditional testing, questioning its appropriateness for their purposes. The discomfort has not been the traditional concerns over reliability and validity so much as a concern about whether the right things have been measured. In-Basket tests, for example, became widely used assessment center exercises not because of superior reliability or predictive power, but because they tap the everyday decision-making skills of managerial work.

Some validity coefficients for assessment centers, especially when ratings of potential are correlated with advancement over time, are very high. Others, however, seem ordinary, even low. The corrected validity coefficient for predicting performance was .36–.37 in the meta-analyses. Corrected validity coefficients for some paper-and-pencil methods are as high as .50. Why would one go to the trouble and expense of developing an assessment center that, on the average, might yield a lower validity than achieved by less expensive, more traditional methods? And how much confidence can one have that even the best validities cannot be explained away as the result of common stereotypes?

Skepticism implies questioning, not rejection. The questions should lead to research, not to abandonment or undue abbreviation of assessment centers. Some explanatory research has been done, and suggestions for redesign of assessment procedures have been offered. Many researchers explain the problem of construct validity as caused by excessive cognitive demands on assessors. Perhaps assessors should be specialists for a few dimensions or a few exercises.

Two questions seem to require data. How many dimensions are needed to arrive at a stable OAR? How many are needed to predict the criterion? For either question, assessment center development requires an analog of the item analysis procedures of traditional test development. That is, the developer should determine, for each dimension and for each exercise, what it does, in fact, contribute both to the judgment and to the prediction. This requires pilot studies and a willingness to discard and perhaps replace dimensions or exercises that do not contribute.

Discussion Topics

1. What are some ways to improve the reliability and validity of employment decisions based on individual assessments?
2. How can clinical judgment best be used in the process of assessing the suitability of a candidate for a position?
3. Viewing assessment center exercises as test items, how might we make the assessment center a more reliable and valid test?

Note

1 This was done in an unpublished study by Dennis Sweeney, then at Bowling Green State University.

References

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125–154.

Bentz V. J. (1967). The Sears experience in the investigation, description, and prediction of executive behavior. In F. R. Wickert & D. E. McFarland (Eds.), *Measuring executive effectiveness* (pp. 147–205). New York, NY: Appleton-Century-Crofts.

Bingham, W. V., & Freyd, M. (1926). *Procedures in employment psychology: A manual for developing scientific methods of vocational selection.* New York, NY: McGraw-Hill.

Camerer C. F., & Johnson E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 195–217). Cambridge, England: Cambridge University Press.

Collins, J. M., Schmidt, F. L., Sanchez-Ku, M., Thomas, L., McDaniel, M. A., & Le, H. (2003). Can basic individual differences shed light on the construct meaning of assessment center evaluations? *International Journal of Selection and Assessment, 11,* 17–29.

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic incremental validity. *International Journal of Selection and Assessment*, *17,* 254–270.

Freyd, M. (1923). Measurement in vocational selection: An outline of research procedure. *Journal of Personnel Research, 2,* 215–249, 268–284, 377–385.

Freyd, M. (1925). The statistical viewpoint in vocational selection. *Journal of Applied Psychology, 9,* 349–356.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511.

Hanson, C. P., & Conrad, K. A. (1991). *A handbook of psychological assessment in business*. New York, NY: Quorum Books.

Highhouse, S. (2002). Assessing the candidate as a whole: A historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology, 55,* 363–396.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology, 1*(3), 333–342.

Huse, E. F. (1962). Assessments of higher level personnel. IV. The validity of assessment techniques based on systematically varied information. *Personnel Psychology, 15,* 195–205.

Jeanneret, R., & Silzer, R. (1998). An overview of psychological assessment. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment: Predicting behavior in organizational settings* (pp. 3–26). San Francisco, CA: Jossey-Bass.

Joyce, L. W., Thayer, P. W., & Pond, S. B., III. (1994). Managerial functions: An alternative to traditional assessment center dimensions? *Personnel Psychology, 47,* 109–121.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology, 98,* 1060–1072.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Meyer, H. H. (1956). An evaluation of a supervisory selection program. *Personnel Psychology, 9,* 499–513.

Miner, J. B. (1970). Executive and personnel interviews as predictors of consulting success. *Personnel Psychology, 23,* 521–538.

Morris, S. B., Kwaske, I. H., & Daisley, R. R. (2011). The validity of individual psychological assessments. *Industrial and Organizational Psychology, 4*(3), 322–326.

Office of Strategic Services. (1948). *Assessment of men.* New York, NY: Rinehart.

Oldfield, R. S. (1947). *The psychology of the interview.* London, England: Methuen.

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43,* 71–84.

Ryan, A. M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T., & McCormick, S. (1995). Direct, indirect, and controlled observation and rating accuracy. *Journal of Applied Psychology, 80,* 664–670.

Ryan, A. M., & Sackett, P. R. (1987). A survey of individual assessment practices by I/O psychologists. *Personnel Psychology, 40,* 455–488.

Ryan, A. M., & Sackett, P. R. (1989). Exploratory study of individual assessment practices: Interrater reliability and judgments of assessor effectiveness. *Journal of Applied Psychology, 74,* 568–579.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67,* 401–410.

Sackett, P. R., & Wilson, M. A. (1982). Factors affecting the consensus judgment process in managerial assessment centers. *Journal of Applied Psychology, 67,* 10–17.

Sarbin, T. L. (1943). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology, 48,* 598–602.

Scott, W. D., & Clothier, R. C. (1923). *Personnel management.* Chicago, IL: A. W. Shaw.

Silverman, W. H., Dalessio, A., Woods, S. B., & Johnson, R. L., Jr. (1986). Influence of assessment center methods on assessors' ratings. *Personnel Psychology, 39,* 565–578.

Silzer, R., & Jeanneret, R. (2011). Individual psychological assessment: A practice and science in search of common ground. *Industrial and Organizational Psychology, 4*(3), 270–296.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149–155.

Sparks, C. P. (1990). Testing for management potential. In K. E. Clark & M. B. Clark (Eds.), *Measures of leadership* (pp. 103–112). West Orange, NJ: Leadership Library of America.

Thornton, G. C., III, & Byham, W. C. (1982). *Assessment centers and managerial performance.* New York, NY: Academic Press.

Viteles, M. S. (1925). The clinical viewpoint in vocational selection. *Journal of Applied Psychology, 9,* 131–138.

Woehr, D. J. & Arthur, W. (2003). The construct related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29,* 231–258.

Zedeck, S. (1986). A process analysis of the assessment center method. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 8, pp. 259–296). Greenwich, CT: JAI Press.

This page intentionally left blank

INDEX

Note: Page numbers in *italics* refer to information in figures and tables.