CHAPTER 1

INDIVIDUAL DIFFERENCES THAT INFLUENCE PERFORMANCE AND EFFECTIVENESS

What Should We Assess?

Kevin R. Murphy

Assessment in organizations can be carried out for a variety of purposes, many with high stakes for both individuals and organizations. The stakes can be particularly high when assessments are used to make decisions about personnel selection and placement or about advancement and development of individuals once they have been hired. If assessments focus on traits, attributes, or outcomes that are not relevant to success and effectiveness, both organizations and individuals may end up making poor decisions about the fit between people and jobs. If assessments are appropriately focused but poorly executed (perhaps the right attributes are measured, but they are measured with very low levels of reliability and precision), these assessments may lead to poor decisions on the parts of both organizations and individuals.

In this chapter, I focus on broad questions about the content of assessments (for example, What sorts of human attributes should assessments attempt to measure?) and say very little about the execution of assessments (the choice of specific tests

or assessment methods, for example) or even the use of assessment data. My discussion is general rather than specific, focusing on general dimensions of assessment (whether to assess cognitive abilities or broad versus narrow abilities, for example) rather than on the specifics of assessment for a particular job (say, the best set of assessments for selecting among applicants for a job as a firefighter).

This chapter provides a general foundation for many of the chapters that follow. It sets the stage by discussing broad dimensions of individual differences that are likely to be relevant for understanding performance, effectiveness, and development in the workplace. The remaining chapters in Part One start addressing more specific questions that arise when attempting to assess these dimensions. Chapter Two reviews the range of methods that can be used to assess the quality of measures, and Chapters Three through Eight provide a more detailed examination of specific domains: cognitive abilities, personality, background and experience, knowledge and skill, physical and psychomotor skills and abilities, and competencies.

Part Two of this book discusses assessment for selection, promotion, and development, and Parts Three and Four deal with strategic assessment programs and with emerging trends and issues.

I begin this chapter by noting two general strategies for determining what to assess in organizations: one that focuses on the work and the other that focuses on the person. The person-oriented approaches are likely to provide the most useful guidance in determining what to assess for the purpose of selection and placement in entry-level jobs, and work-oriented assessments might prove more useful for identifying opportunities for and challenges to development and advancement.

Two Perspectives for Determining What to Assess

A number of important decisions must be made in determining what to assess, but the first is to determine whether the focus

around the characteristics of individuals that influence what they do and how well they do it in the workplace (person oriented). For example, it is common to start the process of selecting and deciding how to use assessments with a careful job analysis on the assumption that a detailed examination of what people do, how they do it, and how their work relates to the work of others will shed light on the knowledge, skills, abilities, and other attributes (KSAOs) required to perform the job well. An alternative strategy is to start by examining the individual difference domains that underlie most assessments and to use knowledge about the structure and content of those domains to drive choices about what to assess.

The choice of specific assessments is a three-step process that starts with choosing between a broadly person-oriented or work-oriented approach, then making choices about the domains within each approach to emphasize (for example, whether to focus on cognitive ability or on personality), and finally narrowing down the choice of specific attributes (say, spatial ability) and assessment methods (perhaps computerized tests). As I noted earlier, this chapter focuses on the first two of these steps.

Work-Oriented Strategies

Different jobs involve very different tasks and duties and may call on very different sorts of knowledge or skill, but it is possible to describe the domain of work in general terms that are relevant across a wide range of jobs and organizations; such a wide-ranging description provides the basis for worker-oriented strategies for determining what to assess. Starting in the late 1960s, considerable progress was made in the development of structured questionnaires and inventories for analyzing jobs (for example, the Position Analysis Questionnaire; McCormick, Jeanneret, & Mecham, 1972). These analysis instruments in turn helped to define the contents and structure of the O*NET (Occupational Information Network; Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1999) Generalized Work Activities Taxonomy, argu-

Table 1.1. O*NET Generalized Work Activities

Information input	Looking for and receiving job-related information
	Identifying and evaluating job-relevant information
Mental processes	Information and data processing
	Reasoning and decision making
Work output	Performing physical and manual work activities
	Performing complex and technical activities
Interacting with others	Communicating and interacting
	Coordinating, developing, managing, and advising
	Administering

If you were to ask, "What do people do when they work?" Table 1.1 suggests that the answer would be that they gather information, process and make sense of that information, make decisions, perform physical and technical tasks, and interact with others. The specifics might vary across jobs, but it is reasonable to argue that Table 1.1 provides a general structure for describing jobs of all sorts and for describing, in particular, what it is that people do at work. Each of these major dimensions can be broken down into subdimensions (which are shown in this table), most of which can be broken down even further (for example, administering can be broken down into performing administrative activities, staffing organizational units, and monitoring and controlling resources) to provide a more detailed picture of the activities that make up most jobs.

and it is not clear whether competency modeling is really anything other than unstructured and informal job analysis. Nevertheless, the business world has adopted the language of competencies, and competency-based descriptions of work are becoming increasingly common.

Some competency models are based on careful analysis and compelling data, most notably the Great Eight model (Bartram, 2005):

Great Eight Competency Model

- Leading and deciding
- Supporting and cooperating
- Interacting and presenting
- Analyzing and interpreting
- Creating and conceptualizing
- Organizing and executing
- Adapting and coping
- Enterprising and performing

Bartram summarizes evidence of the validity of a range of individual difference measures for predicting the Great Eight. Unlike some other competency models, assessment of these particular competencies is often done on the basis of psychometrically sound measurement instruments.

Drilling Deeper

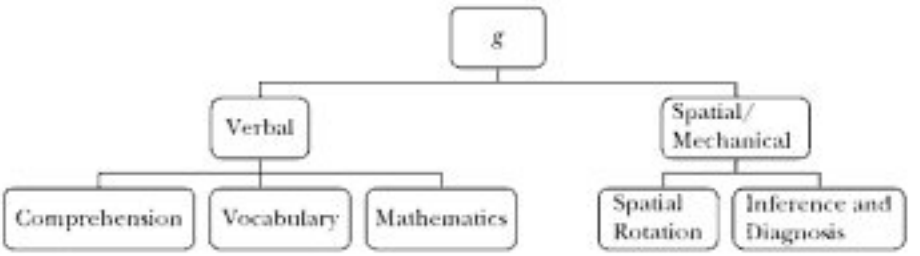
Work can be described in general terms such as the competencies detailed in the previous section. A more detailed analysis of what people do at work is likely to lead to an assessment of more specific skills and an evaluation of background and experience factors that are likely to be related to these skills. In this context, *skill* has a specific meaning: the consistent performance of complex tasks with a high level of accuracy, effectiveness, or efficiency. Skills are distinct from abilities in three ways: (1) they involve the performance of specific tasks, (2) they involve automatic rather than

in performance on a wide range of cognitively demanding tasks. At the next level (the broad stratum) are a number of areas of ability, which imply that the rank ordering of individuals' task performance will not be exactly the same across all cognitive tasks, but rather will show some clustering. Finally, each of these broad ability areas can be characterized in terms of a number of more specific abilities (the narrow stratum) that are more homogeneous still than those at the next highest level.

The hierarchical structure of the domain of cognitive abilities has important implications for understanding three key aspects of cognitive ability tests: (1) the validity of these tests as predictors of job performance and effectiveness, (2) the relationships among abilities and the relative importance of general versus specific abilities for predicting performance, and (3) adverse impact. First, abundant evidence shows that cognitive ability is highly relevant in a wide range of jobs and settings and that measures of general cognitive ability represent perhaps the best predictors of performance (Schmidt & Hunter, 1998). The validity of measures of general cognitive ability has been established in all sorts of jobs and settings, and it is reasonable to believe that a good ability test will be a valid predictor of performance in virtually any application of testing.

The hierarchical structure of the cognitive domain is almost certainly a key to the widespread evidence of the validity of cognitive tests. All jobs require active information processing (such as retrieving and processing information, making judgments), and

Figure 1.1. The Cognitive Domain



even when the content of the job focuses on very specific tasks or types of ability (a job might require spatial visualization abilities, for example), the strong intercorrelations among abilities virtually guarantee that measures of general ability will predict performance. This intercorrelation among cognitive abilities also has important implications for evaluating the importance of general versus specific abilities.

A good deal of evidence exists that the incremental contribution of specific abilities (over and above general ability) to the prediction of performance or training outcomes is often minimal (Ree, Earles, & Teachout, 1994). Because of the correlation among measures of general and specific abilities, payoff for the specific abilities required in a job is usually small. Measures of general ability will usually be as good as, and often better than, measures of specific abilities as a predictor of performance and effectiveness.

The strong pattern of intercorrelation among cognitive abilities poses a strong challenge to the hypotheses that many types of intelligence exist (Gardner, 1999) or that important abilities have not yet been fully uncovered. In particular, the overwhelming evidence of positive correlations among virtually all abilities raises important questions about the nature of emotional intelligence.

Organizations have shown considerable interest in the concept of emotional intelligence (EI: Murphy, 2006). There are many different definitions and models of EI, but the claim that it is a distinct type of intelligence is at the heart of the debate over its meaning and value. On the whole, little evidence exists that emotional intelligence is related to other cognitive abilities, casting doubts on its status as an "intelligence." Some evidence suggests that EI is related to a variety of organizationally relevant criteria, but on the whole, the claim that EI is a distinct type of intelligence and an important predictor of performance and effectiveness does not hold up to close scrutiny (Murphy, 2006). More generally, the idea that there are distinct types of intelligence does not square with the evidence.

Finally, the hierarchical structure of the cognitive domain has

consistently receive lower scores on cognitive ability tests than white examinees, and the use of cognitive ability tests in personnel selection or placement will normally lead to adverse impact against black and Hispanic examinees (Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). Some differences in the amount of racial disparity are expected with measures of different specific abilities (in general, the stronger the correlation of specific abilities with *g*, the larger the racial disparities), but one consequence of the positive manifold among measures of various abilities is that adverse impact will be expected almost regardless of what specific abilities are measured. The hierarchical structure of the cognitive ability domain has several implications for research and practice in personnel assessment, including:

- General abilities have broad relevance in most settings.
- Identifying the right specific abilities is not necessarily important.
- The faults of general abilities will be shared with specific ones.
- The belief in multiple types of intelligence or newly discovered intelligences is not consistent with the data.

First, the hierarchical structure of cognitive abilities means that general abilities are more likely to be useful for predicting and understanding behavior in organizations than more narrowly defined specific abilities. This structure guarantees that even if it is the specific ability that is important, general abilities will also turn out to be good predictors in most settings. Because general abilities are usually measured with more reliability and more precision, making the case for focusing on specific rather than on general abilities is often hard.

Second, if the goal is predicting future performance and effectiveness, this structure suggests a diminishing payoff for getting it exactly right when drawing inferences about the abilities required by a job. For example, the spatial-perceptual branch of most hierarchical models of cognitive ability includes a number of specific abilities (say, three-dimensional spatial visualization

three-dimensional spatial visualization), the less difference choices among branches of the ability tree are likely to make in determining the eventual value and criterion-related validity of ability tests.

Third, the use of ability tests in making decisions about people in organizations such as personnel selection or placement will lead to adverse impact against members of a number of racial and ethnic groups, and the use of specific rather than general ability measures will rarely change this fundamentally. Specific ability measures do show slightly lower levels of adverse impact than general ones, but they also typically show lower levels of criterion-related validity. The decision to use cognitive ability tests in organizations is necessarily also a decision to accept a certain level of adverse impact; the decision to refrain from using such tests is almost always also a decision to sacrifice validity.

Finally, the long-standing assumption and hope of many researchers and practitioners (especially in educational settings) that we can identify many separate types of intelligence is exactly that: an assumption and an aspiration. Models that posit multiple intelligences or suggest that specific types of content such as emotional information require their own type of intelligence are popular but not well supported. In the case of emotional intelligence, which has attracted a great deal of attention in both educational and organizational settings, improvements in the models and measures of this construct may eventually lead to the acceptance of EI as a distinct and important domain of human cognitive ability, but there are few data on the immediate horizon to lead us to believe that current conclusions about the structure and nature of human cognitive ability will need to be radically changed to accommodate separate intelligences such as EI.

Personality

The link between personality and behavior in organizations has a long history of interest. In a highly influential review, Guion and Gottier (1965) cast considerable doubt on the value of personality measures, especially as predictors of job performance. They concluded that "there is no generalizable evidence that personality measures can be recommended as good or practical tools for

situations as a basis for making employment decisions about people" (p. 160). This review led to a long period of skepticism about the relevance of personality in understanding performance and effectiveness in the workplace. Not until the 1990s did personality reemerge as a viable tool for understanding and predicting performance and effectiveness (Barrick & Mount, 1991). It is now widely accepted that measures of normal personality have some value as predictors of performance, but the validities of these measures are often low. Nevertheless, they are also viewed as useful measures for helping to structure and manage development and placement programs.

As with cognitive ability, one of the keys to understanding the relevance and value of personality measures is to examine the structure and the contents of this domain. The Five Factor Model, often referred to as the "Big Five," has emerged as a dominant model for describing normal personality. This model has been replicated across a number of methods, settings, and cultures, and it provides a good starting point for describing what exactly *personality* means. This model suggests that normal personality can be described largely in terms of five broad factors that are at best weakly related to one another and (with the exception of Openness to Experience) with cognitive abilities:

- Neuroticism: emotional instability, tendency to experience negative emotions easily
- Extraversion: outgoing, energetic, tending toward positive emotions
- Agreeableness: cooperates with, is compassionate and considerate toward others
- Conscientiousness: reliability, self-discipline, achievement oriented, planfulness
- Openness to Experience: curiosity, imagination, appreciation for new ideas and experiences, appreciation of art, emotion, adventure

integrity tests capture aspects of Conscientiousness, Neuroticism, and Agreeableness (Ones, Viswesvaran, & Schmidt, 1993); the breadth of the domain these tests cover may help to explain their validity as a predictor of a fairly wide range of criteria. In principle, there might be no effective limit to the types of composite personality tests that might be created, and some of these might plausibly show very respectable levels of validity. However, this strategy almost certainly involves a trade-off between the potential for validity and interpretability.

The use of personality assessments to make high-stakes decisions about individuals is controversial (Morgeson et al., 2007), in large part because most personality inventories are self-reports that are potentially vulnerable to faking. The research literature examining faking in personality assessment is broad and complex (Ones, Viswesvaran, & Reiss, 1996), but there is consensus about a few key points. First, people can fake, in the sense that they can often identify test responses that will paint them in the most favorable light. Second, while faking can influence the outcomes of testing, it often does not greatly affect the validity of tests. This is because positive self-presentation biases are often in play when job applicants and incumbents respond to personality inventories, meaning that everyone's scores might be inflated. Although faking is a legitimate concern, it is probably more realistic to be worried about the possibility of differential faking. That is, if some people inflate their scores more than others, faking could change both the mean score and the rank order of respondents. In other words, if everyone fakes, it might not be a big problem, but if some people fake more or better than others, faking could seriously affect the decisions based on personality inventories.

As with cognitive ability, the structure and nature of the domain of normal personality have important implications for research and practice in organizational assessment:

- Personality measures provide information that is distinct from that provided by ability measures.
- The relatively low correlations with ability suggest that personality measures will be poor predictors of performance and effectiveness; the available evidence seems to confirm this prediction.
- Narrow dimensions of personality are easiest to interpret, but are often similarly narrow in terms of what they predict. The broadest dimensions show more predictive power but are hard to sensibly interpret.

First, the broad dimensions that characterize the Big Five are relatively distinct, which poses both opportunities and challenges. On the opportunity side, it is more likely that the complex models (for example, configural models, in which the meaning of a score on one dimension depends on a person's score on other dimensions) will pay off in the domain of personality than in the domain of cognitive ability. In the ability domain, the pervasive pattern of positive correlations among virtually all ability measures means it is hard to go too far wrong. Even if you fail to identify the exact set of abilities that is most important, you can be pretty certain of capturing relevant variance with measures of general abilities. In the personality domain, choices of which dimensions to assess and how to combine them are likely to matter. This also means that identifying the best way to use personality information is likely to be a much more difficult challenge than identifying the best way to use information about abilities.

Second, personality and ability seem to be largely independent domains. There are some broad personality dimensions that may be related to *g*, but most are not. This means that potential exists for personality measures to contribute to the prediction of performance and effectiveness above and beyond the contributions of ability measures. Unfortunately, as noted in our third point, this often does not happen. The validities of personality measures are

and another between ease of use and trustworthiness. Personality measures are usually self-reports, and they are not necessarily hard to develop. They are, however, vulnerable to faking. Ability tests have many defects, but at least it is hard to "fake smart." A personality inventory that shows an applicant to be high on Conscientiousness and Agreeableness might mean exactly what it appears to mean—or it might mean that the respondent knows that high scores on these dimensions are viewed favorably, and is faking.

Interests and Value Orientations

Organizational assessments are used not only to predict performance and efficiency but also to evaluate the fit between people and environments or jobs. Ability and personality measures can be very useful in assessing fit, but many discussions of fit focus on interests and value orientations, based on the argument that the congruence between the interests and the values of an individual and the environment in which he or she functions is an important determinant of long-term success and satisfaction. There are important questions about the extent to which fit can be adequately measured and about the importance of person-environment fit (Tinsley, 2000), but the idea of congruence between individuals and environments is widely accepted in areas such as career development and counseling. Numerous models have been used to describe the congruence between individuals and environments; Lofquist and Dawis's (1969) Theory of Work Adjustment represents the most comprehensive and influential model of fit. The theory examines the links between the worker's needs and values and the job's ability to satisfy those needs, and it also considers the match between the skills an individual brings to the job and the skills required for effective performance in that job.

Assessments of interests have long been an important part of matching individuals with jobs. Strong (1943) defined an interest as "a response of liking" (p. 6). It is a learned affective response to an object or activity. Things in which we are interested elicit positive feelings, things in which we have little interest elicit little

help individuals identify vocations and jobs that are likely to satisfy and engage them.

The dominant theory of vocational choice was developed by Holland (1973), who suggested that vocational interests can be broken down into six basic types: realistic (interest in things), investigative (interest in ideas), artistic (interest in creating), social (interest in people), enterprising (interest in getting ahead), and conventional (interest in order and predictability). The Holland RIASEC model is shown in Figure 1.2.

The hexagonal structure of Figure 1.2 reflects one of the key aspects of the Holland model. Interests that are close together on the Holland hexagon, such as Realistic and Investigative, are more likely to co-occur than interests that are far apart such as Realistic and Social. The great majority of measures of vocational interests and theories of vocational choice are based on the Holland model.

Unlike the field of interest measurement, there is no single dominant model of work-related values. Probably the best-researched

Figure 1.2. Holland Taxonomy of Vocational Interests

model is that proposed by Lofquist and Dawis (1969). Their taxonomy of work-related values, shown in Table 1.3, was adopted by O*NET as a way of characterizing the values most relevant to various occupations.

Like many other taxonomies, the O*NET Work Value Taxonomy is hierarchically structured. At the highest level of abstraction, jobs can be characterized in terms of the extent to which they are likely to satisfy value related to opportunities for achievement, favorable working conditions, opportunities for recognition, emphasis on relationships, support, and opportunities for independence. One of the many uses of O*NET is to match jobs to people's values. For example, individuals who value achievement and recognition can use O*NET to identify jobs that are likely to satisfy those preferences. The lower level of the taxonomy helps to clarify the meaning of each of the higher-order values and provides a basis

Table 1.3. O*NET Work Value Taxonomy

Achievement	Relationships
Ability utilization	Coworkers
Achievement	Social service
	Moral values
Working conditions	Support
Activity	Company policies and practices
Independence	Supervision, human relations
Variety	Supervision, technical
Compensation	
Security	
Working conditions	
Recognition	Independence
Advancement	Creativity

more detailed predictions of judgments. There are many models of person-job fit, and different models often depend on different sets of values. No single agreed-on taxonomy adequately captures the universe of organizationally relevant values. Nevertheless, the general proposition that some jobs are more likely than others to fit an individual's values and that some individuals are more likely than others to fit any specific job seems well established, and the measurement of work-related values has potential for both research and practice.

This chapter has been intentionally broad in its focus, and the implications for assessment laid out in the preceding paragraphs are similarly broad. Chapters Two through Eight examine more specific issues in assessments of domains ranging from abilities to personality to background and experience. Chapters Nine through Fourteen show how assessments of these domains are used in making decisions in occupations ranging from hourly or skilled work to executive and managerial positions. Chapters Fifteen through Twenty-Four discuss a wide range of questions encountered when developing and using assessments in a range of organizational contexts.

References

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

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

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

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment, 64*, 21–50.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st*

CHAPTER 2

INDICATORS OF QUALITY ASSESSMENT

Fritz Drasgow, Christopher D. Nye, Louis Tay

Assessment, whether for selection or development, can play a critical role in elevating an organization from mediocrity to excellence. However, this is true only if the assessment is excellent. In this chapter, we describe the characteristics and features that differentiate outstanding assessment programs from mediocre systems. With this information, organizational practitioners can thoughtfully consider how assessments can be implemented in their organizations, evaluate any current uses of tests and assessments, and move toward state-of-the-art measurement.

When an organization decides to begin an assessment program, its first decision concerns whether to purchase a test from a test publisher or consulting firm or develop the assessment tool in-house. We begin the chapter by reviewing the issues to consider when making this important decision. We next discuss the test construction process, which begins with the question, "What should the test measure?" and addresses item writing, pretesting, and psychometric analyses. The next two sections examine the critical quality issues of reliability and validity. Obviously organizations want their assessments to be reliable and valid, but there are some subtleties that test users should understand in order to make informed judgments; we summarize these issues.

We then discuss operational models for assessment programs. With advances in computer technology and the Internet, organizations have a dizzying array of choices. Some of the topics discussed

include testing platform (paper-and-pencil versus computer), unproctored Internet testing, cheating, and score reporting.

The next section of the chapter addresses quality control, a topic that receives little attention in many testing programs. There have been several highly publicized fiascos in high-profile testing programs in recent years and undoubtedly many other problems that were kept under wraps. Consequently, we discuss issues to consider and steps organizations can take to ensure high quality. Finally, we end with a few brief conclusions.

We expect that people with diverse backgrounds will read this chapter. We encourage those with psychometric training, including classical test theory (CTT) and item response theory (IRT), to dig into the technical details that are needed to fully address the quality of a testing program. To this end, the equations that are referenced throughout the chapter are in Table 2.1 for convenience. For those who do not have this background, we encourage looking at the big picture to gain an understanding of critical issues. We have attempted to give conceptual descriptions of each topic so that all readers can understand important problems; a flowchart of the key processes related to quality assessment is shown in Figure 2.1. Organizational leaders can then consult with either internal or external measurement professionals for guidance on technical concerns.

Buy Versus Build

If a decision is made to implement an assessment program, the organization must decide whether to purchase a commercially available test or develop a measure in-house. To make this decision, organizations need to weigh the costs and benefits of each approach to determine which will be more appropriate, and a number of questions must be addressed. First, do any currently available tests meet the needs of the organization? Specifically, do the commercially available tests validly measure the requisite knowledge, skills, and abilities (KSAs) and have rigorous empiri-

is not directly observable). However, many organizations do not have expertise in IRT. Additional difficulties may be encountered when using computerized or computer-adaptive tests (CATs). (See Drasgow, Luecht, & Bennett, 2006, for a description of the complexities of a technologically sophisticated testing program.)

Other decision factors are the time lines and the breadth of the assessment program. Organizations with an immediate need may benefit from purchasing a commercially available test because test development and validation can be time-consuming. An organization should also consider how frequently the test will be used. For example, if a brief unproctored Internet test is used as an initial screening for tens of thousands of job applicants, a commercially available test may become very expensive as costs accrue with each administration. In contrast, organizations that use assessments only intermittently may not recoup the cost of developing the test in-house.

Finally, test security should also be considered. Are cheating conspiracies likely? Test security is enhanced by using multiple forms of paper-and-pencil tests. Even more effective is the use of a CAT with a large item pool and an effective item exposure control algorithm. To illustrate the significance of this problem, note that there have been significant cheating conspiracies involving college entrance and licensure exams. Even multiple conventional forms and CATs can be susceptible to large-scale cheating conspiracies, such as online sharing sites and companies devoted to cracking the tests. One benefit of commercially developed assessments is that the developers are well positioned to ensure the security of the exam because their business success is severely affected by cheating conspiracies. Some professional test developers may even employ individuals with the sole responsibility of searching for and eliminating item-sharing sites and companies.

In sum, the buy-versus-build decision involves considerations of availability, feasibility, timeliness, in-house expertise, cost, and so forth. Clearly this is a critical and complex choice. Regardless of the buy-versus-build decision, a quality assessment must be cre-

perform the steps we describe. And before buying, the organization should examine documentation from the test publisher to ascertain whether the criteria we describe next are satisfied.

Test Construction Considerations

Several steps in the development process have a critical impact on the quality of an assessment. Integrating these steps provides a systematic approach to test development and ensures a high-quality result. A less-systematic approach may produce a test that misses important aspects of the KSAs to be assessed, which is likely to reduce the effectiveness of the assessment.

The first step in test development lies in identifying what a test is intended to measure. Here, test developers establish the content that will be assessed. In an employment setting, this is most frequently done with a thorough job analysis. Test developers may survey or interview subject matter experts, examine critical incidents, or rely on expert judgment. Because it is usually impossible to assess all important KSAs for a particular job or job family, the criteria for including content should be based on information provided by the job analysis regarding the importance of each dimension. For psychological phenomena such as intelligence, personality, and attitudes, inclusion criteria should also be based on a careful definition of the trait to be assessed, followed by a thorough review of the literature on the topic.

The second step is to determine the testing format that is most appropriate for the purposes of the test. With the large number of administration formats now available for psychological testing, this issue is fundamental to the test construction process. In addition to the traditional paper-and-pencil format, a conventional test (one in which all examinees are administered the same set of items) may also be administered by stand-alone computers or

unproctored Internet testing allows unsupervised examinees to take the exam at a time and place of their convenience.

Each of these testing formats has implications for the type and number of items used. In a computerized format, novel item stimuli may be presented interactively as audio, video, pictures, or some combination of media. For CATs to operate effectively, a large pool of items is required to ensure accurate ability estimates and increase test security. Similar security issues are salient for unproctored tests. As a result, it is often advisable to administer both an unproctored selection test and, later, a proctored confirmation test to verify results. Here, the proctored confirmation test may be a parallel form of the unproctored exam.

The choice between administration methods may also affect the third step in test construction where test specifications are formulated. These guidelines should be used as a road map for item writers. For example, test specifications would detail the number of items assessing verbal, quantitative, and spatial abilities in a measure of cognitive ability. In addition, these plans may specify the item difficulty and discrimination levels required for accurate ability estimates. These criteria are particularly important for CATs, where the quality of ability estimates improves when the item pool contains items with a wide range of difficulties.

The test specifications give the appropriate number and content of items as well as the format for the test. The number of items should be chosen based on considerations for reliability, content coverage, and test security. However, workplace assessments must effectively balance content sampling with space and time limitations. Assessments with too few items may not adequately measure the entire domain of the trait (content validity) or provide consistent results (reliability). And assessments with too many items may result in test-taker fatigue or negative reactions and may not be appropriate for situations with strict time constraints.

The choice of the item format may mitigate some of the disadvantages traditionally associated with measurement. For example, forced-choice response formats, where respondents must choose between two or more items matched on social desirability, may reduce the prevalence of faking on personality items. Other novel

information considered when making a high-stakes decision, reliability should be at least .80. A measure that does not reach an adequate level of reliability should be revised.

Traditional Forms of Reliability

In this section we review traditional measures of reliability and their limitations. These reliability indexes all range from 0 to 1, with 1 indicating perfect reliability.

Test-retest reliability is estimated by administering a test or scale to a sample at two points in time and then correlating the scores. It is an important index for characteristics that should be stable across time. For example, intelligence is a highly stable trait, and consequently a minimal requirement for an intelligence test is to have substantial test-retest reliability.

Internal consistency reliability includes split-half reliability, the Kuder-Richardson KR20 and KR21 reliabilities, and Cronbach's coefficient alpha. All of these measures are functions of the intercorrelations of the items constituting a test. Thus, for a fixed test length, internal consistency reliability is higher when the test's items are more strongly correlated.

Reliability coefficients can be manipulated and artificially inflated. Therefore, it is important to consider several factors when interpreting a reliability coefficient, including test content, interitem correlations, test length, and the sample used to estimate reliability.

By incorporating highly redundant items, it is possible to manipulate reliability (and particularly internal consistency reliability) to produce substantially inflated values. Therefore, before giving credibility to a measure of reliability, it is important to examine the content of the measure for substantive richness and breadth. A narrow and excessively redundant measure may have an internal consistency reliability in excess of .95 but nonetheless be lacking in regard to other important properties, such as construct validity, which would reduce its correlation with job performance

redundancy. For example, when assessing conscientiousness, two items might be, "I am careful in my work" and "I am meticulous in my work." Or in assessing math ability, two items could be restatements of the same problem but employ different numbers. Because variants of the same item will be answered by applicants in similar ways, such redundant items should be excluded because they ostensibly increase reliability but do not truly add new information.

Classical test theory shows that reliability can be increased by adding more items. Some high-stakes licensing exams, for example, consist of several hundred items. If a test has a long form and a short form, the reliability of the long form should be larger than the reliability of the short form, and it is important not to confuse the two. Unfortunately, high reliabilities of long tests are sometimes mistaken as indicating unidimensionality.

Reliability also depends on the characteristics of the sample. Range restriction, which occurs when the selection process has resulted in a sample that displays a truncated range of test scores, lowers inter-item correlations and results in lower reliability. Conversely, an artificially broad sample for example, using a sample of third-, fourth-, and fifth-grade students to estimate the reliability of a math achievement test designed for fourth graders, will inflate reliability. Because estimates of test reliability are sample dependent, it is important to ask whether the sample that was used to estimate test reliability is similar to the sample used for a specific organizational assessment purpose. If it is not, then the reliability estimate will be less informative for the organization.

Perhaps the greatest limitation on test-retest reliability results from the fact that reliability is sample dependent. Test-retest reliability is often estimated in a small, experimental study because it is difficult to administer the same test twice to a random sample under operational conditions. Thus, the question arises of whether results from the sample in the small research study can be generalized to other groups of test takers. Answering this question can be difficult or impossible.

Because reliability is subgroup dependent, it is inappropriate to say, "The reliability of test X is .92." Instead, a statement about reliability should include information about the group for which it was computed.

Additional concerns can be seen by looking at the technical definition of *reliability* as defined with classical test theory (the squared correlation between true scores, that is, the hypothetical scores people would receive if assessed with a perfect test, and observed scores). For example, the traditional reliability index is uninformative as to test precision at different score levels; one value of reliability is given for the test. Similarly, the standard error of measurement (the standard deviation of observed scores around the examinee's true score) of a test score X is given by equation 2 in Table 2.1, where $\hat{\sigma}_x$ is the standard deviation of test scores, and r_{xx} is the test's reliability; no differentiation is made between high, low, or moderate values of X. In many situations, it is critical to determine the test's precision at important cut scores where high-stakes decisions are made (AERA/APA/NCME, 1999).

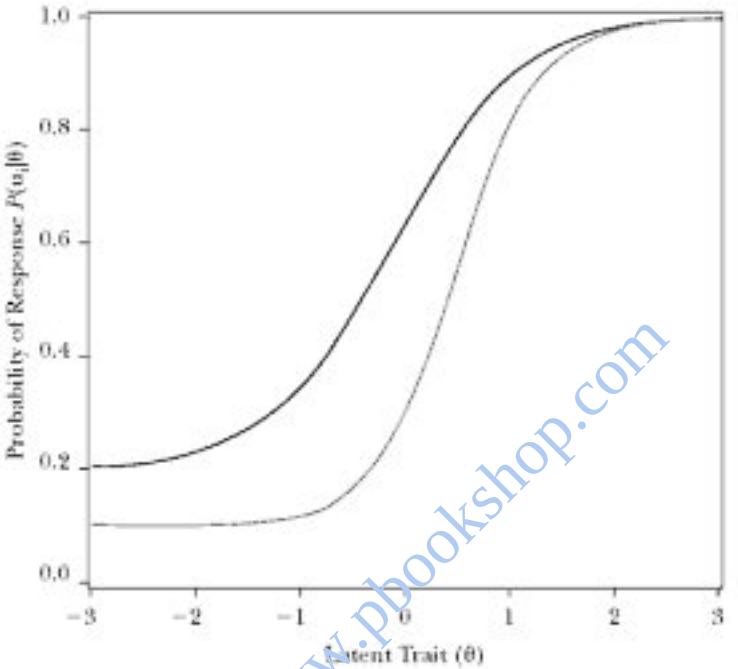
Although CTT provides only a single standard error of measurement for a test, the standard error in IRT is conditional on the level of the latent trait. Thus, we denote the conditional standard error of measurement at a given true score τ by $SE(X|\tau)$ and the conditional standard error at a given latent trait score θ by $SE(X|\theta)$. These values, computed using IRT, allow test users to understand the magnitude of measurement error at critical score ranges.

Modern Forms of Reliability

A modern perspective on reliability is grounded in IRT, so the details are more complicated.

If number-right scoring is used on a test, the total test score is determined by counting the number of items answered correctly. Mathematically, the number-right score can be defined by equation 3 in Table 2.1, where X is the total score on the n item test and the score on item i is coded $u_i = 1$ if correct and 0 if incorrect. It can be shown that there is a one-to-one correspondence between the true score τ of classical test theory and the θ of IRT when the assumptions of IRT hold (see equation 4 in Table 2.1). Using θ_τ to indicate the value of θ corresponding to a particular true score τ, the conditional standard error of measurement is given in equation 5 in Table 2.1.

Figure 2.6. Hypothetical IRFs for Men and Women



Note: Black and grey lines represent hypothetical IRFs for men and women respectively.

of responding correctly. Measurement equivalence occurs when individuals with the same standing on the latent trait have equal chances of responding correctly, regardless of their group. Thus, establishing measurement equivalence is an important prerequisite to producing accurate scores across subgroups.

Conclusion

We have outlined many exemplary testing practices as well as numerous pitfalls. Of course, no testing program should be evaluated with a checklist mentality, where strong features and weaknesses are simply added up. But as the 1985 *Test Standards*

Tests should be reliable for their intended use. By this we do not mean that coefficient alpha should be large for some unspecified sample. Instead, tests should consistently classify test takers into meaningful groupings such as "recommended for hiring" versus "not recommended for hiring," "high potential" versus "not high potential" for development, "recommended for a training program" versus "not recommended for training," and so forth. A small conditional standard error of measurement at the cut score for the categories is needed for consistent classification. More generally, small conditional standard errors of measurement are needed at all important cut scores for a test to validly fulfill its intended function.

Modern views describe validity as a unitary construct. We have described how the test content and its internal structure, convergent and discriminant validity, and test-criterion relationships are all important facets of validity. We urge test users and judges of test use to look beyond the statistical significance and magnitude of a test-criterion correlation when evaluating validity. The U.S. Army example in Figure 2.5 clearly demonstrates that focusing on a small correlation can obscure an important and powerful relationship.

A reliable and valid test constitutes one of the most extraordinary returns on investment available to organizations. For example, a thirty-minute test of sales skills, purchased for perhaps $10 per test taker, may increase revenue per salesperson by $100,000 per year. In this era of growing domestic and international competition, organizations cannot forgo this competitive advantage.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.

most useful for summarizing the value of general cognitive ability tests in the prediction of training performance.

The first type of particularly valuable studies are validity generalization studies that have examined the value of general cognitive ability in the prediction of training performance in specific job families. For example, Barrett, Polomsky, and McDaniel (1999) reported a validity of .77 for general cognitive ability tests and firefighter training performance. Pearlman, Schmidt, and Hunter (1980) reported a validity of .71 for general cognitive ability tests and clerical training performance. Hirsh, Northrop, and Schmidt (1986) reported a validity of .71 for a general cognitive ability composite for police and detective training performance. There are many additional studies, including ones that that have examined semiprofessional occupations (Trattner, 1985) and blue-collar jobs where apprenticeship training was common (Northrop, 1988). Lilienthal and Pearlman (1983) reported general cognitive ability validities for training performance in entry-level aid and technician occupations in the health, science, and engineering fields. Hunter and Hunter (1984) reported training validities for a large number of jobs grouped by job complexity (the cognitive demands of the jobs). These validities varied from .50 to .65.

The second type of valuable studies are primary validity studies that use military data. Such studies are of particular value because they use well-developed tests of general cognitive ability, everyone who enters the military receives some job training, and the sample sizes are huge. Olea and Ree (1994) reported validities for general cognitive ability tests and pilot and navigator training in the U.S. Air Force. General cognitive ability predicted training performance (validity = .31) for the pilots and .46 for the navigators. Earles and Ree (1992) examined the validity of general cognitive ability in 150 military training programs using data from 88,724 U.S. Air Force recruits. The general cognitive ability test showed impressive validities across all 150 training programs.

those at the lowest levels of general cognitive ability are not job applicants. Second, jobs often have educational or experience requirements. Job applicants who meet the educational and experience requirements will be more equal in general cognitive ability than a random sample of the population. Thus, one would expect large mean ethnic differences for jobs with low educational and experience requirements and smaller mean ethnic differences for jobs with extensive educational and experience requirements. Thus, for example, mean ethnic differences should be smaller for those whose highest educational credential was a four-year college degree than for those who did not graduate from high school. Roth, BeVier, Bobko, Switzer, and Tyler (2001) provided the best estimates of mean ethnic differences in general mental ability among applicant groups. The mean difference is expressed as standardized mean differences (d) in a standard deviation metric. Thus, if $d = 1$, the white mean on general cognitive ability is one standard deviation higher than the black mean. Based on 125,654 applicants, the mean d for low-complexity jobs was .86. For medium-complexity jobs (31,990 applicants), the mean d was .72. Based on 4,884 applicants, the mean d for high-complexity jobs was .63. This pattern of larger mean differences for lower-complexity jobs is consistent with the expected mean differences varying as a function of educational and experience requirements. At all complexity levels, these differences are large and will cause disparate hiring rates by race if hiring is solely based on scores on general cognitive ability tests in a race-blind manner. Adding other predictors to a cognitive ability test battery may reduce the disparate hiring rates.

Single-Group Validity, Differential Validity, and Differential Prediction

The large mean differences between whites and blacks in general cognitive ability raise the possibility that general cognitive ability tests are biased against blacks. The hypothesis of single-group validity holds that a general cognitive ability test has validity for one group (for example, whites) but zero valid-

In principle, then, personality measures can be used to forecast performance in all jobs in every organization. In practice, however, entry-level employees are typically screened using short and inexpensive measures of integrity/reliability, resilience/stress tolerance, and service orientation. Applicants for more senior positions in sales and management are more often screened with more extensive batteries.

How Are These Measures Developed?

The best-known personality inventories (CPI, 16PF, HPI) were developed following accepted professional standards and guidelines. But the scoring systems for interview questions, observer rating forms (including 360-degree appraisals), and the various projective tests are typically ad hoc, and the measures are often used with little concern for reliability and validity.

For What Purposes Are These Measures Used?

In the workplace, personality assessment is used for three purposes. First, assessment is frequently used to provide individuals with feedback that is intended to increase their self-awareness and improve their ability to work as part of a team. The Myers-Briggs Type Indicator in particular is widely used for these purposes; it tells people how they tend to perceive the world, as compared with how others see the world, and how these perceptual differences can lead to unintended misunderstandings and conflict. The so-called 360-feedback process is also used to enhance self-awareness. A target person is described by subordinates, peers, and superiors using a standardized rating form. The results are summarized and reported back to the target person, a process that is as often disheartening as it is enlightening—because most people think more highly of themselves than others actually do.

Second, and related to the preceding point, assessment results

can be evaluated in terms of potential strengths and developmental needs. Management and leadership development programs rely heavily on assessment for career guidance. The relevant profile for effective performance in managerial or leadership roles is reasonably well understood; in addition, a vast armada of leadership assessment material is commercially available. Leadership development is a $50 billion a year industry in the United States alone, and it is increasing with worldwide demand.

Finally, personality assessment is used for personnel selection. As noted above, we can profile the psychological requirements of every job in the U.S. economy. We can then match the assessment profile of an applicant to the required profile of any job and determine the likelihood of the person succeeding in the job. Employment interviews are an informal way of doing personality-based personnel selection; assessment centers and standardized psychometric batteries are a more formal way of doing it.

How Well Does Personality Assessment Work?

The question of how well personality assessment works is, of course, the bottom-line issue for this chapter. A naive reader might think this is a straightforward empirical question, but in academic life, few important questions seem capable of being answered in a straightforward manner. Especially on this question, opinion is bitterly divided.

Framing the Question

Some writers (like the authors) are enthusiastic advocates of personality assessment. Others doubt that personality assessment works and vigorously criticize its use for selection purposes (Morgeson et al., 2007). As always, conceptual confusions cloud these discussions. Persons new to this field need three pieces of background information in order to understand the debate.

The Measurement Problem

First, the accepted method for determining the validity of psycho-

(Hunter & Schmidt, 2004). The correlation between test scores and performance criteria in any single research study is regarded as one data point, and that data point is assumed to be contaminated by a variety of statistical artifacts. The process of meta-analysis collects as many studies as are available on the relationship between a type of test (or measurement dimension) and performance criteria. The results of all the studies are combined and then corrected for statistical artifacts, such as the degree of measurement error and range restriction in the test and in the outcome measure. These "corrected" results are assumed to provide the best possible estimate of the "true" relationship between a type of test (or measurement dimension) and a class of outcomes (or criteria). To summarize, meta-analysis is used to answer the question of how well a test works in predicting occupational performance.

Second, meta-analysis is well suited for estimating the validity of measures of cognitive ability, popularly known as intelligence quotient (IQ). The reason meta-analysis works well for measures of intelligence is that the various measures are all highly intercorrelated, which means that collectively, they represent the same large general factor. Because measures of intelligence are so statistically similar, one can compare the results from different studies with little regard for the particular measure that the researchers use.

In the case of personality, however, it is essentially impossible to combine measures across studies. The number of scales or dimensions on the best-known personality inventories varies widely, from three on the Eysenck Personality Inventory to twenty-one on the California Psychological Inventory. Furthermore, scales with the same names on different inventories (for example, Agreeableness) are not highly correlated—unlike the different measures of intelligence—and they predict different outcomes differently. This fact makes it virtually impossible to combine studies across different personality inventories; consequently,

The third problem concerns how to define job performance—more specifically, how to assign numbers to individual differences in performance—in a way that is comparable across organizations. Job performance is typically defined in one of three ways. In the first case, it is defined in terms of performance in training. When performance is defined this way, intelligence always outperforms personality in terms of validity because intelligence is the best single predictor of training performance. In the second case, performance is defined in terms of supervisors' ratings of overall job performance. In our view, this is problematic because overall ratings of performance tend to be political judgments; they reflect how much a supervisor likes an employee rather than how well he or she is performing. In the third and ideal case, performance is defined in terms of the relevant components of the job (for example, showing effort, maintaining personal discipline, facilitating team performance), which can be identified statistically (Campbell, McCloy, Oppler, & Sager, 1993). Once the relevant dimensions of performance have been defined and measured, the predictors should be aligned with these dimensions in ways that make conceptual sense. For example, one would not use a measure of intelligence to predict ethical behavior because many smart people have problems with integrity (financial managers, for example). However, a measure of Conscientiousness would be expected to predict ethical behavior because following rules is a key element of Conscientiousness.

The Comparative Validity of Personality

When the performance dimensions have been defined explicitly and the predictors have been aligned with the performance dimensions, it becomes possible to evaluate how well personality assessment works. As always, these judgments are relative. Consider Table 4.2, which presents meta-analytic estimates for the uncorrected validity coefficients for seven commonly used predictors of occupational performance. These validity coefficients vary between .11 and .28, and these values should be con-

- Whether self-evaluative or factual information is gathered.
- The amount of structure provided in the response format. For example, "Describe a time when . . ." allows an open-ended response, whereas, "How often have you performed a specific task?" is typically followed by multiple-choice response options.
- Whether the response options are continuous—for example, "most" to "least"—or categorical—for example, terminated from employment, retired, employed, or volunteer.
- The source of the information. Questions might be asked of the individual (self-report) or of some other person (other report). Or the data might be archival.
- How the information is scored. Scoring procedures might be empirically developed, resulting in algorithms that generate scores or rationally developed by experts.
- The response mode. Responses might be provided in written form or orally in an interview.

Validity of Biodata Measures

The ways in which the information is gathered and scored and the purposes for which the scores are used are evaluated by amassing validity evidence—criterion-related, construct, and content—as well as other information, such as applicant reaction, cost, and mean score differences between groups. The fundamental validity issue is the degree to which the interpretations of scores are appropriate for their intended or actual use. This chapter examines validity and research evidence for biodata measures with a focus on what we can learn from the research and how we can put that knowledge to use. The research findings provided the foundation for a set of questions that appear at the end of the chapter. The questions are intended to be used when evaluating biodata measures that assess an individual's past behavior and experience.

The evidence is clear: biodata predict valued organizational criteria as well as other criteria, such as life expectancy, divorce, and illness. This conclusion is true across virtually all kinds of biodata measures, scale development methods, criteria, crite-

are found in Bliesener, 1996; Huffcutt & Arthur, 1994; Hunter & Hunter, 1984; McDaniel, Schmidt, & Hunter, 1988a, 1988b; Ng, Eby, Sorensen, & Feldman, 2005; Quiñones, Ford, & Teachout, 1995; Schmitt, Gooding, Noe, & Kirsch, 1984). The expected criterion-related validity (observed validity) of a well-developed biodata measure is in the mid-.20s to low .30s. When compared with other predictor measures, the criterion-related validity of biodata measures ranks among the best, rivaling the validity of general cognitive ability measures, considered by many to be the best predictor of overall job performance. The cumulative evidence is that biodata predicts virtually all criteria that I/O psychologists have examined.

Of course, aggregating correlation coefficients across so many different studies, types of biodata measures, predictor and criterion variables, and scale construction methods ignores all the variables that are known or thought to moderate the validity of biodata measures. For example, research informs us that:

- Validity study design affects validity. Concurrent validity studies result in higher criterion-related validities than predictive validity designs.
- Population studied affects validity. Studies involving job incumbents result in higher criterion-related validities than studies involving job applicants.
- Type of criterion affects validity. Criterion-related validities are higher for training criteria than job performance criteria.

Nonetheless, biodata measures predict valued organizational criteria even when summaries report results from studies that included research designs, populations, and so forth known to result in lower criterion-related validities. Importantly, research also indicates that certain scale development features and item characteristics can enhance the validity of biodata measures.

to be predicted. If, for example, the purpose is to measure leadership, an analysis of what is meant by leadership is needed. Is it charismatic leadership, thought leadership, ethical leadership, and so forth? What precisely is to be measured? A good definition of the characteristic is required. It is a first step in developing a measure that has construct validity. If, however, the purpose is to predict turnover, an analysis of the causes or reasons for turnover in the particular setting should be examined. If the purpose is to predict job performance, an analysis of the job (work) and the context within which the work is performed is required. An analysis of the criterion is a first step in developing a measure that has criterion-related validity. If both construct and criterion-related validity are desired, the individual difference characteristic and the criterion, such as job performance, need to be carefully defined. An analysis of the characteristic to be measured and criterion to be predicted informs the test developer (or test user) of the content to be included in the item pool for the measure (scale)—a critically important initial step. Good theories, good job and work analyses, good definitions, and good thinking are needed. The sections that follow provide examples of item generation strategies that are likely to generate valid biodata measures.

Functional Job Analysis Approach

This method of job analysis produced the occupational classification system in later editions of the *Dictionary of Occupational Titles* (DOT; U.S. Department of Labor). It results in task-, attribute-, and behavior-based information very helpful for generating job-relevant biodata items (and scales). The task-based component includes work functions (what gets done) and working conditions. The attribute-based component focuses on individual difference characteristics needed to perform the job effectively, and the behavior-based component focuses on what workers do with things, data, and people. An important part of the functional job analysis approach is the information generated about levels of skill needed in each of three types of skills:

change related to physical, interpersonal, organizational, and working conditions
- *Functional skills*—mental, interpersonal, and physical capacities related to data, people, and things
- *Specific content skills*—competencies that enable a person to perform the specific job tasks, including environmental and work conditions that relate to procedures, standard operating procedures, machines, and equipment

This approach brings into focus the context of the task or work and the adaptive skills that include working conditions, effort, and responsibility. (For more information about this method, see Fine & Cronshaw, 1994.)

Each task statement in a functional job analysis contains information about the behavior, knowledge, skills, and abilities needed for effective job performance. That information, in conjunction with the context of the job, results in biodata items such as these:

- When given a work assignment, you prefer to have:
  a. discretion to do the work on your own
  b. direction from a supervisor or others
- As a student, you preferred homework assignments that:
  a. were detailed and explicit
  b. allowed you to define what to do and how to do it

Functional task statements can lead directly to biodata items such as, "How many times have you repaired the starter on an automobile?" or "How many years have you worked in a job in which you repaired the starter on an automobile?" that yield observed criterion-related validities in the .20s and sometimes higher (meta-analysis; Quiñones et al., 1995). Such biodata items appear to be very good for predictors of task performance but less good for predictors of organizational citizenship behav-

The effectiveness of the algorithm can be evaluated by examining the relationship of the computer-generated scores with expert-generated scores, as well as the relationship of the computer-generated scores with valued external criteria such as job performance. At this time, development costs and the validity of computerized scoring of text are still hurdles, although they are becoming less so. Meta-analysis of the relationship between scores on the Accomplishment Record (sometimes called the behavioral consistency method) and job performance indicates observed criterion-related validities are in the mid-.20s (McDaniel et al., 1988a).

The behavior-based interview method is similar to the Accomplishment Record method. The important difference is, of course, that a person asks the question and the respondent typically responds orally rather than in written form. Multiple meta-analyses indicate that the more structured the interview questions and rating process are, the higher the criterion-related validity is. Given that the interview questions are job relevant and the questions and rating process are structured, observed criterion-related validities can be expected to be in the .20s and even higher. As with the Accomplishment Record, the process involves human judgment. The interview, however, takes significantly more human time than the Accomplishment Record.

Comparison of Inductive, Deductive, and External Strategies

The purpose and circumstances of the assessment are important when comparing and evaluating scale construction methods. Depending on the purpose and circumstances of the assessment, one method may be better than the others. If, for example, construct validity is important, the external (empirical) scale construction method is not a good choice. The internal or rational scale construction methods are more appropriate. Frequently the purpose of measurement is to predict some future outcome. Thus, the expected level of accuracy with which each of the three scale construction methods predicts valued outcomes is of vital

and individuals involved. Unlike the inductive or deductive (rational) methods, which seek to understand and measure a construct, a presumed strength of the external method is its focus on predicting an external criterion of interest such as job performance, training performance, turnover, or satisfaction.

Hough and Paullin (1994) gathered criterion-related validities for each of the three scale development methods and compared the results. Their conclusion was that the three scale construction strategies produce reasonably similar levels of criterion-related validity for predicting external criteria, although the internal construction technique fared somewhat less well than the others. The criterion-related validities of the rational method were, on average, slightly higher (.01) than the cross-validities of the empirical method and about .06 higher than the validities of the internal method.

Yet even when the purpose is selection and criterion-related validity is of paramount importance, construct validity is also important. An important goal of science is to understand and predict phenomena. Meta-analyses of validities of construct-oriented scales for predicting specific criteria are more likely to lead to useful generalizations than are meta-analyses of heterogeneous scales that lack a coherent theme. Construct-oriented scales are building blocks in the development of scientific knowledge. Although both science and practice benefit from the development and use of construct-oriented scales, no one scale construction method is inherently better than the others.

All three approaches have advantages and disadvantages and are more or less appropriate depending on the purpose of the assessment. There is no clearly right or wrong approach, and scale developers often blend features of the three strategies, producing hybrid scale construction methods.

New Developments in Scale Construction Methods

A hybrid approach that appears to solve some of the problems