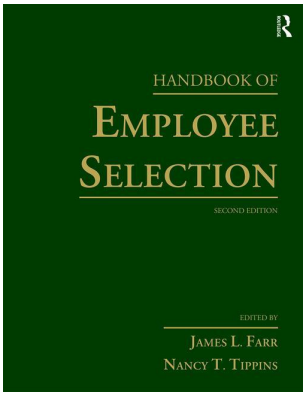

Handbook of Employee Selection

James L. Farr, Nancy T. Tippins, Walter C. Borman, David Chan, Michael D. Coovert, Rick Jacobs, P. Richard Jeanneret, Jerard F. Kehoe, Filip Lievens, S. Morton McPhail, Kevin R. Murphy, Robert E. Ployhart, Elaine D. Pulakos, Douglas H. Reynolds, Ann Marie Ryan, Neal Schmitt, Benjamin Schneider

Decisions in the Operational Use of Employee Selection Procedures

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-16
Nancy T. Tippins, Emily C. Solberg, Neha Singla
Published online on: 22 Mar 2017

16

Decisions in the Operational Use of Employee Selection Procedures

Choosing, Evaluating, and Administering Assessment Tools

NANCY T. TIPPINS, EMILY C. SOLBERG, AND NEHA SINGLA

INTRODUCTION

Most organizations have strategic goals that determine the kind of selection program they need. An organization that achieves its competitive advantage by delivering goods to consumers with a high level of service may focus on how its selection program can identify the best service workers. In contrast, another organization that pursues a goal of low-margin, high-volume sales of goods may be less concerned about a high level of service skills among employees and instead value a low-cost program for identifying employees with minimal skills to do the job efficiently. These overall objectives for the selection program, in turn, determine the goals for a specific assessment tool or test. An organization whose selection program goals relate to high levels of job performance is likely to set goals for the tests it uses related to validity and reliability and may attend less to their costs. There can be many selection program goals. Some examples include enhancing employee productivity, minimizing error, reducing accidents, complying with Equal Employment Opportunity (EEO) regulations, minimizing staffing costs, and supporting the employment brand. Once the organization's strategic goals are understood and the selection program goals defined, the organization can begin the process of determining the characteristics of the assessment tools constituting a selection program that meets those goals.

Just as there are many goals for a selection program, there can also be a number of test goals that typically fall into two broad classes: test characteristics and administrative goals. Test characteristics that often influence the choice of instruments include the validity and reliability that are typically found for a type of test used in the applicant population of interest and the estimated utility. Some organizations consider the appropriateness and feasibility of different validation strategies for a test in the context of their organization. Many organizations pay a great deal of attention to the potential group differences in test scores and adverse impact, and they search for alternative selection procedures that might have equal or greater validity and less adverse impact than another instrument. Some carefully review the past history of the test or type of test in litigation and attempt to avoid assessment tools that will be difficult to defend. Often, applicant reactions are also an important concern.

Administrative goals relate to issues around test administration and scoring. Organizations must consider how to measure the important job relevant constructs given their staffing

environments, financial and personnel resources, and time constraints. For many organizations, significant concerns about costs, including the costs of personnel with the necessary skills to administer and score the test in every location necessary, the costs of equipment needed for test administration, scoring, and data storage, the costs of test purchase, the costs of test development and maintenance, and the costs of validation, arise. In some environments, the availability of personnel and equipment cannot be assumed. In other environments, time concerns are very important. Organizations must consider the time requirements for development and validation as well as the amount of time necessary for test administration and the length of time between test completion and the availability of results.

A few test goals blend the two categories. The feasibility of the use of a test with particular characteristics within a specific staffing context is particularly important for some employers. For example, some organizations want to use tests that can be administered in an unproctored environment. Other employers, particularly those in public services such as police officers and firefighters, use tests for only one round of hiring to lessen the amount of information sharing that occurs over multiple administrations. In addition, organizations must decide how many of the critical knowledge, skills, abilities, and other personal characteristics (KSAOs) should be measured to adequately cover the domain of job performance and assure an acceptable level of accuracy in prediction, as well as how many constructs it can afford to measure. Other test goals that blend the two categories include the question of how test scores will be used. Typically, the employer must decide what form of test score to use (e.g., raw scores, scaled scores, percentiles), the combination and weighting of test scores, and the type of guidance to provide to hiring managers (e.g., cutoff scores, bands and expectancy tables, if used).

Sometimes, these test goals conflict. Choices made with the objective of identifying the best applicants may not be the same as those made to minimize costs. Furthermore, different constituencies in the same organization may have different goals. While the department receiving new employees may want tests that result in the best prediction of future performance, the staffing organization may want to minimize administration costs, and the legal team may want to avoid challenges to the selection process. Occasionally, the source of budget may determine which group's point of view prevails. For example, if the department needing the workers is funding test development and validation and the Human Resource department bears the cost of administration, costs for development may have different limitations than ongoing administrative costs. Optimization of all goals is often not possible, and organizations must usually make some trade-offs. For example, an employer that wants the most accurate predictor using the lowest cost instrument that takes the least amount of time is unlikely to achieve all three goals and will need to find the right compromises for the organization.

Many decisions about selection programs have ramifications for other decisions, and it is important to note that none of the decisions to be made should be considered independently from the others. For example, a requirement to have one, very short test to minimize the amount of time test takers will spend on a test will have an effect on the validity and reliability of the selection program. Or, a desire for positive applicant reactions could preclude a lengthy testing process composed of abstract measures of problem solving and suggest a choice of instruments that are more face valid. In such cases, the hiring organization must decide which goal is more important because both cannot be maximized. Often, decisions that are already made will need to be revisited as new decisions are made and additional criteria are considered.

In addition, there is no defined sequence to the decision-making process. Each employer tackles the problem of determining what test to use in its own way. While most testing experts would recommend determining the organization's strategic goals first, then the selection program's goals, and then the goals for specific tests as the most efficient process, many experienced professionals have reversed the order and deduced the strategic goals and selection program goals through discussion of the test goals. Research is needed on how goal choices are made and what goal hierarchies exist in relation to selection programs. Figure 16.1 displays some examples of how organizational goals, selection program goals, and test goals can be related to each other.

The remainder of the chapter reviews basic decisions about employment tests and discusses the considerations that influence those decisions. This chapter reviews five sets of decisions that must be made when developing and implementing a selection system: (1) What constructs

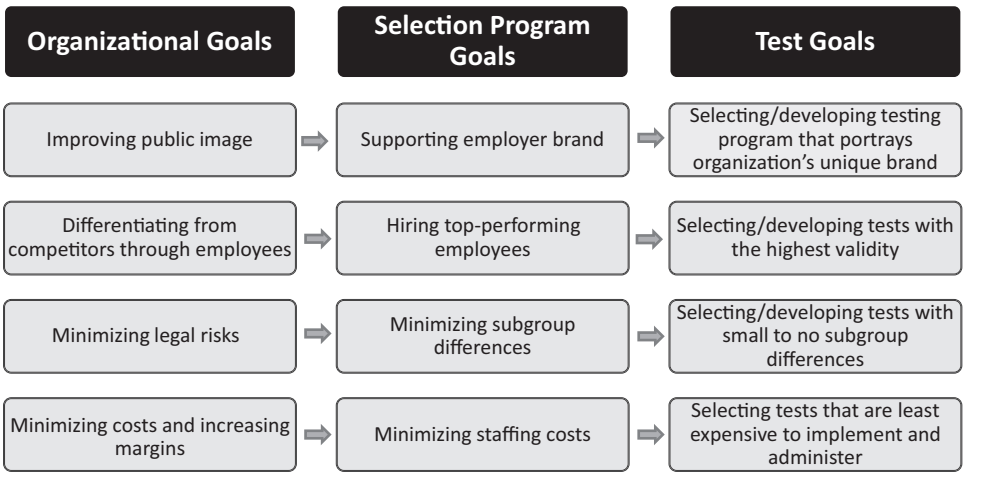

FIGURE 16.1 Organizational Goals, Selection Program Goals, and Test Goals

should be measured in the selection system? (2) How should the chosen constructs be measured? (3) How should the validity of assessments be evaluated? (4) How should the test be administered? (5) How should the resulting scores be used?

WHAT CONSTRUCTS SHOULD BE MEASURED?

One of the initial decisions to be made when developing and selecting assessment tools concerns the constructs that should be measured in the selection program and by individual tests constituting the selection program (see Chapters 11–15 in this volume for more information on the measurement of specific constructs). The test user must determine both which constructs to measure and how many to measure. Additionally, the test user must consider the advantages and disadvantages of measuring a single KSAO or broader subset of the entire content domain.

Importance and Needed at Entry

According to legal and professional guidelines, such as the *Uniform Guidelines on Employee Selection Procedures* (Uniform Guidelines; Equal Employment Opportunity Commission, 1978), the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), and the *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP *Principles*; SIOP, 2003), tests should measure KSAOs that are job relevant and necessary at entry to the job. Often, job relevancy is operationalized as those KSAOs that are identified as important and required to perform important tasks by subject matter experts who are typically also asked to rate the extent to which a KSAO is needed at entry into the job.

In choosing or developing appropriate tests, the test user should generally avoid a test that measures an appropriate construct but requires one or more nonrelevant KSAOs to complete it. If a measure of manual dexterity requires the applicant to read detailed instructions and reading at the level of the test instructions is not a job requirement, then the manual dexterity test is not likely to be an accurate predictor of performance in the job, because such a test will confound the candidate's manual dexterity with his/her reading proficiency. Steps must be taken to communicate the test directions in another form if reading is not also required for the job.

When the applicant has a disability and meets the requirements for protection under disability laws in the U.S. (e.g., the Americans with Disabilities Act [ADA] and the ADA Amendments Act

of 2008), the need for skills that are not directly relevant to the job in test administration in the testing situation can be particularly important. For example, if an applicant for a job requiring manual dexterity skills and minimal reading skills beyond the initial training period has a visual disability that makes reading printed instructions impossible but is not so severe as to limit the manipulations performed on the job, the employer should find an alternative method for evaluating manual dexterity that does not require reading. Such issues highlight the importance of making a clear distinction between the KSAOs to be measured and the KSAOs required to take the test when deciding the type of test to use as well as considering appropriate accommodations for candidates with qualified disabilities.

It is important to note that not all employment tests are designed to predict job performance. Frequently, employers want to know how likely an applicant is to turnover or be absent or to exhibit organizational citizenship behaviors or counterproductive work behaviors. (See Chapters 20–24, in this volume for a discussion of criterion constructs in employee selection.) In such cases, a test may not be directly related to a KSAO required to perform an important job task but is nevertheless a predictor of an outcome that is important to the organization. Regardless of the criterion, it is incumbent upon the test user to demonstrate the relationship between the test and the criterion of interest.

Feasibility of Measuring the Construct

Some important constructs can be notoriously difficult to measure for a variety of reasons, such as lack of a clear definition, psychological and/or statistical multidimensionality of the construct, and subjectivity of scoring (Shute & Wang, 2016). For example, highly predictive measures of an individual's creativity are difficult to find or develop. As a result, a key factor in deciding which constructs to measure will be the extent to which the constructs can be assessed validly and reliably.

In addition, there are usually organizational constraints (e.g., budget, staffing context) that limit the feasibility of assessing some constructs. For example, if an employer has no employment offices and only administers computer-based tests in an unproctored environment, then a direct measure of oral communication skills or physical abilities would not be possible. Similarly, if an employer plans to test a large volume of candidates, then a test designed to assess each candidate's physical strength may not be a feasible option unless some screening to narrow the applicant pool is done first. Additional details regarding organizational constraints are also covered in the Administrative Concerns section later in this chapter.

Number of KSAOs to Measure

A job analysis often results in many more KSAOs that are important and required at entry for a job than are feasible to measure (see Chapter 6 in this volume for more information regarding work analysis), but there is no clear guidance regarding the degree to which the job content domain should be covered in the selection program. While all would argue that the KSAOs that are measured in an employment test must be important, few would suggest that all important and needed at entry KSAOs should be measured. The SIOP *Principles* (2003) indicate that measurement of all the important KSAOs is not necessary: "Not every element of the work domain needs to be assessed. Rather, a sample of the work behaviors, activities, and worker KSAOs can provide a good estimate of the predicted work performance" (p. 24). In contrast, Goldstein, Zedeck, and Schneider (1993) suggest using a guide of measuring KSAOs that linked to at least 50% of the tasks (in a content-oriented validity study) and view measuring only 10–20% of a job as problematic. However, they also acknowledged that measuring KSAOs linked to 50% of the tasks might not be possible in some cases.

In the U.S., when a selection practice is challenged under Title VII of the Civil Rights Act of 1964 as amended in 1991, the user of a test(s) that is supported on the basis of evidence from

a content-oriented validation study may need to defend how much of the job content area is measured by the test(s). Despite the litigation, court opinions have varied on what is sufficient job content representation. Thus, in practice, it is not clear how many KSAOs should be measured or how much of the job content domain should be measured.

Perhaps, the most important considerations in determining the number of KSAOs to measure are the criticality of each KSAO relative to job performance and the extent to which one KSAO may compensate for another. When two or more KSAOs are critical for job performance, all may need to be measured. For example, in jobs that are physically demanding and require cognitive skills, incumbents need both sets of skills. A cable splicer may need the cognitive skills associated with splicing cables as well as the physical abilities associated with ascending and descending utility poles. One cannot do the job if he/she cannot climb the pole or if he/she cannot splice a cable correctly. In some jobs, one skill compensates for another. A customer service job may require both interpersonal skills and problem-solving skills; however, in some cases, a lower level of problem-solving skill may be compensated by higher interpersonal skills and vice versa, although a minimum level of each may be required.

In many situations, the number of KSAOs is expanded to measure the broader job because multiple criteria are valued. For example, employers that are concerned about job performance and prosocial behaviors may measure KSAOs related to both criteria. In the U.S., where litigation concerns prevail, another rationale for increasing the number of KSAOs measured is rooted in the hope of minimizing subgroup differences, which open the door to legal challenges. For example, an employer might add a reading test to a math test (assuming reading is a job-relevant KSAO) if the historical mean group differences on the tests indicate that women do better on reading and men do better on math even when the reading test does little to improve the level of prediction.

Another way to determine the number of tests that should be used is to evaluate the incremental validity of each test when criterion-oriented validity data are available. However, it merits noting that test users often find little quantitative support for multiple predictors in a criterion-related validation study beyond the first few. When a content-oriented validation strategy has been used, incremental validity data are not available, and the test user can only rely on data regarding the KSAOs that are important, needed at entry, and linked to one or more critical tasks. As noted in the SIOP *Principles*, "The sufficiency of the match between (the) selection procedure and work domain is a matter of professional judgment based on evidence collected in the validation effort" (SIOP *Principles*, 2003, p. 25).

As a final cautionary note, often practitioners argue that a test measuring a single, important KSAO can be demonstrated to be job-related and a business necessity by virtue of the results of the job analysis and a criterion-oriented validity study. Employers that strive to minimize costs may minimize the number of KSAOs measured and focus only on those tests that have the greatest payoff in terms of prediction. However, when large mean subgroup differences exist for the selected tests, this approach can be considered risky as regulatory agencies and courts may question the decision to use a test that measures a single, albeit important KSAO, or only a few important KSAOs, based on the rationale that even strong criterion-oriented validity coefficients do not explain a great deal of the variance in performance.

Relationship Between the Goals of the Organization and the Number of KSAOs Measured

As noted above, all constructs measured in an employment test must be important and required at entry; however, organizations have differing views on the number of KSAOs to measure that are related to their goals for their selection program (see Chapter 10 in this volume for more information regarding employee selection and organizational strategy). Many organizations attempt to balance their needs for accurate evaluation of candidates' skills with cost-effective staffing procedures and legal compliance. Organizations that are focused primarily on the cost-effectiveness of their selection programs will pay a great deal of attention to the number

of constructs measured and their utility, choosing to use only those that contribute substantially to the prediction of performance (or other criteria), and they manage costs related to test development, validation, and administration partially by using fewer measures. On the other hand, organizations that are more concerned about the legal defensibility of a selection process may be more likely to include tests that provide broader coverage of the domain of critical KSAOs.

Breadth of KSAOs

In addition to determining the number of KSAOs to measure, the organization must also define the breadth of the critical KSAOs to be measured. Some test users will choose to measure a narrow, homogeneous construct (e.g., addition), whereas others will measure a broader combination of constructs (e.g., math, including addition, subtraction, multiplication, division, fractions, decimals).

A test that measures a unidimensional variable has questions that require similar thought processes and result in similar types of answers. Tests that evaluate a multidimensional construct may involve several different processes and contain questions that elicit different kinds of answers. Different types of items (e.g., math word problems and word analogies) can be found in the same, multidimensional test that measures "mental ability." Some multidimensional tests such as a problem-solving test may measure a single construct that has multiple components. For example, a test user might employ a business case to determine how well job candidates solve problems that require the collection of data from multiple sources and the analysis of quantitative and qualitative data.

HOW SHOULD THE CONSTRUCTS BE MEASURED?

Once the constructs to be measured are identified, the test user must determine the best way to measure them. There are multiple ways to measure most content areas. For example, job knowledge might be assessed through a multiple-choice test of cognitive abilities, work samples and simulations, or interviews. Each measurement approach has its advantages and disadvantages that are relative to the population for which the test is being used. A test format that is acceptable for selecting applicants into an entry-level position in a fast-food restaurant may not be acceptable to executives seeking promotion in their own company. Some of the criteria that should be considered when determining their measurement options are discussed in the following sections.

Timing

One of the primary factors that organizations have to consider when choosing a selection process is the time that will be needed to implement it. Once a need for a new selection process is uncovered, many organizations are impatient to implement the new process. Some organizations lack an existing selection process and need to rapidly develop and deploy one to meet the staffing requirements associated with their strategic direction. Others have detected some problem with their existing program and are anxious to replace the current selection process, which is flawed. Only a few organizations seem to take a continuous improvement approach to employee selection and develop and validate a new selection procedure when the existing one is working well.

The immediate need for a selection process may guide an organization toward off-the-shelf tests and/or tests that can be validated quickly. Rather than creating its own test, an employer may eschew the development process and choose an off-the-shelf test that is ready for a validation study. Some employers will gravitate to a test and rely on a publisher's generalizability study or choose a test that can be validated using a content-oriented validity strategy relatively quickly.

For example, an organization may choose an off-the-shelf test for which a large-scale meta-analysis has been conducted and implement the test based on evidence from other validity studies while planning a local criterion-related validation study based on applicant data to be collected in the coming months. Another may choose a work sample test that can be validated using a content-oriented strategy to avoid the time and costs of a local criterion-oriented validation study.

Group Differences in Test Score Means and Adverse Impact in Selection Decisions

Many organizations embrace a diverse workforce because it contributes to the achievement of their strategic objectives, and in the U.S., they want to avoid unnecessarily eliminating members of protected groups. Thus, these organizations consider the available evidence of differences in score means among subgroups of interest when choosing a test.

Although mean score differences are not the same as adverse impact calculations, they are often related. Adverse impact may be assessed in several ways, ranging from four-fifths ratios to statistical tests of significant differences between pass rates. Regardless of how adverse impact is assessed, organizations must decide whether to avoid, reduce, or eliminate it through their choice of tests, decisions on cutoff scores, or some other approach such as alternate recruitment strategies. For tests of some constructs, group mean differences cannot be easily eliminated, and the organization must prioritize its goals and decide if the construct should be measured at all.

Consideration of Alternatives

In the U.S., employers are required to search for alternative selection procedures that have equal or greater validity and less adverse impact:

> Where two or more selection procedures are available which serve the user's legitimate interest in efficient and trustworthy workmanship, and which are substantially equally valid for a given purpose, the user should use the procedure which has been demonstrated to have the lesser adverse impact.
>
> (Uniform Guidelines, Section 3B)

Although the Uniform Guidelines do not clearly outline the steps that should be taken to comply with the alternatives provision, practitioners often consider a variety of options, such as different measures of the same construct, measures of different constructs, different ways of combining measures, different methods of making selection decisions (e.g., setting pass/fail cutoff scores, banding scores, using top-down selection), and different cutoff scores.

Practical issues make the comparison of alternatives difficult. A comparison of the validity and adverse impact of two alternatives often assumes the availability of criterion-oriented validity data and adverse impact data for both instruments. However, such data are not always available. For example, validity data may have been collected for different jobs or with different criteria (e.g., job performance, organizational citizenship behaviors). Or, adverse impact statistics may have been collected on an applicant population in one situation and on the incumbent sample that participated in a criterion-related validity study in another. Moreover, content-oriented validity strategies do not yield validity coefficients, and many content-oriented validity studies do not produce estimates of adverse impact unless the tests have been administered in a pilot study. Question 51 of the *Uniform Guidelines* indicates that the strength of validity evidence is determined by the proportion of critical job behaviors and the extent to which the test resembles work behaviors. An additional complication is the lack of operational specificity in the Guidelines about what qualifies as "substantially equally valid" or "lesser impact." Thus, even if two tests have comparable validity coefficients and adverse impact data available, there is no commonly accepted method or standard for determining how much of a difference in validity coefficients or adverse impact findings is sufficient to warrant the use of one test over the other; instead, the decision must be based on the test user's professional judgment. These

difficulties with the process of identifying alternatives with equal or greater validity and lesser impact may explain why, in practice, some test users do not make thorough reviews of alternatives, as pointed out by Guion (1998).

Reviews of a wide array of alternative selection procedures may be limited in some circumstances. In practice, some organizations choose a consulting firm to develop and validate a selection program, knowing that the tests and alternatives considered will be limited to the firm's own proprietary tests. Theoretically, at least, the organization may have considered a broader set of tests in the process of selecting a consulting firm, but this process of evaluating the pros and cons of different test publishers is rarely documented. If a human resources professional or operating manager is choosing the tests without the assistance of a qualified testing professional, he or she may not be aware of the need to consider alternatives and may lack a sufficient understanding of the concepts of validity and adverse impact to make nuanced judgments about different tests (Murphy & Davidshofer, 1988). Furthermore, decision makers lacking a testing background may be easily swayed by test advertisements that broadly claim the "validity" of a test, promise no adverse impact, indicate approval by a governmental regulator or professional association, or extol the ease of administration. Test users who fail to understand that validity relates to the inferences made from test scores for a particular job and is not a characteristic of a test, or who are not familiar with different ways of calculating adverse impact or the effect of different applicant populations on adverse impact often have difficulty evaluating alternative procedures.

Consideration of Past Legal and Administrative Challenges

Another factor that is often considered in the U.S. is a test's history of legal and administrative challenges. Although theoretical information and research findings supporting a test that evaluates a required KSAO, as well as the administrative requirements of the test, are primary drivers of test choice, the outcome of previous litigation or grievance and arbitration procedures involving the test user's own organization or others' can inform the test user about the potential liabilities associated with a particular test (see Chapters 26–30, in this volume, for a more detailed account of legal issues related to employee selection).

Several characteristics of tests seem to increase the likelihood of some form of legal scrutiny. Tests that generally produce large group differences (e.g., multiple-choice measures of cognitive ability) are more often challenged than are those with smaller group differences (all other factors being equal). Certain ways in which a test is used also tend to attract more attention. For example, high-volume selection programs seem to be challenged more frequently than those tests used for smaller numbers of applicants. A test used for selection into an entry-level position in a large retail firm appears more likely to be challenged than one used for the selection of executives in a small professional services organization. Tests used for promotions or upgrades and transfers in a unionized setting are often the catalyst for a grievance. Some test practices also appear to be lightning rods for challenges. Test users often avoid selecting a personality test that uses a lie scale or a social desirability scale for promotions because of the implications of dishonesty on the part of the candidate and the test takers' negative reaction to these scales. Indeed, many organizations with a represented labor force will avoid any selection tool that does not have universally right and wrong answers. Additionally, some test formats are challenged less frequently than others. For example, structured or unstructured interviews are less frequently reviewed than multiple-choice tests in general, possibly because of the ubiquity of the use of interviews. Similarly, high-fidelity work samples that are obviously similar to the job for which they are used seem to be challenged less frequently than more abstract tests of basic skills.

The type of test used in combination with the feasibility of an appropriate validity study may also have legal implications for the test user. When a criterion-oriented validity study is not feasible because of the size of the available sample, test users may gravitate to a test that can be appropriately validated using a content-oriented strategy under the *Uniform Guidelines*. Fewer questions may be raised about a work sample test that is validated using a content-oriented strategy than a personality inventory or a more abstract measure of cognitive ability that is validated

using a content-oriented strategy. Thus, when a criterion-oriented validation strategy is not possible, some organizations will consider only those assessment tools for which a content-oriented strategy will provide compelling evidence of their relevance and validity. Nevertheless, some organizations believe a criterion-related study produces better evidence of validity in terms of legal defensibility and choose instruments for which there is a history of successful criterion-related studies.

Although past challenges may influence test choice, test users often do not have access to detailed information about these challenges in a timely manner. The decisions in many court cases can be a long time coming, and grievance information is not often shared outside of a particular organization or labor organization. By the time the test user has had an opportunity to access publicly available information about testing litigation, the test in question may be out of date or there may be new and better options.

Administrative Concerns

For many organizations, a major concern is whether or not the test can be administered appropriately in their staffing environment (see Chapter 8, in this volume, for additional information on administrative concerns associated with employee testing). Frequently, an organization's staffing process and resources will dictate the choice of a particular test. For example, a multiple-choice test of cognitive ability with a fixed number of items that must be proctored is not practical in an organization that processes applications via the Internet and lacks the facilities and staff available to administer tests in proctored settings. Similarly, a lengthy assessment process that requires one-on-one assessors for scoring is not practical for high-volume jobs that have substantial amounts of turnover. Some common administrative concerns are discussed below.

Administrative Personnel The test user must consider whether the personnel required to administer and score the test are available and affordable. Some tests require an administrator who facilitates the test administration by reading instructions, distributing and collecting the testing materials, timing the test, scoring the test by comparing answers to a scoring template, etc. In other situations, the administrator's role may be more complicated and require more complex skills and training, such as setting up equipment, serving as a role player, or making judgments about a candidate's performance. For example, a work sample test measuring knowledge and skill in welding metal parts may require that a certified welding expert score the sample. When such an expert is not available, a work sample may not be feasible.

Cost of administration personnel is another related factor that is critical. Even if personnel with the prerequisite skills are available to administer tests, an organization may find the cost of using many of these employees prohibitive or at least greater than the return on the investment warrants. For example, an organization using a structured interview as a first screen for a high-volume, high-turnover position may well find the cost of the interviewer exceeds the value of the interview as a screening tool.

Some organizations make the mistake of failing to consider all of the associated costs of test administration. In addition to time spent administering and scoring tests, many tests require administrative personnel to be trained, retrained, calibrated, and monitored. For example, many structured interview programs include extensive training to ensure that all interviewers, regardless of location, are administering the interview properly and using the behavioral anchors in the same manner for scoring. Some of these companies also offer "refresher" training to reinforce the standards for the interview. Some organizations bear the additional cost of monitoring test administrators and scorers to promote strict adherence to administration rules, testing policies and procedures, data recording requirements, etc.

In the past 20 years, many organizations have abandoned proctored testing in favor of unproctored Internet testing (UIT). Although most testing professionals who have adopted UIT agree that one of the primary advantages is the reduction in administration costs (e.g., Tippins, Beatty, Drasgow, & Pearlman, 2006), most also agree that administration costs are not eliminated entirely, and significant costs related to IT personnel remain. Thus, the costs of

developing or buying a computer-based test administration system, maintaining it, and monitoring it must be factored into the testing decision. Although many organizations purchase these services from an outside vendor, these costs are embedded into the cost of the test.

Testing Facilities and Equipment In addition to adequate personnel to administer selection tests, equipment and facilities appropriate for the tests are also necessary. Some tests rely on paper copies of test materials and pencils; others are computer-based or require telephones or video equipment. Today, many computer-based tests require an Internet connection. Although most in the U.S. assume the availability of reliable electric power and an Internet connection, that assumption may not be true in all parts of the world, and the test user may need to take extraordinary efforts to find appropriate facilities. Mobile testing can also present challenges related to Internet connections. Work samples can involve almost any equipment that might be used on the job, ranging from simple machinery to complex electronics. One of the more expensive forms of testing in terms of the facilities required is an assessment center that requires space for participants to work independently, assessors to privately score exercises and/or conduct interviews, as well as additional space for group exercises with other participants at the center.

When administration costs are a factor in test choice, the test user must consider not only the initial cost of the facilities and equipment but also the cost for ongoing maintenance and upgrades. For example, an assessment center facility has to be acquired and maintained through the life of the assessment process. Video-based assessments containing pictures of individuals can become quickly outdated because of clothing or hairstyles. Technology-enabled assessments must be implemented with the necessary equipment and facilities, but many programs will be upgraded, requiring additional, newer, or more powerful equipment and the places to put it. In addition, user interfaces may need to be overhauled to achieve a "modern" look even when the existing interface is functional. Physical abilities testing can also require expensive facilities or equipment. High-fidelity physical abilities tests may include equipment such as that found on the job (e.g., ladders, utility poles, stretchers, stair chairs), which must be maintained. Measures of physical capability may include abstract kinds of physical ability tests that use equipment such as tensiometers, which can be expensive to buy and maintain because they must be recalibrated regularly to ensure accurate ratings.

Although cost may be the overriding concern for many organizations when considering testing facilities, the mobility of facilities and equipment can also be important in certain situations. When large numbers of applicants must be tested in multiple places, movement of bulky equipment (e.g., computers or equipment for work samples) may not be feasible, or if feasible, it may not be cost effective to purchase redundant equipment or move it around. Similarly, facilities for complex physical abilities tests or assessment centers may be expensive and time-consuming to replicate in multiple locations.

Proctored and Unproctored Internet Testing An issue that has continued to gain more attention from test users is the question of whether to use proctored or unproctored Internet testing. The advantages and disadvantages of UIT are well-documented (Tippins et al., 2006). In a nutshell, advantages usually include lower administration costs, standardized instructions, faster time to testing, and broader applicant pools. Disadvantages frequently cited include nonstandardized test environments that have the potential to affect test scores, cheating, including the inability to identify the test taker, and threats to the security of test materials. In choosing to use a proctored or unproctored test, the test user must weigh the pros and cons to determine what works best for the organization.

Time Requirements for Test Administration The time to administer a test has two important implications for staffing programs: (1) the cost of administration and (2) the impact on applicant flow. The time spent testing is related to the costs of the personnel required for administration and the use of facilities and equipment and can be driven by several factors. In proctored testing or testing that requires an assessor or interviewer to participate in the test administration, the longer the test, the greater the personnel cost. A biodata form that takes 30 minutes to complete will cost less in terms of administration time than will a one-hour structured interview

or a three-hour business case. Tests that have high reading demands can increase the administration time and the costs. For example, situational judgment inventories (SJIs) are often more time-intensive than other forms of tests because of the reading load. There are often trade-offs with respect to test time. Both high- and low-fidelity simulations often convey a great deal of information about the job and organization. Although the testing time may be somewhat longer than for more abstract tests, there are additional benefits in the form of information about the job, work environment, and company associated with the additional time.

Another concern about lengthy tests is their impact on applicant flow. A common perception is that applicants have a low tolerance for lengthy evaluations, especially when they are administered in unproctored Internet settings, although recent literature indicates this may not be a correct assumption (Speer, King, & Grossenbacher, 2016). Recruiters often point to the abandonment rate on UITs; however, there may be multiple reasons why a candidate chooses to stop testing. Some applicants may be exploring jobs and not have a sincere interest in the one for which the test is required; some may begin the test and realize the job has requirements that do not match their skills; and some may be distracted and leave the test. In addition, lengthy applications confound the applicant's perception of the amount of time spent testing. Abandonment may be the result of the test in addition to the application process and not the test alone. Even when the applicant is asked to take a test at an employer's site, the amount of time spent testing can be a deterrent to maintaining the applicant's interest in employment. Employed applicants may be particularly reluctant to invest significant amounts of time in face-to-face testing in another firm. The number of tests administered can also pose a challenge to keeping an applicant engaged. When the selection program is based on multiple hurdles and the applicant is asked to return multiple times to take tests in the sequence, the problem of keeping the applicant engaged is exacerbated.

Consequences of Poor Hiring Decisions

Another source of costs related to selection programs comes from hiring an applicant without the necessary skills. When the consequences of hiring someone who does not have the necessary skills are severe (e.g., an error caused by an employee without the requisite skills leads to injury, death, or widespread property damage, or the cost of training a replacement for inadequate employees is high), accurate prediction of future job performance is critical. Thus, test users often look for a selection procedure that covers more of the KSAOs required for the job or measures them in a more reliable and valid manner. For example, if extensive on-the-job training is necessary, the cost of training a replacement for an unsuccessful employee could justify the cost of a more elaborate hiring system. Similarly, a more comprehensive selection system might be chosen when hiring individuals for highly critical positions with no or minimal margin for error, such as flying commercial airplanes or operating heavy machinery. In contrast, an organization might choose other, less extensive selection instruments when the repercussions of an error are relatively minor or when the cost of training a replacement employee is minimal.

Organization Reactions

Many experienced testing professionals who have been in a position to review and select the appropriate assessment tools to implement within their organizations have learned that the organizations for which they work often have strong likes and dislikes for various types of tests. As noted above, many of these preferences are related to the goals of the organization. Some of the preferences that are expressed by members of an organization collectively are related to the image they want to project to applicants. For example, some organizations promote the idea that anyone can perform any job with some hard work and a little coaching. Thus, tests that measure relatively immutable traits (e.g., personality tests) or are based on past experiences (e.g., biodata) and that evaluate skills that generally cannot be developed are not acceptable. Instead,

tests measuring skills and abilities (e.g., skills tests, achievement tests) or knowledge that can be acquired are preferred. Some organizations espouse the idea that selection should be based only on skills related to performance and that measures of how that work is accomplished are not appropriate. In such situations, the organization might use tests that only relate to performance outcomes and not to the way the outcomes are achieved. For example, a sole job knowledge test might be used instead of a job knowledge test in combination with a measure of interpersonal skills. Some organizations promote the idea of selecting the best by hiring individuals who have proven their skills by graduating from top schools. Consequently, asking individuals to demonstrate their mental prowess through measures of cognitive ability is anathema. Many organizations have a decided preference for face valid tests in hopes of avoiding the challenge of explaining the relevance of a less face valid test. So, instead of using a personality inventory to gauge interpersonal skills, an organization might use a customer service simulation. While these organizations are likely to select work samples and simulations, other organizations want to assure test takers of the company's objectivity in selection and choose only instruments that involve no human judgment. These organizations might avoid the simulations that must be evaluated by an individual and rely instead on objectively scored multiple-choice tests.

Applicant Reactions

Sackett and Lievens (2008, p. 439) characterized the lack of evidence for a relationship between applicant reactions and individual or organizational outcomes as "the Achilles heel of this field." Nonetheless, many believe that applicants' reactions to the testing experience can have important implications for organizations, such as influencing the applicants' intention to remain in the selection and hiring process and accept job offers, affecting their attitude if hired, increasing the possibility of legal action if the selection process is deemed inappropriate, and increasing the likelihood of sharing their negative experience with the organization to others (Bauer, McCarthy, Anderson, Truxillo, & Salgado, 2012).

Applicant reactions have become increasingly important to organizations in recent years, with many organizations focusing their efforts on creating and promoting an employer brand. As part of this effort, employers are looking for tests that are shorter, more modern looking, or more entertaining (e.g., simulations, games). Test users must be aware, however, that shorter tests often have lower validity than their longer counterparts, and more entertaining tests often cost a great deal more to develop or implement and may not be any more valid than less entertaining tests.

Several research studies have investigated the role of various factors, such as test type, administration format, procedural characteristics, and personal variables, on applicant reactions to the testing event (see Gilliland, 1993, for a theoretical model of applicant reactions). In general, research indicates that tests are perceived more positively when the relationship between the content of the test and the duties of the job is clear to the applicants. However, as is the case with other test selection criteria, an applicant's reaction is not the only factor in deciding which test to use. If a job requires cognitive ability, the finding that cognitive ability tests are not perceived as favorably by applicants as interviews and work samples may be irrelevant. More research is needed before concluding that applicants' reactions to selection procedures actually predict applicant behaviors (e.g., withdrawal from the selection process, job acceptance, job performance). Nonetheless, applicants should be treated fairly and consistently because of legal, moral, and ethical constraints on the organization. Moreover, applicant reactions to the testing procedures are likely to result at least in more positive perceptions of the organization.

HOW SHOULD ASSESSMENT TOOLS BE EVALUATED?

There are several ways to evaluate the validity of assessment tools and a number of factors that influence the choice of validation strategy, which are described in the following sections. In addition, professional guidelines such as the *Standards* and the *Principles* suggest that an accumulation of evidence of validity strengthens the support for the inferences made from a test score.

Information on the Validity of the Test

Validity is a critical factor when selecting assessment tools (Chapters 2 through 4 of this volume contain more detailed discussions of validity.) When choosing test instruments, testing professionals often review past validity research to help identify which selection instruments will be useful to measure certain constructs. Data from past research can provide information regarding the validity of a particular test type in predicting various outcomes and the incremental validity of using various assessment types in conjunction with other forms of tests (see Schmidt & Hunter, 1998). Although innovation in testing processes can be helpful, it is often unwise to use a test for which the extant evidence provides little or no support for the kind of inference to be made (e.g., using a typing test to measure conscientiousness).

A review of the validity evidence for a particular test use is sometimes overlooked in practice, particularly when individuals who do not have training in industrial-organizational (I-O) psychology choose the tests for the experimental battery (Rynes, Colbert, & Brown, 2002). There are times when untrained test users attempt to review validity evidence from publishers or the I-O literature, but they fail to understand technical issues well enough to make sound judgments. For example, a test user may believe a test is valid for a particular use because a study of the test indicates a seemingly large correlation between predictor and criterion, not grasping the importance of significance testing or effect sizes. Or, the test user may not understand the importance of cross-validation when items are selected based on their correlations with the criterion measure. In addition to the test user lacking the knowledge and skill necessary to understand the concept of validity and review related technical materials, the test user may fail to conduct a review of the literature because he/she does not know where to get information about the validity of a test or there is no information. A few test users may even discount the value of such information, instead relying on idiosyncratic beliefs about the constructs and tools that predict job performance and other outcomes of interest (e.g., turnover, absenteeism, integrity).

Appropriateness and Feasibility of Validation Strategies

Several validation strategies can be used to demonstrate the validity of inferences made from tests (e.g., content-oriented strategies, criterion-related strategies, validity generalization techniques). However, the feasibility and appropriateness of different validation strategies vary based on a number of factors, including those outlined as follows.

Type of Test—Although content-oriented studies and criterion-related validity studies can be conducted for any test that produces a score, different validation strategies are often used for different types of tests. For example, evidence of validity for a structured interview often comes from a content-oriented study, and evidence for a numerical reasoning test frequently comes from a criterion-related study. There are several likely reasons for this choice. First, the relationship between the constructs measured by the test and the critical KSAOs is probably easier for subject matter experts (SMEs) to evaluate when the test constructs are more similar to the KSAOs. For example, SMEs may be more likely to see the relationship between a work sample test that measures electronic repair and a critical KSAO such as knowledge of electronics than the relationship between a number series test and knowledge of electronics. Second, instruments like structured interviews and work samples are often developed for use with a smaller number of applicants than more abstract measures that are often used for screening purposes. Small sample sizes make criterion-related validity studies technically infeasible.

One of the primary determinants of appropriate validation in the U.S. is the perception of which validation strategy is legally defensible. The *Uniform Guidelines* state that "A selection procedure can be supported by a content validity strategy to the extent that it is a representative sample of the content of the job." The *Guidelines* also reject content-oriented validation strategies for measures of "traits or constructs, such as intelligence, aptitude, personality, commonsense, judgment, leadership, and spatial ability" (Section 14.C.1). Thus, some tests may be technically validated using a content-oriented strategy only with some concern for legal defensibility if the

test is challenged. Although not consistent with professional guidelines (e.g., *Standards for Educational and Psychological Testing*, *Principles for the Validation and Use of Employee Selection Procedures*), some believe that criterion-related validity is the "gold standard" for successful legal defense.

Some organizations consider factors such as those outlined above when choosing a type of validation effort, but other organizations ignore the process of validation altogether. These organizations put themselves at a disadvantage not only in terms of missing the benefit of identifying the most predictive hiring tools and collecting evidence of the effectiveness of the selection program but also with respect to opening themselves up to potentially costly litigation in the event that their hiring practices are challenged and found wanting.

Size of Incumbent Population—Organizations developing tests for jobs that have few incumbents and low hiring volume are unlikely to be able to execute a criterion-oriented validation study because these studies require relatively large sample sizes to obtain the sufficient power for statistical analyses. In such a circumstance, the test user sometimes resorts to content-oriented validation. In other situations, alternatives such as a validity generalization strategy, including a transportability study or a meta-analysis of validity studies involving relevant measures and criteria, are employed to establish evidence of validity. Additionally, organizations with small incumbent populations may use a synthetic or job component validity approach in which validity inferences are based on the synthesis of the relationships between scores on a test and measures of performance on a component of the job.

New Jobs—New jobs can pose special problems for test validation. A concurrent criterion-oriented validation study is clearly not feasible due to the lack of incumbents and supervisors available to complete test and performance ratings. A traditional content-oriented validation approach may not be possible either due to the lack of SMEs available to provide input about a job that does not exist. Occasionally, another source of expertise about the job is used to provide task and KSAO ratings for the new job and to establish linkages between the tasks and the KSAOs, and the KSAOs and the proposed tests. For example, information can be gathered from those who designed the job about the work tasks, processes, and equipment as well as the impetus for the newly created job, proposed minimum qualifications, jobs from which current employees will be promoted, proposed training, and similar jobs from external sources (e.g., O*NET™, the I-O literature).

Test Security—Another factor that might limit the type of validation process selected is the level of test security required. Some validation strategies (e.g., concurrent criterion-oriented) require that internal employees complete the tests experimentally. When the need for test security is high, the involvement of organization personnel in the test design process or validation effort may raise questions about the security and confidentiality of the test content.

Existence of Robust Database of Validation Studies—When a sufficient database of validation studies is available to the user, validation based on generalization strategies instead of criterion-related or content-oriented validity approaches may be an option for the test user. In some cases, this database will come from a test publisher that maintains records of validation studies conducted using the firm's tests as predictors. This type of database may provide sufficient evidence for the test user to reasonably believe a test is likely to be valid in the local setting so that the test can be used on an interim basis until the test user's company can gather additional local validity evidence to support the test use. In other cases, the validation data may come from a database of validation studies internal to the organization that will facilitate validity transportation.

Cost of Test Development and Validation Studies and the Utility of the Selection Program

In virtually every organization, costs are a consideration when developing and validating selection tools. These processes can be expensive when an outside consulting firm is used to develop and validate a test; however, even when the test development and validation work is conducted in-house, the validation effort can be expensive as qualified professionals still cost the organization money. Regardless of who performs the test development and validation

work, internal personnel must perform many tasks, such as coordinating study participants, ensuring appropriate communications, and collecting background data on populations and other archival data relevant to the study. Job incumbents, supervisors, and other SMEs may take time away from the job to participate in various components of the project (e.g., answer job analysis surveys, make linkage ratings, complete experimental tests, or provide criterion data). There can also be costs associated with the equipment and supplies required to construct the test (e.g., work samples) and conduct validation efforts (e.g., laptops for employee testing). As a result of these many costs, organizations often seek ways to minimize their expenditures and take steps to reduce the cost of test development (e.g., use off-the-shelf tests) and validation effort (e.g., rely on validity generalization strategies such as the transportation of validity).

The source of funding for these efforts can become an important factor. For example, in some organizations, test development and validation expenses are paid from a limited, centralized human resources budget, whereas test administration costs may come from richer, decentralized operational budgets. In other organizations, the opposite is true. In the first case, an organization might be motivated to select tools that are less costly to develop (e.g., commercially available tests, interviews) or less costly to validate (e.g., those that can be justified through a transportability study or a content-oriented validity strategy). Because budgets are usually managed on a yearly basis, organizations may use off-the-shelf tests even when the ongoing licensing fees are more costly overall than the development of proprietary tests that have fewer recurring costs. Alternatively, the organization that has a higher budget for test development and validation may develop and validate a custom test tailored to its industry, core values, or culture rather than buy a commercially available test with ongoing licensing fees. The volume of test use may also be related to cost considerations. Under circumstances in which the volume of test use will be extremely high, the cost of ongoing test licensing may so greatly exceed the initial upfront costs of developing a proprietary test that test development becomes more economically viable.

Another factor that can influence the organization's approach to test development and validation is its perceptions of the test's value. When an organization uses company-specific equipment or process, or has a unique culture that is not reflected in off-the-shelf tests, a proprietary test tailored to the needs of a specific business may be needed. Similarly, if the organization has confidence in the value of its selection program and believes the selection process offers a competitive advantage, then the business may seek to develop a test that is specific to it. Occasionally, organizations simply want to avoid the repercussions of other companies' poor testing practices. For example, if a competitor in the same geographic market has poor testing practices, an organization may seek a different off-the-shelf test or develop its own unique test. When the value of a business's services is derived from something other than its employees (e.g., natural resources), a test shared with other similar companies may be sufficient for its needs. In a few situations (e.g., utility companies), where one organization dominates a geographic area and applicants come primarily from regional pools, tests shared with other organizations from different geographic areas tend to have little effect on the organization's competitive advantage.

As a final note, it can be argued that the ultimate measure of a test's value to the organization is its utility, which takes into account not only its costs but also its benefits. Although testing professionals often struggle to identify and estimate all costs and the value of all benefits, both tangible and intangible, they should consider the costs to develop, validate, and administer a test relative to its benefits.

HOW SHOULD TESTS BE ADMINISTERED?

There are multiple ways to administer a test. Currently, one of the most discussed questions about test administration is whether or not the test should be proctored. However, other dimensions of test administration affect the selection of tests. Several of the more common questions about test administration that affect the choice of test are discussed as follows.

Proctored or Unproctored Testing

The essential question about unproctored testing appears to be whether the risk of cheating in any form, which can decrease the validity of the test, justifies the advantages resulting from the tests that are administered on the candidates' equipment at times and places of their convenience (see Chapters 39 and 41, in this volume). Ideally, the user considers the types of items used in the test as well as the consequences of bad hires when deciding whether to test in unproctored environments. While there are few ways to cheat on unproctored self-description inventories that cannot also be used in proctored settings, a number of maleficent behaviors can be used when there are clear right and wrong answers that can increase test scores in ways that do not reflect the test taker's ability. When the consequences of failure to perform are significant, many employers will avoid unproctored testing and opt for monitoring test takers during administration. When the staffing context requires unproctored testing, test users should consider carefully both the type of test to administer and the implications of the test taker's opportunity to cheat and choose tests accordingly.

Speeded Test or Power Test

Another frequent consideration in test administration is the use of time limits. Although some constructs (e.g., measures of perceptual speed and accuracy) require speeded tests, others do not. In such cases, the test user must decide what, if any, time limit to place on the testing time. Test users often impose a time limit for administrative reasons. In proctored settings, the time limit allows for efficient scheduling. In unproctored settings, a time limit may inhibit some forms of cheating.

There are few rules about how to set a time limit on a test, but several factors should be considered. In the U.S., where accommodations for individuals with disabilities can be an important element of test administration, time limits are often generous to reduce the need for adjustments in administration times. For example, a user might set a time limit that allows 90% of test takers to complete 90% of items. Firms concerned with test taker reactions may set generous time limits to avoid negative test taker reactions when the test is difficult for most candidates to finish. For some tests like business case assessments, time limits are set to standardize the exercise and allow the organization to learn what candidates can do in a set amount of time.

Group or Individual Administration

Many tests can be administered either individually or in a group setting, and the choice of which to use may depend entirely on the staffing model the employer uses. However, some tests (e.g., many physical abilities tests, structured interviews, and work samples) require individual administration and scoring. When resources are insufficient to allow for this, alternative forms of testing must be found.

Test Preparation

Many employers provide test preparation materials that explain what is being measured, how the test is scored, what can be done to prepare for the test, what are the rules regarding testing, etc. Other employers offer practice tests that familiarize test takers with the test and provide them with some idea of how their practice scores compare to the test standard for the job to which they are applying. At times, the practice test feedback is accompanied by developmental suggestions intended to improve the skill being measured. For example, employers who use physical abilities tests may offer a practice test, feedback, and developmental suggestions on improving upper and lower body strength, flexibility, etc. The intent of many of the preparation efforts is

to inform test takers of what to expect so that their scores more closely reflect their ability and not their comfort with or savviness for taking tests.

In choosing the type of test to use, the test user must consider whether or not to offer test preparation materials and how to provide access to all candidates. Because test preparation materials and practice tests represent another source of costs and may have implications for applicant reactions and adverse impact, the amount of test preparation required and the guidance to test takers needed can be factors in the choice of test. The test user must also decide what kind of guidance to provide for tests that measure skills that are difficult to develop (e.g., personality tests).

HOW SHOULD SCORES BE USED?

After tests have been identified or developed and validated, the test user must consider how to calculate, report, and use the resulting test scores. (Chapters 8 and 18 in this volume contain additional information regarding the use of test scores.) Considerations related to these decisions include the form of the test score used (e.g., raw score, percentile score, score bands), the method for combining test scores (e.g., compensatory, multiple hurdle), and the operational use of test scores (e.g., top-down selection, banding). Additionally, decisions need to made regarding what type of feedback (if any) to provide to test takers.

Calculation and Form of Reported Test Score

A variety of methods can be used to calculate test scores (e.g., points are given for a single correct answer, points are given differentially for each possible response to a question, a different number of points is given for answers to different questions depending on the difficulty of the question, points are subtracted for guessing). Additionally, the final test score can be presented in a variety of forms (e.g., raw score, percent score, percentile score compared to a norm group or to the current group of test takers, standardized score). Various factors should be considered when determining what kind of score to report, including the type of test (e.g., power versus speeded), the construct measured (e.g., cognitive ability versus personality), the number of competencies measured, the availability of appropriate normative groups, the ability of the test score recipient to understand the score, the purpose of the test score, and the reliability and validity of the test score.

Different test types require different score formats. A score indicating the number or the percentage of items the test taker answered correctly may be effective when communicating the extent to which an individual possesses a body of knowledge. In contrast, a number or percent correct score on a personality inventory would be difficult to interpret as there are not technically right or wrong answers; instead, there are responses that describe the test taker's standing on a construct to varying degrees. Similarly, a percent correct would be appropriate on a power test but would be less useful on a speeded test. A standardized score or percentile score might be helpful when information about an applicant's standing relative to other test takers is needed, but less useful when the question posed is how much of some ability or skill a person possesses. In such a case, the number correct or the percent correct might be more useful. If there is no relevant normative group, then the use of a percentile score or standard score that is based on a sample of individuals in an irrelevant group is not informative.

The ability of the test score recipient to understand various types of scores can also influence the decision of how to calculate and present scores. For example, test takers and hiring managers may have difficulty interpreting some forms of test scores (e.g., norm-referenced percentile scores with multiple norm groups), whereas testing professionals may prefer more complex forms of the score that convey more information about the individual.

The purpose for which the test is given can influence the type of score to be provided. If the test is used for selection, all the test taker and hiring manager may need to know is whether or

not the test taker has met the qualifying standard. If, however, there is a developmental component to the test, more detailed information may be warranted. For example, an employee seeking promotion may need to know how far from the test standard his/her score is or what his/her score on each scale is so that he/she can focus his/her developmental activities.

Finally, the reliability and validity of a test or scale should also be considered when determining how to present test results. It may be more appropriate to present more general feedback regarding overall test performance (e.g., pass/fail) than to provide individual scale scores that lack sufficient reliability.

Combining Scores Across Tests

When multiple tests are used in a selection process, a decision needs to be made regarding how to use multiple test scores to make a selection decision. One option is to weight and combine the separate test scores in a compensatory fashion. Another option is to use a multiple-hurdle model in which a cutoff score is applied to each test and applicants must score above each cutoff score to be qualified on the overall assessment. Another alternative is to use a mixed model in which a minimum level of performance is required on certain tests and then the scores are also combined into a single score and a cutoff score is applied to the overall score as well. The method used to combine scores should take into account the requirements of the job as well as available data that may support the decision. For example, a selection procedure for a technical sales job that requires technical skills and sales skills may involve a multiple-hurdle approach when job analysis data indicate that high levels of technical skills do not compensate for low levels of sales skills or vice versa. In another job that requires lower levels of technical skills along with sales skills, combining these test scores in a way that high levels of persuasiveness compensate for lower technical skills may be more appropriate. In still another job in which technical skills are 80% of the job and sales is 20% of the job, a compensatory model that weights scores on tests measuring technical skills more highly than tests measuring sales skill (e.g., 80/20) may be appropriate.

Use of Test Scores

Test scores can be used in many ways for hiring decisions (or progression to the next step in the hiring process). Test scores are often distributed to individuals making hiring decisions as one source of job-relevant information that they use according to their own understanding of the meaning of the test score and the job requirements. Test users can also be provided with an expectancy table and accompanying guidance regarding how the test score should be used. For example, a candidate falling into the top score range may be hired without any other education or experience credentials, whereas another candidate with a score in the lower range may be hired only if he or she has certain levels and kinds of relevant education or experience. Alternatively, strict cutoff scores (for individual tests or a battery) can be established, and decision makers are only given pass/fail information without the opportunity to deviate from the company-wide rule. A common variation to a single cutoff score is score bands that theoretically take into account the unreliability of individual test scores and treat all scores within a band as though they predict the same level of performance. Finally, some organizations use top-down selection by hiring the individuals with the highest scores first.

The best method for using test scores depends on a variety of factors, such as the goals of the organization, the frequency of hiring, and the qualification level of the applicant pool. Top-down selection, for example, can work well when testing occurs infrequently and the employees are drawn from a single pool of qualified applicants; however, it may be less appropriate when testing occurs frequently because the candidate pool changes daily and the top candidate may be different in terms of qualification level from one day to the next. When an organization strives to hire the best of the applicant pool, top-down hiring can help ensure the organization achieves its goals. When an organization uses a test with large group mean differences and desires a

diverse workforce, top-down hiring can present a barrier to achieving the diversity goal. Additionally, top-down hiring for a test with large group mean differences can exacerbate the level of adverse impact and lead to legal challenges. In addition, where the cutoff score is set may have legal implications related to the extent of adverse impact. Despite the desire of some organizations to upgrade their workforces, cutoff scores on tests with large group mean differences that reflect skills that exceed the minimum required to perform the job can be difficult to defend (see *Lanning v. SEPTA*). Organizations may need to collect evidence on the minimally acceptable level of performance to justify a cutoff score.

Another important consideration involves the requirements of the job. In situations where a high level of skill is required on the job, an organization may need to set a floor on test scores to ensure minimum skill levels in all new hires. Top-down hiring could be appropriate if there is a wide range of skill in the applicant population but may result in the employment of unqualified individuals if there are few highly skilled individuals in the applicant pool. Occasionally, organizations will set a minimum score while using top-down hiring to identify qualified candidates for a job.

As noted earlier, decisions regarding many of the factors described in this chapter influence decisions on other factors. For example, consider an organization that chooses to use a multiple-hurdle approach for selection that includes a numerical reasoning test, a reading test, and a situational judgment inventory. On the basis of data from a concurrent criterion-oriented validity study, the organization decides to set a cutoff score on the numerical reasoning test that results in 95% of candidates who pass the numerical reasoning test also passing the reading test and 80% also passing the situational judgment inventory. In this scenario, there is virtually no value in retaining the reading test and little value for using the situational judgment inventory. Thus, the decision regarding the cutoff score for the numerical reasoning test essentially alters the decision of which constructs to measure and what tests to use. Therefore, these types of interactions should be considered when making decisions regarding all of the factors described above, and the test user should be prepared to revisit these decisions repeatedly.

While it can be challenging to identify all of the goals for the testing program and the individual tests and prioritize them, it is important to use tests in ways that meet the organization's goals. Few job aids exist to facilitate the user in determining how to use a test other than an understanding of the organization's goals and knowledge of the impact various decisions have. Recently, some organizations have turned to Pareto optimization methods when considering multiple goals to maximize the levels of goals achieved.

Feedback

When deciding what kind of feedback to offer, if any, most organizations consider a wide array of factors, including the size of the applicant pool, the type of candidate (e.g., internal or external), the expectations of the candidate, the resources of the organization, the employment brand the organization wishes to project, the level of the position (e.g., entry level, executive), and the type of test(s) administered. When organizations test a large number of individuals from outside the organization for an entry-level role that traditionally has high turnover, they frequently provide candidates with basic pass/fail information regarding whether or not they successfully progressed to the next stage in the hiring process.

At the other extreme, an internal candidate applying for a higher-level position who completes tests may expect more detailed feedback (e.g., percentile score for each test/scale) to guide his/her development. The specificity of feedback is particularly important when the internal candidate is not promoted into the new role and is expected to develop in the deficient areas to prepare for the role in the future. Another consideration related to feedback is the type of test completed. It is more appropriate to provide feedback on constructs that can be improved with effort (e.g., knowledge areas) than on more stable attributes (e.g., personality). Additionally, when feedback is provided on tests measuring constructs that can be developed by the individual (e.g., knowledge tests), developmental suggestions are often given in addition to detailed test performance information.

Some employers recognize the role of feedback in shaping test takers' feelings about the company. They strive to provide accurate and constructive feedback in a sensitive manner in hopes of reducing the likelihood of a challenge to the selection program or decreasing negative comments about the organization's staffing process.

CONCLUSIONS

This chapter has reviewed many of the issues test users consider in selecting or developing a test for validation or for operational use on an interim basis. The issues, framed around five questions, are many, and hard-and-fast answers are few. As noted earlier in the chapter, none of these factors can be evaluated without consideration of the others. For example, the feasibility of test development and validation and their costs are significant factors in the choice of tests. An organization with few incumbents in a job for which tests are being considered may not be able to supply enough job incumbents to complete tests for a concurrent study, or perhaps even SMEs for a content-oriented study. Even enterprises with many incumbents may not be able to relieve employees from their job duties for the time needed to assist with test development and validation and maintain smooth operations.

In addition, the answer to a question may need to be revisited depending on the answers to the other questions. An organization that decides to measure only problem-solving ability because it was the most important KSAO for a particular job and then decides to use a work sample test may find that the work sample measures a broader array of KSAOs than just problem-solving abilities. Conversely, an organization that decides to measure all of its important KSAOs may find that the number of tests required is so large that testing requires three days and consequently is unaffordable to the organization and intolerable to applicants. Or, the organization may find that after the first few tests the latter tests add little incremental validity.

A particularly difficult, overarching concern is how to arrive at one decision in the face of many competing demands on the organization. Optimization of all factors is challenging, if not impossible, in most cases. For example, increasing validity while minimizing adverse impact and meeting organizational constraints of time and cost associated with validation and administration remains a balancing act rather than a series of discrete decisions. Minimally, it is imperative that those who are tasked with identifying or developing successful selection systems are familiar with the many decision points in the process. The test user responsible for designing selection systems must consider these issues and their ramifications, weigh the tradeoffs, and make fully informed final decisions.

In many organizations, these questions and their answers must be revisited regularly. Many things about an organization can change quickly. The business needs and strategies change; the staffing context changes; the applicant pool changes; etc. What exists today may not exist tomorrow. Thus, the skilled test user will continually evaluate each selection program to ensure it meets as many needs of the organization as possible.

REFERENCES

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.

Bauer, T. N., McCarthy, J., Anderson, N., Truxillo, D. M., & Salgado, J. F. (2012). What we know about applicant reactions on attitudes and behavior: Research summary and best practices. *International Affairs Committee of the Society for Industrial and Organizational Psychology, Inc.* Society for Industrial and Organizational Psychology, Inc.

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

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*, 694–734.

Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job analysis-content validity process. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 3–32). San Francisco, CA: Jossey-Bass.

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

Murphy, K. R., & Davidshofer, C. O. (1988). *Psychological testing: Principles and applications*. Englewood Cliffs, NJ: Prentice-Hall.

Rynes, S. L., Colbert, A. E., & Brown, K. G. (2002). Human resource professionals' beliefs about effective human resource practices: Correspondence between research and practice. *Human Resource Management, 41*, 149–174.

Sackett, P. R., & Lievens, F. (2008). Personnel selection. In S. T. Fiske, A. E. Kazdin, & D. L. Schacter (Eds.), *Annual review of psychology* (pp. 419–450). Palo Alto, CA: Annual Reviews.

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.

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

Shute, V. J., & Wang, L. (2016). Assessing and supporting hard-to-measure constructs. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and application* (pp. 535–562). Hoboken, NJ: Wiley.

Speer, A. B., King, B. S., & Grossenbacher, M. (2016) Applicant reactions as a function of test length: Is there reason to fret over using longer tests? *Journal of Personnel Psychology, 15*, 15–24.

Tippins, N. T., Beatty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored Internet testing in employment settings. *Personnel Psychology, 59*, 189–225.