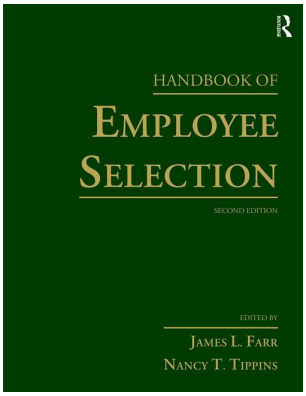

Handbook of Employee Selection

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

Professional Guidelines/Standards

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-27
P. Richard Jeanneret, Sheldon Zedeck
Published online on: 22 Mar 2017

27

PROFESSIONAL GUIDELINES/STANDARDS

P. RICHARD JEANNERET AND SHELDON ZEDECK

INTRODUCTION[1]

Three primary sources of authoritative information and guidance that can be relied upon in the development, validation, and implementation of an employment selection procedure are the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; *Standards*), the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003; *Principles*), and the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, Civil Service Commission, Department of Justice, & Department of Labor, 1978; *Uniform Guidelines*). The term *selection procedure* in this instance should be interpreted broadly to include any process or information used in personnel decision making. Selection procedures would include (but not be limited to) all forms and types of tests (e.g., cognitive, personality, work samples, and assessment centers), interviews, job performance appraisals, and measures of potential. These procedures may be administered, scored, and interpreted as paper-and-pencil or computer-based instruments and/or by individuals internal or external to the organization. This broad view is consistent with the interpretations expressed by the authoritative sources. The term "test" is often used in one of the sources. For the purposes of this chapter, a test is synonymous with a selection procedure.

A number of other guidelines, standards, and legal requirements exist both in the United States and in other countries around the world. Relevant standards and guidelines include (but are not limited to) the following:

- U.S. Department of Labor guide regarding testing and assessment (International Standards Organisation, 2011)
- International Standards Organisation standards for assessment delivery (ISO-10667–2, 2011)
- New guidelines for assessment center operations (International Taskforce on Assessment Center Operations, 2015)
- European Federation of Psychologists' Associations model for description and evaluation of tests (EFPA, 2013)
- Guidelines for test use and adaptation from the International Test Commission (2001, 2005)

Additionally, many countries have statutes, rules, and regulations governing employment practices that may explicitly include testing or incorporate assessment procedures under broader requirements governing all employment practices. Because of the length of their histories and breadth of applicability, this chapter will focus on the primary sources noted in the introduction. However, the interested reader, especially those practicing in international settings would be well advised to review additional resources that may apply in their specific situations.

Purpose and Chapter Flow

The central focus of this chapter is to describe the history and substance of each of the three primary sources, compare and contrast their technical content, and provide some guidance as to how they might be particularly useful to those directly associated with employment selection procedures. Each of these three sources will be discussed separately in chronological order defined by the date of initial publication. The discussion will begin with the purpose and brief history of each document. Then information will be presented that describes the content relevant to employment decision making. After describing each document, the three sources will undergo comparisons with indications of inconsistencies and how they might be resolved. Finally, suggestions are made as to what additions or changes would be appropriate given the current state of relevant research.

Application to Employment Selection Only

The *Standards* in particular and the *Principles* to a lesser extent have potential relevance to settings outside of employment selection. Such venues include forensic, academic, counseling, program evaluation, and publishing that involves psychological instruments and measurements. This chapter does not address these applications. The focus is strictly on organizational settings and employment-related selection decisions.

Importance of the Authorities

For the most part, the authorities are retrospective rather than prospective. By necessity they must rely on the state of knowledge in the fields of measurement and applied psychology. Reality, of course, is that knowledge changes as research in the field develops more information about the strategies and psychometrics of employment selection procedures. Therefore, the authoritative sources become outdated and either include guidance that is no longer relevant or do not offer guidance that is very important in current times. Nevertheless, there are several reasons why the three authoritative sources are valuable resources that can be relied upon by individuals associated with employment selection:

1. *The study of employment-related psychometrics has been taking place for about 100 years*. Accordingly, there is a body of knowledge that is stable, well researched, and directly relevant to understanding the measurement properties of employment-based selection procedures. Much of this knowledge, with varying degrees of specificity, is embedded in all three authorities with little, if any, contradiction. Consequently, the authoritative sources are able to provide accurate information about the state of the science, at least at the time they were written, which can support the proper development and use of an employment selection procedure.
2. *The three documents describe and discuss several specific concepts and terms associated with the psychometric qualities of a selection procedure*. Although not intended as teaching documents per se, they do frequently summarize bodies of research that are otherwise buried in textbooks and research journal articles.
3. *The current editions of the* Standards *and the* Principles *have undergone extensive professional peer review*. Although the initial preparations of the documents were accomplished by committees of experts in the field (the *Standards* jointly by three psychological, educational, and measurement organizations and the *Principles* by a committee of Society for Industrial and Organizational Psychology (SIOP) members), both documents were open for comment by the membership of the American Psychological Association (APA) and, especially in the case of the *Principles*, the document was subject to review by the entire membership of SIOP, a division of APA. The *Standards* and the *Principles* were adopted as policy by APA and hence have formal professional status. Accordingly, there were much greater levels of scrutiny and approval of the scientific content of the *Standards* and *Principles* than typically occurs for a textbook or journal article.
4. *The* Uniform Guidelines *was authored by the Equal Employment Opportunity Commission (EEOC), the Civil Service Commission (CSC), the Department of Labor (DoL), and the Department of Justice (DoJ)*. The

preparation of the *Uniform Guidelines* also relied upon input from individuals with expertise in psychological measurement, but others (e.g., attorneys) were influential in creating the document as well. Given this complement of authors, it is understandable that there was less psychometric content and greater emphasis on the documentation of validity evidence that would be satisfactory in a judicial proceeding. Interestingly, when the *Uniform Guidelines* was under development and when the U.S. House of Representatives was holding hearings on revisions to the *Uniform Guidelines*, the APA (Division 14) submitted information that was, for the most part, not incorporated into the final document. Subsequently, in 1985, an APA representative gave congressional testimony that psychologists disagreed with four technical issues as these topics were addressed in the *Uniform Guidelines*: (a) validity generalization, (b) utility analysis, (c) differential prediction, and (d) validity requirements and their documentation. Similarly, SIOP believed the *Uniform Guidelines* was incorrect with respect to requiring fairness studies, the definition of construct validity, and how validity generalization and utility analyses were considered (Camera, 1996). Nevertheless, the EEOC and the Office of Federal Contact Compliance Programs (OFCCP) currently rely on the *Uniform Guidelines* to determine whether or not a selection procedure is discriminatory.

5. *For those who are involved in the judicial process (particularly judges and lawyers), the authoritative sources are additional reference sources to case law and other judicial writings*. The three sources have been relied upon by experts in the fields of personnel, industrial, organizational, and measurement psychology when formulating opinions about selection procedures. In such instances, the authoritative sources have become benchmarks that help define sound professional practice in the employment setting. Unfortunately, the apparent use of the three sources is rather limited, as indicated by the judicial interviews in Chapter 15 of Landy (2005).

STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING

Brief History

The *Standards* has a history dating back more than 60 years. The first edition was titled *Technical Recommendations for Psychological Tests and Diagnostic Techniques* and was authored by a committee of APA members and published in 1954. A similar publication was prepared by a committee comprising members from the American Educational Research Association (AERA) and the National Council on Measurement Used in Education (NCMUE). The document was titled *Technical Recommendations for Achievement Tests* and was published in 1955 by the National Education Association.

In 1966 the two separate documents were revised and combined into a single document, the *Standards for Educational and Psychological Tests and Manuals*, authored by a committee representing the APA, AERA, and the National Council on Measurement in Education (NCME). These three organizations have continued to jointly publish revisions. In a revision completed by a subsequent joint committee in 1974, the document title was changed to *Standards for Educational and Psychological Tests*. The 1966 document delineated about 160 standards, and this number was increased to more than 225 standards in 1974. However, the number of standards declined to about 180 in 1985 after a revision and publication of the *Standards for Educational and Psychological Testing* (*Standards*). This title has remained with the subsequent 1999 revision.

In 1991, the APA began an initiative to revise the 1985 *Standards*. In 1993, a joint AERA, APA, and NCME committee was formed, and after six years of effort the final document was published. It incorporates 264 standards and was adopted as APA policy. The *Standards* is intended to be prescriptive but does not have any associated enforcement mechanisms. More so than with past versions, the 1999 *Standards* devoted considerable attention to fairness; testing individuals with disabilities; scales, norms, and score comparability; reliability; and the responsibilities of test users.

After six years of revision and review, the latest version of the *Standards* (2014) presents an up-to-date wealth of psychometric information and places expanded emphasis on three topics: fairness, new and emerging technology, and holding users (and especially those associated with high-stakes testing) accountable for proper test use. The 2014 edition contains 45 pages of new material that was not included in the 1999 edition of the *Standards*.

Application

The 2014 *Standards* is applicable to the entire domain of educational and psychological measurement. Because the *Standards* provides a comprehensive wealth of information on psychological measurement, it is not possible to adequately discuss all of the content in this chapter. So, this review will focus on those components of the *Standards* that are most applicable to psychometric issues in employment selection.

Purpose of the *Standards*

"The purpose of the *Standards* is to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses" (p. 1). It is further emphasized that the evaluation of a test or its application should rely heavily on professional judgment and that the *Standards* provides a set of references or benchmarks to support the evaluation process. Finally, the *Standards* is not intended to respond to public policy questions that are raised about testing; however, the psychometric information embedded in the *Standards* may be very useful to informing those involved in debates and decisions regarding testing from a public policy perspective. This relevance exists because the initial version of the *Standards* (1954) preceded and was, in part, foundational to the *Uniform Guidelines* and the *Principles*.

Validity Defined

A key term that will appear throughout this chapter is "validity" or one of its derivatives (e.g., validation process). The *Standards* has established the most current thinking regarding validity and provides a definition that should receive broad acceptance by all professionals concerned with the psychometrics of selection procedures.

According to the *Standards*: (p. 11)

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretation. It is the interpretation of test scores required by the proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way (e.g., both to describe a test taker's current level of the attribute being measured and to make a prediction about a future outcome), each intended interpretation must be validated.

Validity is a unitary concept and can be considered an argument based on scientific evidence that supports the intended interpretation of a selection procedure score (Binning & Barrett, 1989; Cronbach & Meehl, 1955; McDonald, 1999; Messick, 1980, 1989; Wainer & Braun, 1988). There are 25 specific standards regarding validity incorporated into the 2014 document.

Generally, if a test does not have evidence for its validity for a particular purpose, it also will not have utility. Utility is an estimate of the gain in productivity or other practical value that might be achieved by use of a selection procedure. Several measures are used to estimate utility, including increases in job proficiency, reduced accidents, reduction in turnover, training success, etc. (Cascio & Boudreau, 2011; Cronbach & Gleser, 1965; Hunter & Hunter, 1984; Naylor & Shine, 1965; Schmidt, Hunter, McKenzie, & Muldrow, 1979). Consequently, it would be an unusual situation for an organization to want to use a test that lacked validity and (therefore) utility. Furthermore, if a test lacks validity, it is possible that unintended consequences may result from its use. Thus, reliance on test scores that are not valid will not yield results intended by the selection process and may yield outcomes that are detrimental to the organization.

Application to Selection Decision Making

Of the three authoritative sources, the *Standards* offers the greatest level of detail regarding psychometric properties and use of selection procedures. However, all standards are not necessarily equally important in a given situation, and no attempt is made to categorize some standards as "primary" and others as "secondary," as occurred in earlier versions. An entire chapter that incorporates 20 standards is focused on fairness; another chapter devotes 12 standards to the rights and responsibilities of test takers; and still another chapter is focused on individual psychological assessment (18 standards), whereby tests have been categorized into six groups: cognitive and neuropsychological tests; problem behavior measures; family and couples tests; social and adaptive behavior tests; personality measures; and vocational tests. This level of description is at times less precise in the *Principles* and *Uniform Guidelines*, particularly with respect to testing and assessment in the employment domain.

Cautions Offered by the *Standards*

The *Standards* (p. 7) sets forth five cautions that are intended to prevent misinterpretations:

1. Evaluation of a selection procedure is not just a matter of checking-off (or not) one standard after another to determine compliance. Rather the evaluation process must consider (a) professional judgment, (b) satisfaction of the intent of a relevant standard, (c) alternate selection procedures that are readily available, (d) feasibility of complying with the standard given past experience and research knowledge.; and (e) applicable laws and regulations (Note: this edition is the first time such a basis for acceptability has been expressed in the *Standards*.)
2. The Standards offers guidance to the expert in a legal proceeding, but professional judgment determines the relevance of a standard to the situation.
3. Blanket statements about conformance with the Standards should not be made without supporting evidence. Otherwise, care should be exercised in any assertions about compliance with the Standards.
4. Research is ongoing and knowledge in the field will continue to change. Accordingly, the Standards will be revised over time and the use of older Standards may be a disservice to test users and takers.
5. The Standards is not intended to mandate use of specific methodologies. The use of a "generally accepted equivalent" is always understood with regard to any method provided in the Standards.

Sources of Validity Evidence

There are multiple ways in which validity evidence might be assembled for a selection procedure, and no one method is necessarily superior to another. Rather, the validation strategy should be consistent with the nature and intended use of the selection procedure.

The *Standards* (pp. 13–21) describes five validation strategies or sources of validity evidence:

Content
Response processes
Internal structure
Relations to other variables
Consequences of testing

Comprehensive information on validation evidence and strategies may be found in Chapters 2 and 3 of this Handbook.

Convergent and Discriminant Validity

When selection procedure scores and other measures of the same or similar constructs are correlated, convergent validity evidence is demonstrated. When selection procedure scores are not correlated with other measures of purportedly different constructs, there is evidence of

discriminant validity (Campbell & Fiske, 1959; McDonald, 1999). Although both types of evidence are valuable in evaluating tests, convergent validity has been the more frequently studied. For example, in the typical criterion-related validity study, the relationship between a cognitive selection procedure and a measure of job performance is purportedly concerned with the same or very similar constructs (i.e., convergent validity). However, if a selection procedure comprised a cognitive measure and a test of interpersonal skills, and there were two job performance indices (decision making and teamwork), a lack of relationship (or low relationship) between the cognitive measure and teamwork (or a low correlation between the interpersonal skills test and decision making) would provide discriminant evidence.

Validity Generalization

An issue that arose early in research on selection measures was whether or not validity evidence obtained in one situation can be generalized to a new situation without further study of the validity of that procedure in the new setting. When criterion-related validity evidence has been accumulated for a selection procedure, meta-analysis has provided a useful statistical method for studying this generalization question. There are numerous methodological and statistical issues associated with meta-analytic studies, and these matters are too lengthy to be addressed here. The interested reader is referred to Cooper (2010) or Hunter and Schmidt (2004).

Integrating Validity Evidence

A comprehensive and sound validity argument is made by assembling the available evidence indicating that interpretations of scores from a well-developed selection procedure can accurately predict the criterion of interest. Although the various sources of validity evidence discussed above are directly relevant, there are many other valuable information sources, including information obtained from prior research; reliability indices; information on scoring, scaling, norming, and equating data; standard settings (e.g., cut scores); and fairness information. All of these information sources, when available, contribute to the final validity argument and decision regarding the use of a selection procedure (Barrett, Phillips, & Alexander, 1981; Bemis, 1968).

Validity Standards

There are 25 specific standards presented in the validity chapter of the *Standards*. Although all 25 standards are important, certain themes are particularly relevant in the context of employment selection. A brief summary of these themes follows:

- The rationale and intended interpretation of selection procedure scores should be stated at the outset of a validity study. When new interpretations or intended uses are contemplated, they should be supported by new validity evidence.
- Descriptions of individuals participating in validation studies should be as detailed as is practical. If subject matter experts (SMEs) are used, their qualifications and the procedures they followed in developing validation evidence should be documented.
- When criterion-related validity studies are completed, information about the quality and relevance of the criterion should be reported.
- When several variables are predicting a criterion, multiple regressions should be used to evaluate increments in the predictive accuracy achieved by each variable. Results from the analyses of multiple variables should be verified by cross-validation whenever feasible.
- If statistical adjustments (e.g., the correction of correlations for restriction in range) are made, the unadjusted and adjusted correlations and the procedures followed in making the adjustments should be documented.

- If meta-analyses are relied upon as criterion-related validity evidence, the comparability between the meta-analytic variables (predictors and criteria) and the specific situation of interest should be determined to support the applicability of the meta-analytic findings to the local setting. All assumptions and clearly described procedures for conducting the meta-analytic study should be reported.
- If effect size indices are used to make inferences beyond the validation sample, indicators of the amount of uncertainty regarding those indices (e.g., confidence intervals, standard errors, or significance tests) should be reported.

Reliability and Measurement Errors

Part I, Chapter 2 of the *Standards* describes reliability and errors of measurement and sets forth 20 standards related to the topic. The chapter is concerned with understanding the degree to which a selection procedure score is free from error. To the extent that a score is unreliable, it is due to errors of measurement that are usually assumed to be unpredictable and random in occurrence. There are two sources of error: (1) within individuals subject to the selection procedure and (2) conditions external to the individuals, such as the testing environment or mistakes in scoring the selection procedure.

Reliability is an index indicating the degree to which selection procedure scores are measured consistently across one or more sources of error such as time, test forms, or administrative settings. Reliability has an impact on validity in that to the extent the selection procedure is not reliable it will be more difficult to make accurate predictions from the selection procedure scores. Excellent treatments of reliability may be found in McDonald (1999), Nunnally and Bernstein (1994), Pedhazur and Schmelkin (1991), Putka and Sackett (2010), Traub (1994), and Chapter 1 of this Handbook.

The reliability chapter of the *Standards* develops many of the basic concepts embedded in psychometric theory. It is important to note that no single index of reliability measures all of the variables that influence the accuracy of measurement. The two major theoretical positions regarding the meaning of reliability are classical reliability theory and generalizability theory. What is important is that the method used to determine reliability be appropriate to the data and setting at hand and that all procedures be clearly reported. Furthermore, various reliability indices (e.g., test-retest, internal consistency) are not equivalent and should not be interpreted as being interchangeable; accordingly, one should not state that the "reliability of test X is . . .", but rather should state "the test-retest reliability of test X is . . .". Finally, the reliability of selection procedure scoring by examiners does not imply high candidate consistency in responding to one item versus the next item that is embedded in a selection procedure. In other words, just because the scoring of a test is reliable does not mean that the test itself is reliable.

Standards for Employment and Credentialing Tests

Chapter 11 of the *Standards* describes testing used for employment, licensure, and certification. In the employment setting, tests are most frequently used for selection, placement, and promotion. Sixteen standards are set forth in Chapter 11. They address the collection and interpretation of validity evidence, the use of selection procedure scores, and the importance of reliability information regarding selection procedure scores. The chapter's introduction emphasizes that the contents of many other chapters in the *Standards* also are relevant to employment testing. One point of emphasis in Chapter 11 is the influence of context on the use of a selection procedure. Ten contextual features are identified, which by their labels are self-explanatory (see *Standards*, pp. 170–171):

- Internal versus external candidate pool
- Trained versus untrained candidates
- Short-term versus long-term focus

- Screening in versus screening out
- Mechanical versus judgmental decision making (when interpreting test scores)
- Ongoing versus one-time use of a test
- Fixed applicant pool versus continuous flow
- Small versus large sample size
- Application to a new job
- Size of applicant pool relative to the number of job openings (selection ratio)

The *Standards* indicates that the validation process in employment settings is usually grounded in two sources of validity evidence: relations to other variables and content. One or both types of evidence can be used to evaluate how well a selection procedure (predictor) predicts or is directly linked to a relevant outcome (criterion). Furthermore, the *Standards* describe limited situations (e.g., small sample sizes, a new job without incumbents) in which validity evidence might be established on the basis of generalizability to include transporting validity using job analysis or statistical analyses across validation studies that encompassed similar jobs (e.g., meta analysis). Importantly, the *Standards* assert that there is no methodological preference or more correct method of establishing validity; rather, the selection situation and professional judgment should be the determiners of what source(s) of evidence are appropriate.

Evaluating Validity Evidence

Perfect prediction does not occur, and the evaluation of validity evidence is often completed on a comparative basis (e.g., how an observed validity coefficient compares to coefficients reported in the literature for the same or similar constructs). Consideration may be given to available and valid alternative selection procedures, utility, concerns about applicant reactions, statutory or regulatory requirements, fairness, strategies to achieve workforce diversity, and organizational values. Any or all of these types of considerations could influence the final conclusions drawn about the validity evidence as well as the implementation of the selection procedure.

Professional and Occupational Credentialing

In Chapter 11, the *Standards* also address the specific instance of credentialing or licensing procedures that are intended to confirm that individuals (e.g., medical doctors or nuclear power plant operators) possess relevant knowledge or skills to the degree that they can safely and/or effectively perform certain important occupational activities. Credentialing or licensing procedures are intended to be strict to provide the public as well as governmental and regulatory agencies with sound information regarding the capabilities of practitioners. The procedures are designed to have a gate-keeping role and often include written examinations as well as other specific qualifications (e.g., education or supervised experience). Content validity evidence is usually obtained to support the use of the credentialing procedures, because criterion information is generally not available. Establishing a passing score is a critical component of the validation process and is usually determined by SMEs, although empirical methods exist if the relevant data are available. Arbitrary passing scores, such as 70% correct, typically are not useful. They are unlikely to have any relevance to the underlying test psychometrics, and they may not define a level of credentialing procedure success equivalent to acceptable job performance. Thus, they provide no assurance of protection from harm to the public or of fairness to test takers. Finally, issues regarding fairness and accessibility are important and must be evaluated as to test scoring and accommodation, while also considering critical job functions and public interest.

Review of the Standards in Chapter 11

The first four standards are general in nature and apply to both workplace testing and credentialing; the next eight standards apply to workplace testing; the last four standards apply to credentialing. A brief discussion of these standards follows:

- The objective of the employment selection procedure should be set forth, and an indication of how well that objective has been met should be determined.
- Decisions regarding the conduct of validation studies should take into consideration prior relevant research, technical feasibility, and the conditions that could influence prior and contemplated validation efforts.
- When used, the fidelity of the criterion (which could be important work behaviors, work output, or job-relevant training) should be documented.
- Inference about the content validity of a selection procedure for use in a new situation requires that

> critical job content factors be substantially the same (e.g., as determined by a job analysis), and that the reading level of the test material not exceed that appropriate for the new job. In addition, the original meaning of the test materials should not be substantially changed in the new situation.
>
> (*Standards*, p. 181)

- When multiple sources of information are available to decision makers regarding an employment process, the use of each informational component should be supported by validity evidence. Furthermore, the role played by each component as it is integrated into a final decision preferably should be explained. In credentialing situations, the rules and procedures followed when combining scores from multiple information sources should be made available to candidates.
- Cut scores for credentialing tests should be determined on the basis of the skill or knowledge level necessary for acceptable job performance and not on the basis of the number or proportion of candidates passing.

Fairness

Fairness is addressed in Chapter 3 of the *Standards*, where it is described as a fundamental validity issue for all types of measurement, including that of workplace testing. While there is no single technical meaning for fairness, a fair test may be described as one that minimizes the construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some test takers. Fairness must be addressed during both test development and use for individuals from specific subgroups. These subgroups are identified by various characteristics, including disabilities, race, ethnicity, gender, age, culture, language, and socioeconomic status.

The *Standards* asserts that measurement bias is the central threat to fairness, but consideration is also placed on accessibility and universal test design. With these concerns in mind, there are four general aspects of fairness:

1. Equitable treatment of all test takers during the entire testing process
2. Lack of measurement bias
3. Full access to the construct being assessed (e.g., an individual with impaired vision might not be able to read a standard version of a personality test).
4. Validity of individual test score interpretations for their intended use

General threats to fair and valid interpretations of test scores include test content that produces construct-irrelevant variance, test context, test item responses, and opportunity to learn the content and skills measured by the test. Proper test design and adaptations help minimize these threats.

In employment testing, the issue of fairness is typically addressed by statistically examining test results for evidence of bias. It is not simply a matter of whether or not test score averages differ by subgroups but whether or not there are differences in test score predictions by subgroup. Under the most widely used model for analyzing test fairness (Bartlett, Bobko, & Mosier, 1978; Cleary, 1968), if the predictions are equivalent (i.e., no difference in the slopes or intercepts), then there is no bias. It should be noted that a number of concerns have been raised about fairness analyses using moderated regression models, especially with respect to the availability of adequate power in the analyses to detect bias should it actually exist (Aguinis & Stone-Romero, 1997). Another statistical perspective is that of differential item functioning (DIF). In this instance if there is bias, candidates of equal ability differ in their responses to a specific item according to their group membership. Unfortunately, the underlying reason for DIF, when it has been observed, has not been apparent; one group often performs better than another on some items for no explainable reason associated with item content. Use of sensitivity review panels that comprise individuals representative of the subgroups of interest has been one mechanism intended to prevent item content being relevant for one group but not another. Members of such review panels are expected to flag items that will be potentially unfair to a subgroup. However, there is not much research evidence indicating that sensitivity review panels find a great deal to alter in test item content for well-constructed tests.

Selection Procedure Development and Administration

Chapters 4–7 in the *Standards* are concerned with the development, implementation, and documentation of selection procedures. The discussions are quite technical in nature and will not be reviewed in this chapter. However, a couple of topics of particular relevance to employment selection will be mentioned:

- A cut score is used to partition candidates into two groups: one passing or successful and the other not passing or not successful. There is no single or best method for setting a cut score. Furthermore, because selection procedures are not perfect, there will always be errors—some candidates will pass who do not truly have adequate skills (false positives) and some will fail when in fact they do have adequate skills (false negatives). Changing a cut score to correct for one concern will usually increase the occurrence of the other. Thus, professional judgment always must play a significant role when setting a cut score.
- Normative data should be described in terms of demographics, sampling procedures, descriptive statistics, and the precision of the norms.
- The psychometric characteristics of different forms of the same test should be documented, and the rationale for any claim of equivalency in using test scores from different test forms must be reported.
- If the test developer permits different conditions of administration from one test taker or group to another, then a rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.
- Standardization in the administration procedures is extremely important, and all instructions and procedures must be carefully followed.
- The use of computers and the Internet for test administration and scoring result in special cautions. Training may be required to reduce construct-irrelevant variance; explanations and practice may be needed to manage test-specific details such as the test's interface; and managing the testing environment to avoid light reflections on the computer screen that interfere with display legibility may be necessary.
- Technology and the Internet have made it possible to administer tests in which the administration conditions may not be strictly controlled or monitored. Those who allow lack of standardization are responsible for providing evidence that lack of standardization will not affect test-taker performance or the quality and comparability of scores produced.
- Selection procedures and results (including individual scores) should be treated as confidential and kept in a secure manner.
- Documentation for a selection procedure typically includes information about intended purpose; prior research evidence; the development process; technical information regarding validity, reliability, fairness, score interpretation, scaling, or norming, if relevant; administration; and appropriate uses of the results (e.g., pass/fail).

Rights and Responsibilities

Two chapters in the *Standards* (Chapters 8–9) discuss test user and test taker rights and responsibilities. The standards set forth in these chapters are concerned with policy and administrative issues. Generally, these matters become more relevant in specialized circumstances (e.g., an applicant with a verified disability who needs an accommodation for a selection procedure). Professional judgment is typically required because of the individualized nature of the conditions.

Summary

The 2014 *Standards* reflects the state of the science and much of the most current professional knowledge available regarding psychological testing. As in the past, the *Standards* no doubt will be revised in the future. Nevertheless, the *Standards* is extremely informative about current professional thinking and scientific research regarding requirements associated with the development and application of a selection procedure in employment settings. The document has been published to promote the professionally sound and ethical use of selection procedures and to provide a set of standards that can be the basis for developing and implementing a new selection procedure, or for evaluating the quality of an existing selection procedure and practice.

PRINCIPLES FOR THE VALIDATION AND USE OF PERSONNEL SELECTION PROCEDURES

Brief History

The first edition of the *Principles* was published in 1975 in response to the growing concern about the need for professional standards for validation research. Furthermore, because early versions of what became the *Uniform Guidelines* were being prepared by various governmental organizations, Division 14 representatives wanted to set forth the perspective of industrial and organizational (I-O) psychology, particularly with regard to validation studies. The second edition was published five years later and, for the first and only time, cited specific references regarding equal employment opportunity and associated litigation. Because of continuing changes in employment case law, subsequent editions have not attempted to stay current with them. Furthermore, it has not been the purpose of the *Principles* to interpret these cases in terms of the science of I-O psychology

In 1987 the third edition of the *Principles* was published by SIOP. This edition consisted of 36 pages of text and 64 citations to published research to support the various principles contained in the document. An appended glossary defined 76 terms used in the *Principles*.

The fourth edition of the *Principles* was published by SIOP and adopted as policy by the APA in 2003. This edition consists of 45 pages of text and an appended glossary of 126 terms. There are 65 research literature citations that support the scientific findings and professional practices that underlie the principles for conducting validation research and using selection procedures in the employment setting. The increase in glossary terms reflects some of the more recent scientific findings and thinking related to such topics as generalized evidence of validity, work analysis, internal structure validity evidence, models of reliability and fairness, and test development and implementation.

Purpose of the *Principles*

The *Principles* establishes ideals and sets forth expectations for the validation process and the professional administration of selection procedures. The document also can inform those responsible for authorizing the implementation of a validation study and/or selection procedure. The *Principles* does not attempt to interpret federal, state, or local statues, regulations, or

case law related to matters of employment discrimination. However, the *Principles* expects to inform decision making in employment administration and litigation and offers technical and professional guidance that can help others (e.g., human resource professionals, judges, and lawyers) understand and reach conclusions about the validation and use of employment selection processes.

Principles Versus the *Standards*

The *Principles* was revised in 2003 with the full understanding that the document would be consistent with the then-extant *Standards*, especially with regard to the psychometric topics of validity, reliability, and bias. Both documents are grounded in research and express a consensus of professional opinion regarding knowledge and practice in personnel selection. However, there are also some important differences between the two documents.

First, unlike the *Standards*, the *Principles* does not enumerate a list of specific principles in the same manner as the *Standards* sets forth 240 standards. Consequently, the *Principles* is more aspirational and facilitative in content, whereas the *Standards* is more directive in nature. That said, the *Standards* states that it is not a set of legal requirements nor a substitute for legal advice (p. 1).

Second, the *Standards* is much broader than the *Principles* with respect to psychological measurement. For example, although many of the concepts expressed in the *Principles* could be relevant to the field of educational testing, the *Standards* directly addresses the topic. The same is true for such topics as testing in program evaluation and public policy.

Third, the *Standards* is more concerned with the rights and responsibilities of test takers, whereas the *Principles* focuses more on the responsibilities of selection procedure developers and users. This focus reflects the fact that the *Principles* places most of the responsibility for proper selection processes on the employer rather than the candidate, whereas the *Standards* considers a much wider group of test takers to include students, patients, counselees, and applicants.

Finally, the *Principles* provides more guidance on how to plan a validation effort and collect validity evidence within the context of an employment setting. Consequently, there is more discussion of such topics as (a) feasibility of a validation study; (b) strategies for collecting information about the work and work requirements, as well as about job applicants or incumbents and their capabilities; (c) analyzing data, including such topics as multiple-hurdles versus compensatory models, cutoff scores, rank orders, and banding; and (d) information to be included in an administrative guide for selection procedure users.

Application to Litigation

The *Principles* offers relevant information and guidance regarding personnel selection procedures that might be the subject of litigation. Although the document is not written in absolute terms, it provides a wealth of information that defines best practices in the validation and implementation processes required to use selection procedures properly. When examining the qualities of a validation study or the implementation of a selection procedure, a decision maker in litigation proceedings might find that one or more expectations set forth in the *Principles* were not met and ask why. Absent sound and logical explanations, the unexplained issues could be strong indicators that the procedures being scrutinized were not established in accord with accepted professional expectations.

Analysis of Work

Given that the *Principles* is focused on selection procedures in the employment setting, there is a particular emphasis on the analysis of work. Such an analysis establishes the foundation for

collecting validity evidence. More specifically, information from the analysis of work defines relevant worker requirements and determines the KSAOs needed by a worker to perform successfully in a work setting. Second, the work analysis defines the criterion measures that, when appropriate for the validation strategy being used, indicate when employees have successfully accomplished relevant work objectives and organizational goals.

Historically, the analysis of work was labeled "job analysis," and that term is still frequently used. The *Principles* expanded the term to "analysis of work" to give clear recognition to the realization that the concept of a traditional job is changing. Furthermore, the "analysis" should incorporate the collection of data about the workers, the organization, and the work environment, as well as the specific job or some future job if that is relevant to the study. As implied by the various permutations that might be considered, no one preferred method or universal approach is appropriate for completing an analysis of work.

The *Principles* encourages the development of a strategy and a sampling plan to guide an analysis of work. Furthermore, the analysis should be conducted at a level of detail consistent with the intended use and availability of the work information. Any method used and outcomes obtained should be well documented in a written report.

Validation

The *Principles* adopts the same definition of validity as given in the *Standards*. Validity is a unitary concept, and different sources of evidence can contribute to the degree to which there is scientific support for the interpretation of selection procedure scores for their proposed purpose. If a selection procedure is found to yield valid interpretations, then it can be said to be job-related. The *Principles* recognizes the five sources of evidence discussed in the *Standards*. However, the *Principles* places more emphasis on the two sources of evidence most frequently relied upon when studying validity in the employment context—criterion-related and content validity.

Criterion-Related Validity Evidence

The *Principles* emphasizes several issues related to obtaining criterion-related validity evidence:

- *Feasibility*: Is it technically feasible to conduct the study in terms of measures, sample sizes, and other factors that might unduly influence the outcomes?
- *Design*: Is a concurrent or predictive design most appropriate?
- *Criterion*: Is the criterion relevant, sufficient, uncontaminated, and reliable?
- *Construct equivalence*: Is the predictor measuring the same construct underlying the criterion?
- *Predictor*: Is the selection procedure theoretically sound, uncontaminated, and reliable?
- *Participants*: Is the sample of individuals in the study representative of the applicants and/or incumbents, and will it support the generalization of results?
- *Analyses*: Are the analytical methods to be used appropriate for the data collected?
- *Strength of relationships*: What effect size and statistical significance or confidence intervals were hypothesized and observed?
- *Adjustments*: What adjustments are necessary to correct observed validity relationships to avoid underestimating the predictor-criterion relationship? It may be appropriate to adjust for restriction in range and unreliability in the criterion.
- *Combining predictors/criteria*: How are predictor and/or criteria scores weighted if combined?
- *Cross-validation*: Should the estimates of validity be cross-validated to avoid capitalization on chance? Typically, when regression analyses are used and the sample is small, adjustments should be made using a shrinkage formula or a cross-validation design.
- *Interpretation*: Are the results observed consistent with theory and past research findings?
- *Administrative procedures*: Are adequate guidelines established for administering and scoring the selection procedure that will maintain the integrity of the validity evidence?

Content Validity Evidence

The *Principles* also emphasizes several issues related to obtaining content validity evidence:

- *Feasibility*: Are there job determinant conditions (e.g., is the work stable or constantly changing?), worker-related variables (e.g., are past experiences relevant for the current work?), or contextual matters (e.g., are the work conditions extremely different from the testing environment?) that might influence the outcome of the validity study? If so, are they sufficiently controlled so as to not contaminate the study?
- *Design*: Has an adequate sample of important work behaviors and/or worker KSAOs been obtained and analyzed?
- *Content domain*: Has the work content domain been accurately and thoroughly defined and linked to the selection procedure?
- *Selection procedure*: Does the selection procedure adequately represent the work content domain? The fidelity of this relationship is the basis for the validity inference.
- *Sampling*: Is there a sound rationale for the sampling of the work content domain?
- *Specificity*: Has the level of specificity necessary in the work analysis and selection procedure been described in advance?
- *Administrative procedures*: Are adequate guidelines established for administering and scoring the selection procedure that will maintain the integrity of the validity evidence?

The *Principles* also recognizes internal structure validity evidence. The *Principles* points out that evidence based on the structure of a selection procedure is not sufficient alone to establish the validity of the procedure for predicting future work performance or other work-related behaviors (e.g., attendance, turnover). However, consideration of the internal structure can be very helpful during the design of a selection procedure.

Generalizing Validity Evidence

The *Principles* provides considerably more detail regarding the generalization of validity evidence in comparison to the *Standards*. There are at least three strategies for generalizing evidence, known as transportability, job component validity, and meta-analysis. The *Standards* indicates these strategies are especially relevant when a job is new, sample sizes are small, or if research data are available to conduct meta-analyses.

Transportability

This strategy refers to relying on existing validity evidence to support the use of a selection procedure in a very similar but new situation. The important consideration underlying the transport argument is work/job comparability in terms of content and requirements. Also, similarity in work context and candidate groups may be relevant to documenting the transport argument (Gibson & Caplinger, 2007).

Synthetic/Job Component Validity

This type of generalization relies on the demonstrated validity of selection procedure scores for one or more domains or components of work. The work domains or components may occur within a job or across different jobs. If a sound relationship between a selection procedure and a work component has been established for one or more jobs, then the validity of the procedure can be generalized to another job that has a comparable component. As in the transportability argument, the comparability of work content on the basis of comprehensive information is essential to the synthetic/job component validity process (Hoffman, Rashkovsky, & D'Egidio, 2007; Johnson, 2007).

Meta-analysis

The information on meta-analysis in the *Standards* and *Principles* is very similar. In the *Principles*, meta-analysis is acknowledged as a statistical technique that serves as the foundation for validity generalization. Both documents point out that meta-analytic findings may be useful, but not sufficient, to reach a conclusion about the use of a selection procedure in a specific situation. Rather, a local validation study may be more appropriate. Both sources also emphasize that professional judgment is necessary to evaluate the quality of the meta-analytic findings and their relevance to the specific situation of interest. The general conclusion in the *Principles* is that meta-analytic findings for cognitive tests indicate that much of the difference in validity coefficients found from one study to the next can be attributed to statistical artifacts and sampling error (Callendar & Osburn, 1981; Hartigan & Wigdor, 1989; Hunter & Hunter, 1984). Similar but not conclusive evidence is occurring for noncognitive measures (Barrick & Mount, 1991; Barrick, Mount, & Judge, 2001; Hurtz & Donovan, 2000), but the strength of validity may be less for noncognitive tests (Hogan, Davies, & Hogan, 2007; Morgeson et al., 2007).

The *Principles* discuss the appropriateness of the technique and its interpretation in specific situations. In general, reliance on meta-analytic results is most appropriate when the studies contributing to the meta-analysis focus on well-defined constructs. In such instances, the meta-analytic findings reflect the degree to which the measures of the constructs are measuring the same construct. In contrast, when the studies in the meta-analysis focus on methods (e.g., the interview) instead of constructs, several interpretational difficulties arise. Because interviews may measure different constructs, it is difficult to generalize about the general method of the interview unless the features of the interview method "are clearly understood, if the content of the procedures and meaning of the scores are relevant for the intended purpose, and if generalization is limited to other applications of the method that include those features" (*Principles*, p. 30). Generalizing from a meta-analysis of "the" interview method to a new interview method measuring different constructs or to a new interview that addresses a new situation is problematic when constructs do not serve as the foundation of the analysis.

Fairness and Bias

As presented in the *Standards*, the topics of fairness and bias are also prominent in the *Principles*. The *Principles* endorses the definitions and positions taken by the *Standards*.

Predictive Bias

An alternative term to "predictive bias" is differential prediction. Regardless of the terminology, the key is that bias occurs if consistent, nonzero errors of prediction are made for individuals in a particular subgroup that are greater than those for another subgroup. Multiple regression techniques are typically used to assess predictive bias, which is indicated if slope and/or intercept differences are observed in the model. Research on cognitive ability measures has typically supported the conclusion that there is no predictive bias for African American or Hispanic groups relative to Whites, and when predictive differences are observed, they usually indicate overprediction of the performance of the minority group. It is also important to understand that there can be mean score differences on a selection procedure for minority versus majority subgroups that do not result from predictive bias.

Measurement Bias

This form of bias is associated with one or more irrelevant sources of variance contaminating a predictor or criterion measure. There are not well-established approaches to assessing

measurement bias, as is the case for predictive bias, though differential item functioning (DIF) and item sensitivity analyses are suggested as options in the *Principles*, but considerable caution in the value of such analyses is also mentioned. As noted by Sackett, Schmitt, Ellingson, and Kabin (2001), the research results indicate that item effect is often very small, and there is no consistent pattern of items that favor one group of individuals relative to another group. Additionally, the rubric of item sensitivity is very broad and includes concerns about item acceptability and perception, even if no measurement bias has resulted.

Operational Considerations

Almost half of the *Principles* is devoted to operational considerations. The issues discussed are related to initiating and designing validation efforts; analysis of work; selecting predictors, a validation strategy, and criterion measures; data collection and analyses; implementation; recommendations and reports (technical and administrative); and other circumstances that may influence the validation effort (e.g., organizational changes; candidates with disabilities; and responsibilities of selection procedure developers, researchers, and users). There are a few topics discussed in the operational considerations section of the *Principles* deserving particular attention in the development and implementation of an employment selection procedure that are discussed in the following subsections.

Combining Selection Procedures

If selection procedure scores are combined in some manner, the validity of the inferences derived from the composite is of great importance. In other words, it is not sufficient to simply report the validity index for each procedure as a stand-alone predictor; rather, a validity index should be reported for the combined selection procedure score that is used for decision making.

Multiple-Hurdle Versus Compensatory Models

A multiple-hurdle model involves making decisions in a sequence (e.g., applicants who pass one selection procedure move on for further consideration in a following procedure, and those who pass the second procedure move on to a third selection procedure, etc.). In contrast, a compensatory model involves all applicants completing all selection procedures, and their final hiring result is based on a weighted combination of their scores on the components of the procedure. The *Principles* provides no definitive guidance as to which model is more appropriate; rather, each situation must be evaluated on its own merits. Combining scores into a compensatory sum may affect the overall reliability and validity of the process. When multiple predictors (with different reliabilities and validities) are combined into a single weighted composite score, the result produces a single-stage selection decision. How each predictor is weighted will influence the psychometric characteristics of the compensatory selection procedure score, and the final reliability/validity indices may be lower than if used in their individual capacities in a multi-staged selection process (Sackett & Roth, 1996).

Cutoff Scores Versus Rank Order

The *Principles* concludes that a cutoff score may be set as high or low as needed relative to the requirements of the using organization given that a selection procedure demonstrates linearity or monotonicity across the range of predictions (i.e., it is valid). For cognitive predictors, the linear relationship is typically found using a criterion-related validity model and is assumed with

a content validity process. Under these circumstances, using a rank-order (top-down) process will maximize expected performance on the criterion. Whether this same premise holds true for noncognitive measures has not been determined.

In a rank-order model, the score of the last person selected becomes the lower bound cutoff score. A cutoff score set otherwise usually defines the score on the selection procedure below which applicants are rejected. Professional judgments that consider KSAOs required, expectancy of success versus failure, the cost-benefit ratio, consequences of failure, the number of openings, the selection ratio, and organizational diversity objectives are important to setting a cutoff score. In the case of organizational diversity objectives, using lower cutoff scores could result in higher proportions of minority candidates passing some valid initial hurdle, with the expectation that subsequent hurdles might have less adverse impact. In such instances, cutoff scores may be set even lower with the realization that there will be a corresponding reduction in job performance and selection procedure utility, but that the tradeoffs regarding hiring a diverse workforce may be sufficient to overcome such reductions.

Utility

Gains in productivity, reductions in outcomes (e.g., accidents, absenteeism), or comparisons among alternate selection procedures can be estimated by utility computations. Typically, several assumptions must be made with considerable uncertainty to satisfy the computational requirements of the utility models. Thus, caution should be observed in relying upon such utility estimates.

Bands

A band exists when a range of selection procedure scores is established that considers all candidates within the range to be effectively equivalent. Banding may necessarily lower expected criterion outcomes and selection utility when compared to top-down selection, but these consequences may be balanced by increased administrative ease and the possibility of increased workforce diversity.

Technical Validation Report Requirements

Every validation study should be documented with a technical report that contains sufficient information to allow an independent researcher to replicate the study. Such a report should present all findings, conclusions, and recommendations. In particular, the technical report should give information regarding the research sample and the statistical analyses conducted, as well as recommendations on implementation and the interpretation of the selection procedure scores.

Summary

The *Principles* offers a comprehensive resource for use by decision makers when developing and implementing employment selection procedures. Because the *Principles* is focused specifically on employment settings, there is frequently more guidance offered on matters that arise in the development and use of selection procedures than will be found in the *Standards*. Nevertheless, the two documents are very compatible and not at all contradictory. The *Principles* has undergone substantial professional peer review and represents the official policy of the SIOP and APA. Currently, the *Principles* are being revised, with an expected delivery in 2017.

UNIFORM GUIDELINES ON EMPLOYEE SELECTION PROCEDURES

Brief History

When the U.S. Congress passed the Equal Employment Opportunity (EEO) Act of 1972, it created the Equal Opportunity Coordinating Council, which comprised the Directors/Secretaries of the Equal Employment Opportunity Commission (EEOC), the Civil Service Commission (CSC), the Civil Rights Commission (CRC), the Department of Justice (DoJ), and the Department of Labor (DoL). The Council was given the mandate to develop and implement policies, practices, and agreements that would be consistent across the agencies responsible for enforcing EEO legislation. Building on earlier guidelines promulgated by the EEOC and the Office of Federal Contract Compliance Programs (OFCCP), in 1977 the Council began developing the *Uniform Guidelines* document, which was adopted on August 25, 1978, by the EEOC, the CSC, the DoJ, and the DoL's OFCCP, with an effective date of September 25, 1978. On March 2, 1979, the EEOC, Office of Personnel Management (OPM), DoJ, DoL, and Department of Treasury published the Questions and Answers (the Q&As) to clarify and provide a common interpretation of the *Uniform Guidelines on Employee Selection Procedures*. The change in agencies adopting the Q&As was because OPM and, to some degree, the Office of Revenue Sharing of the Treasury Department had succeeded the CSC.

Although some psychologists participated in the development of the *Uniform Guidelines*, there was not consensus from the professional associations (e.g., SIOP, APA) that the document reflected the state of the scientific knowledge regarding the validation and use of employee selection procedures. Ad hoc committees of psychologists from SIOP and APA reviewed draft versions of the *Uniform Guidelines* and offered considerable input, but most of the suggestions were not incorporated (Camara, 1996). When Congress considered revising the *Uniform Guidelines* in 1985, the APA offered testimony that the document was deficient with respect to differential prediction, validity generalization, utility analysis, and validity requirements and documentation. SIOP concurred with the APA's concerns and further argued that the *Uniform Guidelines* was in error in defining construct validity and in determining the acceptable types of validity evidence. Congress declined to revise the *Uniform Guidelines* at that time, though subsequently additional Q&As were adopted regarding Internet testing.

Purpose

The *Uniform Guidelines* is intended to do the following:

Incorporate a single set of principles which are designed to assist employers, labor organizations, employment agencies, and licensing and certification boards to comply with requirements of Federal Law prohibiting employment practices which discriminate on grounds of race, color, religion, sex, and national origin. They are designed to provide a framework for determining the proper use of tests and other selection procedures. These guidelines do not require a user to conduct validity studies of selection procedures where no adverse impact results. However, all users are encouraged to use selection procedures which are valid, especially users operating under merit principles. (Section 1.B 29C.F.R.1607)

The Q&As was prepared "to interpret and clarify, but not to modify, the provisions of the *Uniform Guidelines*" (Introduction, Federal Register 43, 166, 11996–12009, March, 1979).

All subsequent references in this chapter to the *Uniform Guidelines* should be considered to include the Q&As.

Application and Limitations

The *Uniform Guidelines* applies to Title VII of the Civil Rights Act of 1964, Executive Order 11246 (establishing the OFCCP) regarding race, color, religion, sex, and national origin. They do

not apply to the Age Discrimination in Employment Act (ADEA) of 1967, nor to sections 501, 503, and 504 of the Rehabilitation Act of 1973, which prohibit discrimination on the basis of disability. Because the Americans with Disabilities Act (ADA) was not enacted until 1991, the *Uniform Guidelines* was not able to address this legislation and the protection it affords people with disabilities (though courts have applied the *Uniform Guidelines* to subsequent new laws such as ADEA and the Civil Rights Act of 1991). Generally, the *Uniform Guidelines* applies to most public and private-sector employers.

Selection Procedures/Employment Decisions

In general, the *Uniform Guidelines* defines selection procedures (Equal Employment Opportunity Commission, 1979) and employment decisions in a manner similar to the *Standards* and the *Principles*. Thus, processes related to hiring, promotion, retention, and certification are covered. These processes would include tests, assessment centers, interview protocols, scored applications, physical ability measures, work samples, and performance evaluations. Furthermore, the *Uniform Guidelines* applies to any intermediate process (e.g., having to complete a certification program to be eligible for a promotion) that leads to a covered employment decision. Two practices are exempt or are not considered selection procedures: recruitment (excluded to protect the affirmative recruitment of minorities and women) and bona fide seniority systems.

Discrimination/Adverse Impact

The *Uniform Guidelines* explicitly defines discrimination and introduces the term "adverse impact." In essence, discrimination occurs when a selection procedure results in unjustifiable adverse impact. Adverse impact occurs when the selection rate for a protected group is less than four-fifths (80%) of the rate for the group with the highest rate (typically the nonprotected group). To illustrate, if the passing rate for the majority group is 60%, and the passing rate for a protected group is 40%, then the ratio 40/60 yields 67%, which is less than 80%, and the *Uniform Guidelines* says that the enforcement agencies will view that as evidence of adverse impact. If, on the other hand, the passing rate of the protected group was 50%, the ratio becomes 50/60 yielding 83%, resulting in no adverse impact.

This "rule of thumb" is not intended as a legal definition and for good reason, because it is problematic from a couple of perspectives. First, it is highly influenced by sample size. For example, if there are 50 male and 50 female applicants and 20 open positions, the only way a selection process will not violate the 80% rule is to hire at least 9 females (9/50 = 18%) and no more than 11 males (a difference of 2), which does not violate the 80% rule in this case because the passing rate for the males is 22% (18/22 = 82%). Note that if the samples of males and females were each 500, then the same percentages of 22% and 18% hired would yield 110 males and 90 females hired; this difference of 20 would not be considered adverse impact.

Second, and perhaps most important, the 80% rule of thumb is not a statistical test; it is simply a ratio. The null hypothesis is not stated, and there is no estimate of the likelihood of any difference observed being because of chance. Accordingly, an alternative to the 80% rule is a statistical test of significant differences. Such hypothesis testing is accomplished using binomial or hypergeometric probability models. Typically, the .05 level of statistical significance under a two-tailed test (e.g., 1.96 standard deviation units) is considered the threshold of significance (both in the scientific literature and the courts, *Hazelwood School District v. United States*, 1977). Although the 80% value has no standing in the scientific literature, the .05 level of significance is well accepted in social sciences research as indicating statistical significance, but this test also has its practical limitation because statistical significance is also a function of sample size. A difference of 5 points between two groups would be statistically significant if the total sample were in the thousands but would not be statistically significant if the total sample was two digits (e.g., 30). Although the *Uniform Guidelines* recognizes the problems inherent in the rule of thumb

in Section 3D, where it states that statistical significance is impacted by "small numbers," it does not provide guidance as to what is the favored strategy—the 80% rule or statistical difference. Some practitioners have suggested that both analyses should be standard practice (Colosimo, 2010).

Fairness

This concept is introduced in the discussion of criterion-related validity (see Sec. 7.B [3] and Sec. 14.B [8]). The *Uniform Guidelines* requires that a fairness investigation of a selection procedure be conducted if technically feasible before applying validity evidence from one situation to a new situation. Furthermore, if adverse impact is observed and data from a criterion-related validation study are available, the user is expected to conduct a fairness analysis. Unfairness occurs when lower minority scores on a selection procedure are not reflected in lower scores on the criterion or index of job performance. As noted above, the *Standards* and *Principles* consider this a matter of predictive bias, and it is found when consistent nonzero errors of prediction occur for a protected subgroup, but not for other subgroups that are disproportionately selected. Moderated multiple regression is the most frequently used statistical method for examining predictive bias, which occurs if there are slope and/or intercept differences between subgroups. As previously mentioned, there is no consistent research evidence supporting predictive bias on cognitive tests for African Americans or Hispanics relative to Whites. Also, research directed at race differences on noncognitive tests suggests few to small differences (Cascio, Jacobs, & Silva, 2010; Hough & Oswald, 2008; Ployhart & Holtz, 2008; Ryan & Powers, 2012; Schmitt & Quinn, 2010; Schmitt, Keeney, Oswald, Pleskac, Billington, Sinha, & Zorzie, 2009).

Cutoff Scores

Cutoff scores are discussed first in the *Uniform Guidelines* as part of the general standards for validity studies (Sec. 5. H.) and then in the Technical standards section (Sec. 14. B. [6]). According to the *Uniform Guidelines*, "Where cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force" (Sec. 5. H.).

This definition seems to imply the need for professional judgment in setting a cutoff score, and such a stance is consistent with the *Standards* and the *Principles*.

Bottom Line

Another concept introduced by the *Uniform Guidelines* when trying to assess adverse impact or discrimination is the bottom-line approach. If there are multiple components to a selection procedure, then the final decision point is evaluated for adverse impact. According to the *Uniform Guidelines*, only if the adverse impact occurs at the bottom line must the individual components of a selection procedure be evaluated. However, this concept was struck down by the U.S. Supreme Court in *Connecticut v. Teal* (1982). Currently, it is typical for all components of a selection procedure to be evaluated in terms of adverse impact and validity if they can be examined individually (*Hazelwood School District v. United States*, 1977).

Alternative Selection Procedure

The *Uniform Guidelines* introduced the concept that if two or more selection procedures are available that serve the user's interest and have substantially equal validity "for a given purpose," then

the procedure demonstrating the lesser amount of adverse impact should be used. Although conceptually the alternative selection procedure is understandable, it is difficult to contend with in practice. There is no clear definition for "substantially equal valid." Although there may be alternatives, it is not necessarily easy to discern which of them might have lesser adverse impact in a given situation. The degree of adverse impact observed is very specific to the numbers and qualifications of applicants at a particular point in time; furthermore, it is not clear what constitutes "lesser adverse impact." Finally, many selection procedures are available, "which serve the user's legitimate interest in efficient and trustworthy workmanship" but still may not be feasible alternatives (see 3.B.). Examples of concerns affecting feasibility include faking or response distortions of personality and biodata inventories, costs of development and implementation, and the ability to assess very large numbers of applicants at the same time. It is important to consider carefully the purpose served by the selection procedure. It is one thing to substitute a less impactful mechanical aptitude test for one that adversely underselects women, but substituting a reading comprehension test (no matter how valid) for mechanical aptitude may not be appropriate depending on the job tasks and requirements.

Also of note is the general application of the "alternative selection procedure" section of the *Uniform Guidelines*, Section 3B. Whereas most of the attention in the literature and litigation has focused on alternative procedures, the *Uniform Guidelines* also considers "an investigation of . . . suitable alternative methods of using the selection procedure which have as little adverse impact as possible." Thus, application of a particular method in a given situation might be used as pass/fail instead of as top-down selection.

Job-Relatedness/Business Necessity

An employment selection procedure that has adverse impact may be justified in two ways: (a) showing that the procedure is job-related and (b) showing that the procedure is justified by business necessity. Job-relatedness is demonstrated by the validation process. Business necessity is demonstrated when a selection procedure is necessary for the safe and efficient operation of the business entity. Relevant statutes and regulations often define the business necessity argument (i.e., legislation regarding public safety job requirements), but other times information from the analysis of work will demonstrate the business necessity of a selection procedure.

Validity

The *Uniform Guidelines* sets forth what the enforcement agencies consider acceptable types of validity studies and identifies three types: criterion-related, content, and construct. The document notes that new validation strategies "will be evaluated as they become accepted by the psychological profession" (see 5.A.). The *Uniform Guidelines* also states that the validation provisions "are intended to be consistent with generally accepted professional standards . . . such as those described in the *Standards for Educational and Psychological Tests* . . . and standard textbooks and journals in the field of personnel selection" (see 5.C). Of course the *Standards* being referred to were published in 1974, and three major revisions were published in 1985, 1999, and 2014. The *Uniform Guidelines* makes no specific reference to the *Principles*, although the first edition was published in 1975. Consequently, it is easy to understand how the treatment of validity by the *Uniform Guidelines* is not particularly consistent with the state of the scientific knowledge as set forth in the current editions of the *Standards* and the *Principles*.

When introducing validity, the *Uniform Guidelines* offers several warnings or conditions:

- Do not select on the basis of knowledge, skills, and abilities (KSAs) that can be learned on the job during orientation.
- The degree of adverse impact should influence how a selection procedure is implemented, and evidence sufficient to justify a pass/fail strategy may be insufficient for rank order.

- A selection procedure can be designed for higher-level jobs if most employees can be expected to progress to those jobs in about five years.
- An employer can use a selection procedure if there is substantial validity evidence from other applications and if the employer has in progress, if technically feasible, a validity study that will be completed in a reasonable period of time, but reliance on such research, should it not demonstrate validity, will not protect an employer from enforcement actions.
- Validity studies should be reviewed for currency, particularly if alternative procedures with equal validity but less adverse impact may be available.
- There are no substitutes for validity evidence and no assumptions of validity based on general representation, promotional material, testimony, and the like.
- Employment agencies are subject to the guidelines in the same manner as employers.

Criterion-Related Validity

The *Uniform Guidelines*' position on criterion-related validity is very consistent with the information set forth in the *Standards* and *Principles*. Job analysis is important for decisions regarding grouping jobs together and selecting and developing criterion measures. An overall measure of job performance may be used as a criterion if justified by the job analysis; however, the *Principles* and *Standards* emphasize the need for construct equivalence for predictor and criterion measures. Typically, there are criteria with a greater degree of construct specificity developed from work analysis than from "overall performance." Success in training also can be used as a criterion. Concurrent and predictive designs are recognized, and emphasis is placed on the representativeness of the sample of individuals participating in the validity study, regardless of its design.

Criterion-related validity evidence should be examined using acceptable statistical procedures, and the *Uniform Guidelines* establishes the .05 level of statistical significance as the threshold for concluding that there is a relationship between a predictor and a criterion. Usually, the relationship is expressed as a correlation coefficient, which must be assessed in the particular situation: "There are no minimum correlation coefficients applicable to all employment situations" (see 14.B. [6]). Additionally, care must be taken to not overstate validity findings.

Content Validity

The technical standards for content validity studies begin by focusing on the appropriateness of such a study. A selection procedure must be a representative sample of the job content or purport to measure KSAs that are required for successful job performance. Selection procedures based on inferences about mental abilities or that purport to measure traits such as intelligence, common sense, or leadership cannot be supported only on the basis of content validity. Solid job analysis information that is representative of the jobs (and, when necessary, operationally defined) is critical to a content validity argument.

The *Uniform Guidelines* provides for the ranking of candidates assessed by a content-valid selection procedure, given that the procedure is measuring one or more capabilities that differentiate among levels of job performance. This is generally compatible with the guidance offered by the *Principles*, although the Q&As to the *Uniform Guidelines* gives more examples as to when it is, or is not, appropriate to use rank ordering.

Construct Validity

This form of validity is defined in Section 14.D (1) of the *Uniform Guidelines* as "a series of research studies, which include criterion-related and which may include content validity studies." In Section 14.D (1) and (3), it is stated that a "construct" is the intermediary between the selection procedure on the one hand and job performance on the other. A job analysis is required, and one or more constructs that are expected to influence successful performance

of important work behaviors should be identified and defined. To accomplish a construct validity study, it should be empirically demonstrated "that the selection procedure is validly related to the construct and that the construct is validly related to the performance of critical or important work behaviors" (14.D [3]). (This is the definition that drew the objections of the APA and SIOP.) In turn, a selection procedure is developed that will measure the constructs of interest. In a somewhat discouraging note for researchers, the *Guidelines* state that "The user should be aware that the effort to obtain sufficient empirical support for construct validity is both an extensive and arduous effort involving a series of research studies" (*Uniform Guidelines*, Section 14. D[1]).

Documentation Required

The *Uniform Guidelines* sets forth many documentation requirements for a validity study, and many of these requirements are labeled "essential." Generally speaking, the information expected as part of the documentation effort is very consistent with the material presented in each of the various sections of the *Uniform Guidelines*.

Utility

One term—"utility"—does not have a definition in the *Uniform Guidelines*, but it could have many interpretations. Though it is not defined, it is found in the sections dealing with the uses and applications of a selection procedure that has been evaluated by a criterion-related validity study. Specifically, when documenting the methods considered for using a procedure, it "should include the rationale for choosing the method of operational use, and the evidence of validity and utility of the procedure as it is to be used (essential)" (see 15.B. [10]). Identical sentences appear in the uses and applications sections for content and construct validity. Furthermore, in Section 5.G. the *Uniform Guidelines* states:

> If a user decides to use a selection procedure on a ranking basis, and that method of use has a greater adverse impact than use of an appropriate pass/fail basis . . ., the user should have sufficient evidence of validity and utility to support the use on a ranking basis.

COMPARISONS AMONG THE THREE AUTHORITIES

Given different authorships, different purposes, and different dates of adoption, it is useful to make comparisons among the three authorities to identify areas of agreement and disagreement. Such information might be particularly valuable to a user who is deciding about relying on one or more of the authorities or who has relied on one of the authorities and not realized what one or two of the other authorities had to say on the topic of interest.

The common themes across the three authorities are matters of validation and psychometric measurement. To facilitate this discussion, Table 27.1 has been prepared to compare the three authorities on several concepts or terms and their respective definitions or explanations. Before discussing any of the specifics, it is quickly obvious that there are many terms without definitions or explanations under the *Uniform Guidelines* column. There are, no doubt, several reasons for this situation, and two possible explanations may be offered:

- The *Uniform Guidelines* is some 35 years older than the *Standards* and 25 years older than the *Principles*. The latter two documents have undergone two revisions each since the *Uniform Guidelines* was published, but the *Uniform Guidelines* has never been revised or brought up to date, except for inclusion of additional Q&As.
- The *Uniform Guidelines* was written to guide the enforcement of civil rights legislation. The *Standards* and *Principles* were written to guide research and professional practice and to inform decision making

TABLE 27.1

Validation and Psychometric Terminology Comparison

	Standards 2014	*Principles 2003*	*Uniform Guidelines 1978*
Validity (unitary concept)	The degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test	The degree to which accumulated evidence and theory support specific interpretations of scores from a selection procedure entailed by the proposed uses of that selection procedure	Not defined
Sources of validity evidence			
(a) Relations to other variables/ criterion-related	The relationship of test scores to variables external to the test such as measures of some criteria that the test is expected to predict	The statistical relationship between scores on a predictor and scores on a criterion measure	Empirical data showing that the selection procedure is predictive of or significantly correlated with important elements of work behavior
(b) Content	The linkage between a predictor and one or more aspects of a criterion construct domain	The extent to which content of a selection procedure is a representative sample of work-related personal characteristics, work performance, or other work activities or outcomes	Data showing that the content of a selection procedure is representative of important aspects of performance on the job
(c) Internal structure	The extent to which the relationships between test items conform to the construct that is the foundation for test score interpretation	The degree to which psychometric and statistical relationships among items, scales, or other components within a selection procedure are consistent with the intended meanings of scores on the selection procedure	Not defined
(d) Response process	The study of the cognitive account of some behavior, such as making a selection procedure item response	The study of the cognitive account of some behavior, such as making a selection procedure item response	Not defined
(e) Consequences of testing	Whether or not the specific benefits expected from the use of a selection procedure are being realized	Evidence that consequences of selection procedure use are consistent with the intended meaning or interpretation of the selection procedure	Not defined, but possibly referenced in the term "utility"
Construct validity	An indication that a predictor measure represents a predictor construct domain combined with evidence of the linkage between the predictor construct domain and the criterion construct domain. The term "construct validity" is no longer used to define a "type" of validity.	Evidence that scores on two or more selection procedures are highly related and consistent with the underlying construct; can provide convergent evidence in support of the proposed interpretation of test scores as representing a candidate's standing on the construct of interest.	Data showing that the selection procedure measures the degree to which candidates have identifiable characteristics that have been determined to be important for successful job performance
Convergent validity	Evidence based on the relationship between test scores and other measures of the same constructs	Evidence of a relationship between measures intended to represent the same construct	Not defined
Discriminant validity	Evidence indicating whether two tests interpreted as measures of different constructs are sufficiently independent that they do measure two different constructs	Evidence of a lack of a relationship between measures intended to represent different constructs	Not defined
Validity generalization	Applying validity evidence obtained in one or more situations to other similar situations on the basis of methods such as transportability by job analysis or meta-analysis	Evidence of validity that generalizes to setting(s) other than the setting(s) in which the original validation evidence was documented. Generalized evidence is accumulated through such strategies as transportability, synthetic/job component validity, and meta-analysis.	Not defined

Transport of validity	Directly transporting validity evidence from another setting in a situation where sound evidence (e.g., careful job analysis) indicates that the local job is highly comparable to the job for which the validity data are being imported	A strategy for generalizing evidence of validity in which demonstration of important similarities between different work settings is used to infer that validation evidence for a selection procedure accumulated in one work setting generalizes to another work setting	Using evidence from another study when the job incumbents from both situations perform substantially the same major work behaviors as shown by appropriate job analyses; the study should also include an evaluation of test fairness for each race, sex, and ethnic group that constitutes a significant factor in the labor market for the job(s) in question within the labor force of the organization desiring to rely on the transported evidence
Synthetic/job component validity	Not defined	Generalized evidence of validity based on previous demonstration of the validity of inferences from scores on the selection procedure or battery with respect to one or more domains of work (job components)	Not defined
Meta-analysis	A statistical method of research in which the results from several independent, comparable studies are combined to determine the size of an overall effect on the degree of relationship between two variables	A statistical method of research in which results from several independent studies of comparable phenomena are combined to estimate a parameter or the degree of relationship between variables	Not defined
Reliability	The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable; the degree to which scores are free of errors of measurement for a given group	The degree to which scores for a group of assessees are consistent over one or more potential sources of error (e.g., time, raters, items, conditions of measurement, etc.) in the application of a measurement procedure	The term is not defined, but the reliability of selection procedures, particularly those used in a content viability study, should be of concern to the user
Fairness/unfairness	A test that is fair minimizes the construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some individuals. There is no single technical meaning; in the employment setting, fairness can be defined as an absence of bias and that all persons are treated equally in the testing process.	There are multiple perspectives on fairness. There is agreement that issues of equitable treatment, predictive bias, and scrutiny for possible bias when subgroup differences are observed are important concerns in personnel selection; however, there is not agreement that the term "fairness" can be uniquely defined in terms of any of these issues.	When members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group and the differences in scores are not reflected in differences in a measure of job performance
Predictive bias	The systematic under- or over prediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to criterion performance	The systematic under- or over prediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to criterion performance	Not defined, but see "Fairness/unfairness" above
Cut score/cutoff score	A specific point on a score scale such that scores at or above that point are interpreted or acted upon differently from scores below that point	A score at or above which applicants are selected for further consideration in the selection procedure. The cutoff score may be established on the basis of several considerations (e.g., labor market, organizational constraints, normative information). Cutoff scores are not necessarily criterion referenced, and different organizations may establish different cutoff scores on the same selection procedure on the basis of their needs.	Cutoff scores should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the workforce

Note: The above definitions or explanations are taken verbatim from the glossaries or definition section of the authoritative sources whenever possible. Otherwise, the definitions were extracted from document text on the subject.

in applicable areas of employment selection. Hence, the latter two documents have more of a scientific focus and rely heavily on the current research literature; the *Uniform Guidelines* was intended to be consistent with generally accepted professional standards set forth in the 1974 version of the *Standards* but was not necessarily research-based at the time of its preparation.

Standards Versus *Principles*

There are no areas of disagreement between the *Standards* and the *Principles*. In some areas the *Standards* offers more information and guidance than the *Principles*. Examples include (a) discussions of validity evidence based on response processes, internal structure, and the consequences of testing; (b) discussions of reliability and errors of measurement; (c) the test development and revision process; (d) scales, norms, and score comparability; and (e) the rights and responsibilities of test takers. A few topics are more broadly considered in the *Principles* than is true for the *Standards*. Examples include (a) the concept of the analysis of work (to incorporate the work context and organizational setting) rather than job analysis; (b) clarifying that the generalization of validity evidence can be accomplished by several methods, including transportability and synthetic/job component validity, as well as being supported by meta-analysis; and (c) certain operational considerations associated with the conduct of a validation study in organizational settings (e.g., communications, organizational needs and constraints, quality control and security, implementation models, and utility).

Validity (Unitary Concept)

The *Standards* and the *Principles* view validity as a unitary concept, whereas the *Uniform Guidelines* partitions validity into three types: criterion-related, content, and construct. This partitioning of validity was the thinking 40 years ago, but it is clearly out of date now.

Sources of Validity Evidence

(a) *Relations to other variables/criterion-related*: The *Uniform Guidelines*' focus on work behavior as a criterion excludes potential studies of the relationships between a selection procedure of interest and other tests hypothesized to measure the same or different constructs (i.e., other external variables).

(b) *Content*: All three authorities agree that content validity is dependent on a sound determination that the selection procedure is a representative sample of work-related behavior. The analysis of work (or the job) is fundamental to establishing the predictor-criterion linkage. The *Uniform Guidelines* confines job requirements to a study of KSAs; the *Standards* and *Principles* provide for the study of KSAOs and would include "O" variables in a selection procedure subject to a content validity study. The *Uniform Guidelines* precludes use of a content strategy to study the validity of traits or constructs such as spatial ability, common sense, judgment, or leadership. Although it is important to describe the relevant work behavior or KSAO at a level of specificity so there is no misunderstanding about what is being measured, it is unnecessary and unwise to reject content validity evidence simply because it is concerned with linking an ability or personal characteristic (i.e., leadership) to the domain of job performance. Many constructs can be defined in terms of specific work behaviors although they have broad labels. Furthermore, there are many situations in which content validity may be the only option. If leadership capabilities are critical to job performance, validity evidence beyond a content validity study may be infeasible. There may not be adequate numbers of candidates or incumbents to conduct a criterion-related study, and there may not be sufficient and reliable criteria available. Consequently, a content validity study may be the only viable approach to evaluating the validity of a construct of interest.

(c) *Internal structure/response processes/consequences of testing*: These three lines of evidence for a validity argument were not developed at the time the *Uniform Guidelines* was written and hence are not discussed in it.

Construct Validity

The *Uniform Guidelines* treats construct validity as a separate type of validity. In the *Standards* and *Principles*, all selection procedure scores or outcomes are viewed as measures of some construct. Consequently, any evaluation of validity is a "construct validity" study.

Convergent and Discriminant Validity

Although these terms and their implications were well-established at the time the *Uniform Guidelines* was prepared, there was no discussion about the value of these types of evidence in the document.

Validity Generalization

The concept was known at the time the *Uniform Guidelines* was prepared but was not specifically used in the document. Many have interpreted Section 7.B of the *Uniform Guidelines* as providing for validity generalization arguments. The provisions of that section are described under transport of validity evidence in Table 27.1.

Transport of Validity

The three authoritative sources agree that a work or job analysis is necessary to support the transport of validity. However, the *Uniform Guidelines* goes further and requires that there be an existing criterion-related validity and a fairness study of the selection procedure for relevant protected subgroups. However, there is no guidance as to the acceptability of transporting the validity of a selection procedure that has some demonstrated unfairness. Furthermore, as noted previously, in many situations, sample sizes may preclude adequate fairness analyses (Aguinis & Stone-Romero, 1997).

Synthetic/Job Component Validity

This validity generalization strategy has been known for more than 40 years but has not received much attention in validation research conducted outside of the employment arena. Neither the *Standards* nor the *Uniform Guidelines* have defined this strategy of validity generalization.

Meta-Analysis

In 1978 the authors of the *Uniform Guidelines* did not have knowledge of the research findings that have emerged subsequently from meta-analytic research. This, unfortunately, is another void, and a significant amount of research is available today that might not be considered to be within the scope of validation strategies acceptable under the *Uniform Guidelines*.

Reliability

The term *reliability* is not defined in the *Uniform Guidelines* as it is in the other two authoritative sources, but it is considered to be important for selection procedures that have been supported

with a content validity strategy. The *Standards* and *Principles* emphasize that the reliability of any measurement be considered whenever it is technically feasible to do so.

Fairness/Unfairness and Bias

The *Standards* and the *Principles* consider fairness to be a very broad concept with many facets. Alternatively, the two sources consider bias to be a very specific term concerned with under- or overprediction of subgroup performance. This interpretation is basically the one that the *Uniform Guidelines* gives to the term *unfairness* while relying on the 1974 version of the *Standards*.

Cut Score/Cutoff Score

The *Standards* and *Principles* give more attention to developing in detail many of the issues underlying the setting of cutoff scores than does the *Uniform Guidelines*. However, there does not seem to be any significant disagreement across the three documents as to how a cutoff score will function and the intent for a cutoff score to screen out those who will not achieve acceptable levels of job performance.

Summary

There are some levels of consistency or agreement across the three authoritative sources but also consequential areas of disagreement. It is very likely that the advances in selection procedure research and scholarly thinking regarding validity that have occurred over the last 35 years account for these differences. Although the *Uniform Guidelines* is the document that seems most deficient in terms of knowledge of the field, it is also the first document of the three in terms of its adoption. On that basis, its deficiencies can be excused by being out of date; however, as noted earlier in this chapter, the authors of the *Uniform Guidelines* allowed for other procedures and issues to arise and envisioned their potential inclusion in the framework laid out by the document. Sections 5.A and 5.C acknowledge, respectively, that "New strategies for showing the validity of selection procedures will be evaluated as they become accepted by the psychological profession" and that "The provisions of these guidelines . . . are intended to be consistent with generally accepted professional standards . . . and standard textbooks and journals in the field of personnel selection." These clauses can be interpreted to suggest that deference should be given to the *Principles* and *Standards* where they disagree with the *Uniform Guidelines*. Despite these forward-looking provisions, no substantive changes have ever been made in the *Uniform Guidelines*, even though case law has changed various provisions and interpretations (e.g., bottom-line analyses). Arguably, this state of affairs reflects the significant interaction between the *Uniform Guidelines* and the case law as it has developed since its adoption. Indeed, the Supreme Court (1971) indicated that the EEOC's earlier *Guidelines* (predecessor to the *Uniform Guidelines*) was to be given "great deference" by the courts. Changes to the *Uniform Guidelines* will likely be controversial and difficult, if possible at all. Nevertheless, some time in the near future it will be important for the *Uniform Guidelines* to be revised to reflect the current state of the science. Until that time, the decision maker involved in employment selection should look to the *Standards* and *Principles* for guidance on many issues that either are now incorrect or are not addressed in the *Uniform Guidelines*.

FINAL THOUGHTS

Science Versus Litigation Versus Technical Authorities/Guidelines

It is recognized that there are some significant inconsistencies at this time between the technical information provided by the *Standards* and *Principles*, on the one hand, and the *Uniform Guidelines*,

on the other hand, and that these differences can be extremely important in the event of litigation regarding a selection procedure. However, these differences can be resolved. Unfortunately, until a revision to the *Uniform Guidelines* is forthcoming, to the extent that there is more than one authority introduced in litigation that is offered as support to only one side of an argument, resolution of differences that appear in print will need to be part of the judicial decision-making process. In this regard, it is incumbent upon those who do rely on any of the authoritative sources during the course of litigation to be clear about the relevance and currency of the source(s) that are providing guidance to their opinions.

Conclusions

In closing, we want to note several broad, as well as some specific issues. We will start with the broader issues. First, what deference should be given to the *Uniform Guidelines*, *Principles*, and *Standards* in guiding psychologists as they make decisions in employment settings? We ask this question given that the three documents are in many ways static, whereas the field is dynamic. That is, research is constantly being conducted that provides new knowledge and/or influences how we interpret behavioral phenomena. For example, it is a commonly accepted fact that the validity of cognitive ability tests generalizes across situations and jobs (Hunter & Hunter, 1984). Yet, this was not always the "accepted" fact; in the 1960s, validity was described as "situation specific" (Ghiselli, 1966). If there had been three sets of sources promulgated by various agencies in the 1960s, they most likely would have advocated for "situation specificity," and the accepted practice would have been to validate tests in every situation for every job. The point of this example is that perhaps the current sources—*Uniform Guidelines*, *Principles*, and *Standards*—should not be viewed as authoritative regarding knowledge, but rather as primers for "how to conduct research" and what factors to consider when determining the validation of a test.

Reliance on the sources for new research findings may hamper the field. The documents are not "living" and thus cannot account for changes due to new research. However, the practitioner or researcher can rely on the sources with regard to how to establish the validity of a test and what information is needed as part of the research.

Acceptance of the above premise brings us to the second broad issue. Given that the sources are relied upon in litigation, whether introduced directly in testimony in court cases or as authority references when trying to explain to judges and lawyers what and how we conduct our research, the question becomes "How sound are the sources as authoritative documents in court proceedings?"

One potential set of criteria are the "Daubert thresholds" (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993), which set forth rules for determining what is expert testimony or scientific evidence:

1. Testing—adequate testing can be or has been tested by collection of data with an accepted methodology
2. Has a known or potential error rate
3. Has been subjected to peer review and publications
4. Has gained general acceptance in a relevant scientific community

Another concern is the need to consider that the global economy is changing the way in which humans work and with whom they work. Accordingly, future sources should address cultural issues and the changing nature of work. Some examples of these issues include:

1. The need to consider assessment of individuals with diverse linguistic backgrounds as well as the need to accommodate test takers whose first language is not English.
2. The need to consider electronic, Internet, and web-based technology and the fact that the next generation of workers will likely have not been exposed to the same methods of training, operating, and performing at work as the current generation. Advanced technology may provide for greater opportunity to capture actual samples or simulations of job behaviors than are garnered in paper-and-pencil multiple-choice formats.

3. The need to identify criteria that are relatively focused on more short-term gains than those that have been used in the past (e.g., tenure in the position for at least one year). A global pace of competition implies that businesses will need to reduce losses (such as incorrect "hires") and increase gains (such as faster training times) much more quickly than was common in the past.
4. The need to recognize that current tests explain, at most, approximately 25% of the variance in job performance as we measure it today. Although it is appropriate to concern ourselves with searching for additional predictors, we need to consider ways in which to broaden the criterion space and how to combine the criteria in such a fashion as to provide a "comprehensive" picture of the worker. That is, although we can predict to a reasonable degree (15–25% of the variance) how well entering college students may perform as represented by the criterion of final grade point average, we need to examine other factors that measure success in college and how these additional factors can be combined to represent success in the "college experience."

Authoritative sources that incorporate principles, guidelines, and standards have a valuable role to play in the science of employment selection; however, the limitations inherent to such sources must be openly recognized, and to the degree there is disagreement or conflicts among the sources, they should be revealed before they attain a stature that creates a disservice to employees, employers, and I-O psychology professionals.

NOTE

1. This chapter with modifications and considerable updating is based on "Professional and Technical Authorities and Guidelines" by P. Richard Jeanneret, which is found in Landy, F. J. (2005). *Employment discrimination litigation: Behavioral, quantitative and legal perspectives.* San Francisco, CA: John Wiley & Sons.

REFERENCES

Aguinis, H., & Stone-Romero, E. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology*, *82*, 192–206.

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

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

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

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). The FFM personality dimensions and jobs: Meta-analysis of meta-analysis. *International Journal of Selection and Assessment, 9*, 9–30.

Bartlett, C. J., Bobko, P., & Mosier, S. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology*, *31*, 233–245.

Bemis, S. E. (1968). Occupational validity of the general aptitude test battery. *Journal of Applied Psychology, 52*, 240–249.

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

Callender, J. C., & Osburn, H. G. (1981). Testing the constancy of validity with computer-generated sampling distributions of the multiplicative model variance method estimate: Results for petroleum industry validation research. *Journal of Applied Psychology, 66*, 274–281.

Camara, W. J. (1996). Fairness and public policy in employment testing: Influences from a professional association. In R. S. Barrett (Ed.), *Fair employment strategies in human resource management*. Westport, CT: Quorum Books.

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

Cascio, W. F., & Boudreau, J. (2011). *Investing in people: The Financial impact of human resource initiatives* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

Cascio, W., Jacobs, R., & Silva, J. (2010). Validity, utility, and adverse impact: Practical implications from 30 years of data. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 271–288). New York, NY: Routledge, Taylor & Francis Group.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 5, 115–124.

Colosimo, J. (2010). *A primer on adverse impact analysis. White Paper, DCI Consulting Group, Inc*. Downloaded from http://dciconsult.com/whitepapers/AIPrimer.pdf, August 16, 2015.

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

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

Equal Employment Opportunity Commission, Civil Service Commission, Department of Justice, & Department of Labor. (1978). *Uniform guidelines on employee selection procedures*. 29 CFR, 1607.

Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of Labor, & Department of Treasury. (March 2, 1979). *Questions and answers to clarify and provide a common interpretation of the Uniform Guidelines on Employee Selection Procedures*. 44FR, No. 43.

European Federation of Psychologists' Associations. (2013). EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests, Version 4.2.6.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. New York, NY: John Wiley.

Gibson, W. M., & Caplinger, J. A. (2007). Transportation of validation results. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 29–81). San Francisco, CA: John Wiley and Sons.

Hartigan, J. A., & Wigdor, A. K. (Eds.) (1989). *Fairness in employment testing*. Washington, DC: National Academy Press.

Hazelwood School District v. United States, 433 U.S. 299.31 n. 17 (1977).

Hoffman, C. C., Rashkovsky, B., & D'Egidio, E. (2007). Job component validity: Background, current research, and applications. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 82–121). San Francisco, CA: John Wiley and Sons.

Hogan, J., Davies, S., & Hogan, R. (2007). Generalizing personality-based validity evidence. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 181–229). San Francisco, CA: John Wiley and Sons.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial-organizational psychology: Reflections, progress, and prospects. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 272–290.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–88.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed). Newbury Park, CA: Sage.

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

International Standards Organisation. (2011). Assessment serviced delivery—Procedures and methods to assess people in work and organisational settings. ISO-10667–2.

International Taskforce on Assessment Center Operations. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management, 41*, 1244–1273.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing, 1*(2), 93–114.

International Test Commission. (2005). *International Guidelines on Test Adaptation*. Retrieved from www.intestcom.org

Johnson, J. W. (2007). Synthetic validity: A technique of use (finally). In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 122–158). San Francisco, CA: John Wiley and Sons.

Landy, F. J. (Ed.). (2005). *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives*. San Francisco, CA: Jossey-Bass.

McDonald, R. P. (1999). *Test theory*: *Unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York, NY: Macmillan.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683–729.

Naylor, J. C., & Shine, L. C. (1965). A table for determining the increase in mean criterion scores obtained by using a selection device. *Journal of Industrial Psychology, 3*, 33–42.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement*, *design*, *and analysis*: *An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*, 153–172.

Putka, D. J., & Sackett, P. R. (2010). Reliability and validity. In J. L. Farr & N. T. Tippins(Ed.), *Handbook of employee selection* (pp. 9–49). New York, NY: Routledge.

Ryan, A. M., & Powers, C. L. (2012). Workplace diversity. In N. Schmitt (Ed.), *Oxford handbook of personnel assessment and selection* (pp. 814–831). London: Oxford University Press.

Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49*, 549–572.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. *American Psychologist, 56*, 302–318.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64*, 609–626.

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., Zorzie, M. (2009). Prediction of 4-years college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology*, *94*, 1479–1497.

Schmitt, N., & Quinn, A. (2010). Reductions in measured subgroup differences: What is possible. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 425–451). New York: Routledge, Taylor & Francis Group.

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.

Traub, R. E. (1994). *Reliability for the social sciences*: *Theory and applications*. Thousand Oaks, CA: Sage.

Wainer, H., & Braun, H. I. (Eds.) (1988). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.