Industrial and Organizational Psychology, *11*(S1), pp 1–97 December 2018.
 doi:10.1017/iop.2018.195

FOREWORD

The Society for Industrial and Organizational Psychology (SIOP) is pleased to offer the fifth edition of the *Principles for the Validation and Use of Personnel Selection Procedures*, which was approved by the APA Council of Representatives in August 2018 as an authoritative guidelines document for employee selection testing and an official statement of the APA. Over a three-year period, the *Principles* Revision Committee updated this document from the fourth edition to be consistent with the 2014 *Standards for Educational and Psychological Testing*, invited commentary from SIOP and APA that informed subsequent revisions, and solicited a thorough legal review.

The *Principles* Revisions Committee was chaired by Nancy Tippins, PhD, and Paul Sackett, PhD, and its members included Winfred Arthur, PhD; Tanya Delaney, PhD; Eric Dunleavy, PhD; Ted Hayes, PhD; Leaetta Hough, PhD; Fred Oswald, PhD; Dan Putka, PhD; Ann Marie Ryan, PhD; and Neal Schmitt, PhD. Collectively, the committee devoted an enormous number of hours to the revision to ensure that the fifth edition of the *Principles* reflects current research on, and best practices for, the development, validation, and implementation of employee selection procedures.

SIOP is indebted to the *Principles* Revision Committee and to the many members of SIOP and APA who provided commentary.

Nancy Tippins, PHD, and Paul Sackett, PHD
Co-Chairs, SIOP Principles Revision Committee

Fred Oswald, PHD
SIOP President, 2017–2018

Principles for the Validation and Use of Personnel Selection Procedures

INTRODUCTION

Statement of Purpose

The purpose of the *Principles for the Validation and Use of Personnel Selection Procedures* (hereafter referred to as the *Principles*) is to specify established scientific findings and generally accepted professional practice in the field of personnel selection psychology. These include the choice, development, evaluation, and use of personnel selection procedures designed to measure constructs related to work behavior, with a focus on the accuracy of the inferences that underlie personnel decisions. This document is the fifth edition of the *Principles*, which is the official statement of the Society for Industrial and Organizational Psychology (Division 14 of the American Psychological Association [APA] and an organizational affiliate of the American Psychological Society [APS]) concerning validation and personnel selection. The revision is stimulated by theoretical and research developments since the previous edition of the *Principles* (SIOP, 2003) and by the publication of the *Standards for Educational and Psychological Testing* in 2014 (hereafter referred to as the *Standards*) by the American Educational Research Association (AERA), APA, and the National Council on Measurement in Education (NCME). The *Principles* covers many aspects of validation and personnel selection; however, other professional documents may also provide guidance in particular situations (e.g., *Guidelines and Ethical Considerations for Assessment Center Operation* [The International Taskforce on Assessment Center Guidelines, 2015]; *Multicultural Guidelines: An Ecological Approach to Context, Identity, and Intersectionality* [APA, 2017b]; *International Guidelines on Test Use* [International Test Commission, 2013]; and *Professional Practice Guidelines for Occupationally Mandated Psychological Evaluations* [APA, 2018]).

The *Principles* is intended to be consistent with the *Standards*. This revision brings the *Principles* up to date regarding current scientific knowledge, and it further guides sound practice in the use of personnel selection procedures. The *Principles* should be taken in its entirety rather than considered as a list of separately enumerated principles.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Federal, state, and local statutes, regulations, and case law regarding personnel decisions exist both in the U.S. and in many other countries. The *Principles* is not intended to interpret these statutes, regulations, and case law but to provide guidance about psychological methods relevant to contexts that the statutes, regulations, and case law govern.

The *Principles* provides:

1. principles regarding the conduct of selection and validation research;
2. principles regarding the application and use of selection procedures;
3. information for those responsible for authorizing or implementing validation efforts; and
4. information for those who evaluate the adequacy and appropriateness of selection procedures.

Principles as guidance

It is important to recognize that the *Principles* constitutes pronouncements that guide, support, or recommend, but do not mandate, specific approaches or actions. The *Principles* is intended to be aspirational and to facilitate and assist the validation and use of selection procedures. It is not intended to be mandatory, exhaustive, or definitive, and it may not be applicable to every situation. Sound practice requires professional judgment to determine the relevance and importance of the *Principles* in any particular situation. The *Principles* is not intended to mandate specific procedures independent of the professional judgment of those with expertise in the relevant area. In addition, the *Principles* is not intended to provide advice on complying with local, state, federal, or international laws that might be applicable to a specific situation.

The *Principles* expresses expectations toward which the members of this Society and other testing professionals should strive. Evidence for the validity of the inferences from a given selection procedure may be weakened to the extent that the expectations associated with professionally accepted practice, and consequently the *Principles*, are not met. However, circumstances in any individual validation effort or application affect the relevance of a specific principle or the feasibility of its implementation. Complete satisfaction of the *Principles* in any given situation may not be necessary or attainable.

The *Principles* is intended to represent the consensus of professional knowledge and practice as it exists today; however, personnel selection research and development is an evolving field in which techniques and decision-making models are subject to change. Acceptable procedures other than those discussed in this edition of the *Principles* may be developed in the future. In certain instances, references are cited that provide support for the *Principles*, but these citations are selective rather than exhaustive. Testing

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

professionals are expected to maintain an appropriate level of awareness of research developments relevant to the field of personnel selection.

The *Principles* is not intended:

1. to be a substitute for adequate education and training in validation theory and procedures;
2. to be exhaustive (although it covers the major aspects of selection procedure validation and use);
3. to be a technical translation of existing or future regulations;
4. to freeze the field to prescribed practices and so limit creative endeavors; or
5. to provide an enumerated list of separate principles.

Selection Procedures Defined

Depending on one's focus, selection procedures or predictors can be described in terms of what they measure (content/constructs) or how they measure what they are designed to measure (methods). The domain of predictors (i.e., what they measure) can be delineated by theories of psychological constructs (e.g., knowledge, skills, abilities, and other personal characteristics [KSAOs] or competencies), theories of job situations/demands, or even some combination of the two. Predictor methods, on the other hand, refer to the specific processes or techniques by which domain-relevant information is elicited, collected, and subsequently used to make inferences. Examples of these selection procedures methods include, but are not limited to, paper-and-pencil tests, computer-administered tests, performance tests, work samples, inventories (e.g., measures of personality and interests), individual assessments, interviews, assessment centers, situational judgment tests, biographical data forms or scored application blanks, background investigations, education, experience, physical requirements (e.g., height, weight), physical ability tests, and appraisals of job performance. In addition, unproctored internet-based tests, "big data" and machine learning methods (e.g., harvesting information about candidates from social media sites, resumes, or other sources of text or information), gamification, and computer-based simulations of varying levels of technological sophistication are examples of contemporary testing and assessment approaches. In summary, selection procedures can represent a wide variety of methods of measurement that can be used to assess a wide variety of constructs (i.e., KSAOs or competencies) that underlie personnel decision making.

The terms "selection procedure," "test," "predictor," and "assessment" are used interchangeably throughout the *Principles*. Personnel decisions are decisions to hire, train, place, certify, compensate, promote, terminate, transfer, or take other actions that affect aspects of employment.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

The field of personnel selection psychology aims at improving the quality of personnel selection decisions through the systematic development, evaluation, and implementation of job-related selection systems. Doing so is of value to organizations, as it results in a workforce better suited to meet job requirements. Absent the interventions of selection psychologists or other professionals, selection procedures are generally informal and ad hoc, and rely on the judgment of one or more decision makers (e.g., in screening resumes and interviewing candidates). There is an extensive literature on bias in subjective judgments; see, for example, Koch, D'Mello, and Sackett (2014) and Roth, Purvis, and Bobko (2012) for meta-analyses of gender bias in employment decision making, and Quillian, Pager, Hexel, and Midtbøen (2017) for a meta-analysis of racial discrimination in employee screening. Thus, an additional value of systematic selection systems is a reduction in the reliance on subjective decisions and their biases. In short, key goals of the development, evaluation, and implementation of systematic selection systems are thus improved prediction of desired work outcomes and the avoidance of bias in employment decisions.

OVERVIEW OF THE VALIDATION PROCESS

The essential principle in the evaluation of any selection procedure is that evidence be accumulated to support an inference of job relatedness. The job relatedness of a selection procedure has been demonstrated when evidence supports the accuracy of inferences made from scores on, or evaluations derived from, those procedures regarding some important aspect of work behavior (e.g., quality or quantity of job performance; performance in training, advancement, tenure, turnover, or other organizationally pertinent behavior). Although the *Principles* focuses on individual performance, group and organizational performance may also be relevant criteria.

Any claim of validity made for a selection procedure should be documented with appropriate research evidence built on the principles discussed in the *Principles.* Promotional literature or testimonial statements should not be used as evidence of validity.

The *Principles* embraces the *Standards*' definition of validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). Validity is the most important consideration in developing and evaluating selection procedures. Because validation involves the accumulation of evidence to provide a sound scientific basis for the proposed score interpretations, it is the interpretations of these scores required by the proposed uses that are evaluated, not the selection procedure itself.

The *Standards* notes that validation begins with "an explicit statement of the proposed interpretation of test scores, along with a rationale for the

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

relevance of the interpretation to the proposed use. The proposed interpretation includes specifying the construct the test is intended to measure" (AERA et al., 2014, p. 11). Examples of such constructs or concepts include arithmetic proficiency, managerial performance, ability to design a web page, oral presentation skills, conscientiousness, and ability to troubleshoot technical problems. A clear description of the construct or conceptual framework that delineates the knowledge, skills, abilities, processes, and other characteristics to be assessed should be developed.

In the early 1950s, three different aspects of test validity were discussed: content, criterion related, and construct. Since that time, the conceptualization of validity evidence has undergone some modification, moving from three separate aspects of validity evidence to the current *Standards*' view of validity as a unitary concept with different sources of evidence contributing to an understanding of the inferences that can be drawn from a selection procedure. Nearly all information about a selection procedure contributes to an understanding of the validity of inferences drawn from the procedure. Evidence concerning content relevance, criterion relatedness, and construct meaning is subsumed within this definition of validity.

The validity of any inference can be determined through a variety of different strategies for gathering evidence. The *Standards* notes that although different strategies for gathering evidence may be used, the primary inference in employment contexts is that a score on a selection procedure predicts subsequent work behavior. Even when the validation strategy used does not involve empirical predictor–criterion relationships, such as when a user relies on conceptual linkages between test content and job content to provide validation evidence, there is still an implied link between the test score and a criterion. Therefore, even when different strategies are employed for gathering validation evidence, the inference to be supported is that scores on a selection procedure can be used to predict sub- sequent work behavior or outcomes. Professional judgment should guide the decisions regarding the sources of evidence that can best support the intended interpretation and use.

The quality of validation evidence is of primary importance. In addition, where contradictory evidence exists, comparisons of the weight of evidence supporting specific inferences to the weight of evidence opposing such inferences are critical.

The *Standards* discusses five sources of evidence that can be used in evaluating a proposed interpretation of test scores for a particular use: specifically, evidence based on (a) relationships between test scores and other variables, such as test–criterion relationships;

(b) test content; (c) internal structure of the test; (d) response processes; and (e) consequences of testing. Given that validity is a unitary concept, such

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

categorizations refer to various sources of evidence rather than distinct types of validity. It is not the case that each of these five sources is an alternative approach to establishing job relatedness. Rather, each provides information that may be highly relevant to some proposed interpretations of scores and less relevant, or even irrelevant, to others.

Sources of Evidence

Evidence based on the relationship between scores on predictors and other variables

This form of evidence is based on the empirical relationship of predictor scores to external variables. Two general strategies for assembling empirical evidence apply. The first strategy involves examining the relationship between scores on two or more selection procedures measuring the same construct hypothesized to underlie the predictor measure. Evidence that two measures 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. Similarly, evidence that test scores relate differently to other distinct constructs can contribute to evidence of discriminant validity. Note that evidence of convergent and discriminant validity does not in and of itself establish job relatedness, which leads to the second strategy for assembling empirical evidence: relating selection procedure scores to work-relevant behaviors or outcomes. This strategy has historically encompassed two study designs: predictive and concurrent. A predictive study examines how accurately test scores predict future performance. In a concurrent study, predictor and criterion data are collected at roughly the same time although the objective remains to predict performance.

Content-related evidence

Test content includes the questions, tasks, format, and wording of questions, response formats, instructions, and guidelines regarding administration and scoring of the test. Evidence based on test content may include logical or empirical analyses that evaluate the adequacy of the match between test content and work content, worker requirements, or outcomes of the job.

Evidence based on the internal structure of the test

Studies that examine the internal structure of a test and the relationship among its items or tasks (e.g., work samples) can provide additional evidence of how test scores relate to specific aspects of the construct to be measured. Such evidence typically includes information concerning the relationships among items and the degree to which they represent the appropriate construct or content domain. For example, evidence that items on a test

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

represent a single construct or multiple constructs may be evaluated by fitting an appropriate structural model to the items (e.g., a confirmatory factor analysis model). Generic indices of consistency among items (e.g., coefficient alpha) do not provide an evaluation of the internal structure of the test. When a multidimensional factor structure is proposed, evidence supporting inferences concerning the validity of score interpretations for the subcomponents in the predictor may be appropriate. Note that evidence of internal structure provides empirical support for the construct being measured; it does not in and of itself establish job relatedness, which requires additional evidence linking selection procedure scores to work-relevant behavior or outcomes.

Evidence based on response processes

In employment contexts, evidence based on response processes is necessary when claims are made that scores can be interpreted as reflecting a particular response process on the part of the examinee. For example, if a claim is made that a work sample measures the use of proper techniques for resolving customer service problems, then simply assessing whether the problem is resolved is not enough. Evidence based on both cognitive and physical response processes may provide additional evidence of validity. Examining the processes used by individuals in responding to performance tasks or test questions can provide such evidence. Often evidence regarding individual responses can be gathered by (a) questioning test takers about their response strategies, (b) analyzing examinee response times on computerized assessments, or (c) conducting experimental studies where the response set is manipulated. Observations of how individuals engage in performance tasks can also illustrate the extent to which the task is eliciting behavior related to the intended construct as opposed to behavior more related to irrelevant constructs. However, in many employment contexts, such evidence is irrelevant to the proposed use, as is the case where the only claim made is that the scores on the selection procedure are predictive of a particular work-relevant behavior or outcome.

Evidence for validity and consequences of personnel decisions

In recent years, one school of thought has advocated incorporating the examination of consequences of the use of predictors in the determination of validity. This perspective views unintended negative consequences as weakening the validity argument. Although evidence of negative consequences may influence policy or practice decisions concerning the use of predictors, the *Principles* and the *Standards* take the view that such evidence is relevant to inferences about validity only if the negative consequences can be attributed to the measurement properties of the selection procedure itself.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Subgroup differences in test scores and subsequent differences in selection rates resulting from the use of selection procedures are often viewed as a negative consequence of personnel decisions. Group differences in predictor scores and selection rates are relevant to an organization and its personnel decisions; yet, such differences alone do not detract from the validity of the intended test interpretations. If the group difference can be traced to a source of bias in the test (i.e., measurement bias), then the negative consequences do threaten the validity of the interpretations. Alternatively, if the group difference on the selection procedure is consistent with differences between the groups in the work-relevant behavior or outcome predicted by the procedure (i.e., lack of predictive bias), then the finding of group differences could actually support the validity argument. In this case, negative consequences from test use constitute a policy issue for the user rather than indicate negative evidence concerning the validity of the selection procedure.

A different example of negative consequences is also helpful. An organization that introduces an integrity test to screen applicants may assume that based on the validity evidence for the test, this selection procedure provides an adequate safeguard against employee theft and will discontinue use of other theft-deterrent methods (e.g., video surveillance). In such an instance, employee theft might actually increase after the integrity test is introduced and other organizational procedures are eliminated; theft has increased because of a change in procedures, not because of the deficiency of the integrity test. Thus, the decisions subsequent to the introduction of the test may have had an unanticipated negative consequence on the organization. Such consequences may lead to policy or practice decisions to reduce the negative impact. However, such consequences do not threaten the validity of inferences that can be drawn from the integrity test scores.

Planning the Validation Effort

Validation should begin with a clear statement of the proposed uses of a test as well as the intended interpretations and outcomes. Selection procedures should be supported by appropriate validity evidence. When a selection decision is based on multiple components combined into a composite, evidence for the final decision has primary importance. The validation effort should accumulate evidence that generalizes to the selection procedure and work behavior in the operational setting. The design of this effort may take many forms, such as single local studies, consortium studies, meta-analyses, transportability studies, or synthetic validity/job component studies. More than one source of evidence or validation strategy may be valuable in any one validation effort.

In planning a validation effort for personnel decisions, three sources of evidence are most likely to be relevant: evidence of relationships with

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

measures of other variables, content-related evidence, and evidence of internal structure. Under some circumstances, evidence based on response processes and evidence based on consequences may also be important to consider. The decision to pursue one or more of these sources of evidence is based on many considerations, including proposed uses, types of desired selection procedures, availability and relevance of existing information and resources, and strength and relevance of an existing professional knowledge base. Where the proposed uses rely on complex, novel, or unique conclusions, multiple lines of converging evidence may be important.

The design of the validation effort is the result of professional judgment balancing considerations that affect the strength of the intended validity inference with practical limitations. Important considerations include (a) existing evidence, (b) design features required by the proposed uses, (c) design features necessary to satisfy the general requirements of sound inference, and (d) feasibility of particular design features.

Existing evidence

An important consideration in many validation efforts is whether sufficient validity evidence already exists to support the proposed uses. The availability and relevance of existing evidence and the potential informational value of new evidence should be carefully weighed in designing the validation effort. All validity conclusions are generalizations from the results in the validation setting to selection procedures and work behavior in the operational setting. The informational value of existing and possible new evidence is based on the many factors that affect the strength of this generalization.

Existing evidence provides informational value when it establishes a statistical relationship and supports the generalization from the validation setting to the operational setting. When such evidence has been accumulated, it may provide a sufficient rationale for inferring validity in the operational setting and may support a decision not to gather additional evidence. Such inferences depend on evidence of validity rather than mere claims of validity. Advances in meta-analytic methods and a growing knowledge base of meta-analytic results have established considerable validation evidence for cognitive ability measures, and increasing evidence is accruing for some noncognitive measures as well. When a validation study that meets professional standards cannot be conducted, it is particularly important to accumulate evidence of validity from other sources. However, existing evidence alone may not be sufficient to support inferences of validity in a given situation.

Validity conclusions based on existing evidence may be strengthened by evidence from more than one method, especially when the validity inference depends heavily on some underlying or theoretical explanatory concept or construct. However, in some cases, different methods may not support the

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

same conclusions about the underlying explanatory concepts or constructs. For example, factor analyses of test scores may not replicate factor analyses of supervisor ratings of the same attributes. In these situations, convergent and discriminant evidence across multiple methods may be important.

Proposed uses

In designing a validation effort, whether based on existing evidence, new evidence, or both, primary consideration should be given to the design features necessary to support the proposed uses. Examples of such features include the work to be targeted (e.g., one job title or job family), the relevant candidate pool (e.g., experienced or inexperienced candidates), the uniqueness of the operational setting (e.g., one homogeneous organization or many different organizations), and relevant criterion measures (e.g., performance or turnover).

Requirements of sound inference

Primary consideration should also be given to the general requirements of sound validity inferences, including measurement reliability and validity, representative samples, appropriate analysis techniques, and appropriate statistical and design controls over plausible confounding factors. People who provide information in the validation effort should be qualified for the tasks they are asked to perform and knowledgeable about the information they are asked to contribute.

Feasibility

Validation planning must consider the feasibility of the design requirements necessary to support an inference of validity. Validation efforts may be limited by time, resource availability, sample size, or other organizational constraints, including cost. In some situations, these limits may narrow the scope of appropriate generalizations, but in other situations they may cause design flaws leading to inaccurate generalizations. Although validation efforts with a narrow focus may have value, poorly executed validation efforts may lead the employer to reject beneficial selection procedures or accept invalid ones. Misleading, poorly designed validation efforts should not be undertaken.

Analysis of Work

Historically, selection procedures were developed for specific jobs or job families. This often remains the case today, and traditional work analysis methods are still relevant and appropriate in these situations. However, organizations that experience rapid changes in the external environment, the nature of work, or processes for accomplishing work may find that traditional jobs are being transformed or no longer exist. In light of changes to

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

the nature of work over the past decades, increasing numbers of organizations are shifting from job-specific knowledge, ability, and skill requirements when describing work, to a focus on broader competency-based requirements. Competency models are often used by organizations for many different purposes. When they are intended to support the underlying validity or use of a selection procedure, the *Principles* applies. The term "analysis of work" is used throughout the *Principles* and subsumes information that traditionally has been collected through work and job analysis methods, and more recently, competency modeling efforts as well as other information about the work, worker, organization, and work environment. The focus for conducting an analysis of work may include different dimensions or characteristics of work, including work complexity, environment, context, tasks, behaviors and activities performed, and worker requirements (e.g., KSAOs or competencies).

Purposes for conducting an analysis of work

In the context of validation research, there are generally two major purposes for conducting an analysis of work. One purpose is to develop or identify selection procedures. Part of this development process is an analysis of work that identifies worker requirements, including a description of the KSAOs or competencies needed. Such an analysis would determine the characteristics workers need to be successful in a specific work setting or the degree to which the work requirements are similar to the requirements for work performed elsewhere. The second purpose is to develop or identify criterion measures by assembling the information needed to understand the work performed, the setting in which the work is accomplished, and the organization's goals.

There is no single approach that is the preferred method for the analysis of work. The analyses used in a specific study of work are a function of the nature of the work, current information about the work, the organizational setting, the workers themselves, and the purpose of the study. Understanding the organization's requirements or objectives is important when selecting an appropriate method for conducting an analysis of work. The choice of method and the identification of the information to be gathered by that method should be based on the nature of the situation and the relevant research literature.

Level of detail

The level of detail required of an analysis of work is directly related to its intended use and the availability of information about the work. A less detailed analysis may be sufficient when there is already information descriptive of the work, and it may be appropriate when prior research about the job requirements allows the generation of sound hypotheses

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

concerning the predictors or criteria across job families or organizations. When a detailed analysis of work is not required, the testing professional should compile reasonable evidence establishing that the jobs in question are similar in terms of work behavior and/or required KSAOs or competencies, or fall into a group of jobs for which validity can be generalized. An example of situations that require a more detailed analysis of work may include one in which there is little existing work information available, and the organization intends to develop predictors assessing specific job knowledge. Any methods used to obtain information about work or workers should be understood by the participants and should have reasonable psychometric characteristics. Lack of consensus about the information contained in the analysis of work should be noted and considered further. Existing job descriptions or other documents may or may not serve the immediate research purpose; such information needs to be evaluated to determine its relevance and usefulness.

In some instances, an analysis of work may be the basis for developing selection procedures used to assign or select individuals for future jobs that do not exist at present. In other instances, an analysis of work may be used for transitioning workers from current to future work behaviors and activities. In either case, the future work behaviors and activities, as well as the worker requirements, may differ markedly from those that exist at present. Similarly, the work environment in which an organization operates may also change over time. For example, technology has permitted many individuals to work from virtual offices and also has replaced many functions that were previously conducted by individuals. Further, the global environment has expanded geographical boundaries and markets for many organizations. Procedures similar to those used to analyze current work requirements may be applicable for conducting an analysis of work in environments of rapid change; however, approaches that may be more responsive to the complexities of the emerging work environments are more appropriate (Levine & Oswald, 2012; Schneider & Konz, 1989). The central point in such instances is the need to obtain reliable and relevant job information that addresses anticipated behaviors, activities, and/or KSAOs or competencies.

If there is reason to question whether people with similar job titles are in fact doing similar work or if there is a problem of grouping jobs with similar complexity, attributes, behaviors, activities, or worker KSAOs or competencies, then the inclusion of incumbents or other subject matter experts (SMEs) from each of the job titles or families will generally be necessary. Even when incumbents are in positions with similar job titles or work families, studying multiple incumbents may be necessary to understand differences in work complexity, work context, work environment, job behaviors, or worker KSAOs or competencies as a function of shift, location, variations

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

in how work is performed, and other factors that may create differences in similar job titles or work families.

SOURCES OF VALIDITY EVIDENCE

Inferences made from the results of a selection procedure to the performance of subsequent work behavior or outcomes need to be based on evidence. As noted earlier, the *Standards* discusses five sources of evidence that can be used in evaluating a proposed interpretation of selection procedure scores for a particular use: (a) relationships between predictor scores and other variables (e.g., test–criterion relationships), (b) test content, (c) internal structure of the test, (d) response processes, and (e) consequences of testing. Given their relevance to selection practice, the first three sources of evidence will be described in more detail in this section. The generalization of validity evidence accumulated from existing research to the current employment situation is discussed in the "Generalizing Validity Evidence" section.

Evidence of Validity Based on Relationships with Measures of Other Variables

The *Principles* and the *Standards* view a construct as the attribute or characteristic a selection procedure measures. At times, the construct is not fully understood or well-articulated. However, relationships among variables are often assumed to reflect the relationships of their underlying constructs. For example, a predictor generally cannot correlate with a criterion unless there is some conceptual relationship between their respective constructs. Theoretically unrelated constructs may, however, correlate empirically with each other as a result of (a) the constructs having been measured with the same measurement method and/or (b) the constructs and/or measurement methods sharing the same extraneous contaminants. Consequently, all investigations of validity entail an evaluation of constructs to some degree.

Principles for using a criterion-related strategy to accumulate validity evidence in employment settings are elaborated below. Although not explicitly discussed, the following principles also apply to research using variables other than work performance criteria (e.g., turnover, accidents, theft). Some theory or rationale should guide the selection of these other variables as well as the interpretation of the study results.

Criterion-Related Evidence of Validity

Evidence for criterion-related validity typically consists of a demonstration of a relationship between the scores on a selection procedure (predictor) and one or more measures of work-relevant behavior or work outcomes (criteria). The choice of predictors and criteria should be based on an understanding of the objectives for predictor use, job information, and existing knowledge regarding test validity. Predictors are typically standardized procedures;

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

that is, they are consistent in administration, scoring, and interpretation. Because they reduce error variance and enhance reliability, standardized predictor measures and standardized criterion measures are preferred.

The discussion in this section, however, applies to all predictors and criteria, standardized or unstandardized.

Feasibility of a criterion-related validation study

The availability of appropriate criterion measures, the availability and representativeness of the research sample, and the adequacy of statistical power are very important in determining the feasibility of conducting a criterion-related study. Depending on their magnitude, deficiencies in any of these considerations can significantly weaken a criterion-related validation study.

A relevant, reliable, and uncontaminated criterion measure(s) is critically important. Of these characteristics, the most important is relevance. A relevant criterion is one that reflects the relative standing of employees with respect to an outcome critical to success in the focal work environment (e.g., job performance, employee turnover). To the extent that a job performance criterion does not reflect a representative sampling of work behaviors, then generalizations regarding predictor–job performance relationships should be qualified accordingly. If an adequate criterion measure does not exist or cannot be developed, use of a criterion-related validation strategy is not feasible.

A competent criterion-related validation study should be based on a sample that is reasonably representative of the workforce and candidate pool. Differences between the sample used for validation and a candidate pool on a given variable merit attention when credible research evidence exists demonstrating that the variable affects validity.

Statistical power influences the feasibility of conducting a criterion-related validation study. Prior to conducting the study, one should determine whether a large enough sample size can be obtained to meet the desired level of statistical power. The expected magnitude of the predictor–criterion relationship, the standard error of the statistic indexing that relationship (e.g., Pearson correlation, odds ratio, *d* statistic), and the probability level chosen for testing the significance of the chosen statistic (or forming a confidence interval around it), all factor into calculations of the sample size required to achieve a given level of power. Note that statistical artifacts, such as the degree of range restriction present in one's observed data, as well as criterion reliability, will also have implications for estimating required sample sizes. Correcting for these artifacts serves to increase the estimated validity coefficient but also increases the coefficient's standard error. Thus, when judging whether a sufficient sample size is available for a given validation study, care should be taken to differentiate between observed and corrected validity

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

coefficients when it comes to estimating power. If a study is ultimately conducted, the report documenting the validation study should describe procedures and results of relevant power analyses.

Design and conduct of criterion-related studies

If a criterion-related strategy is feasible, attention is then directed to the design of the study. A variety of designs can be identified. The traditional classification of predictive and concurrent criterion-related validity evidence is based on the presence or absence of a time lapse between the collection of predictor and criterion data. The employment status of the sample (incumbents or applicants) also may differentiate the designs. In predictive designs, data on the selection procedure are typically collected at or about the time individuals are selected. After a specified period of time (for retention criteria) or after employees' relative performance levels have stabilized (for performance criteria), criterion data are collected. In concurrent designs, the predictor and criterion data are typically collected on incumbents at approximately the same time.

There are, however, other differences between and within predictive and concurrent designs that can affect the interpretation of the results of criterion-related validation studies. Designs may differ with respect to the basis for the selection decision for participants in the research sample; they may have been selected using the predictor under study, an existing in-use predictor, a random procedure, or some combination of these. Designs also may differ with respect to the population sampled. For example, the design may use an applicant population or a population of recently hired employees, recent employees not yet fully trained, or employees with the full range of individual differences in experience.

The effect of the predictive or concurrent nature of the design on the observed validity may depend on the predictor construct. For tests of cognitive abilities that are expected to be stable over time, estimates of validity obtained from predictive and concurrent designs may be expected to be comparable (Barrett, Phillips, & Alexander, 1981; Bemis, 1968; Pearlman, Schmidt, & Hunter, 1980). In contrast, when dealing with self-report measures of noncognitive constructs (e.g., personality, interests, values, situational judgment) or experience-based measures (e.g., biodata), various factors can potentially lead to differences in validation results obtained from predictive and concurrent designs. For example, in a predictive validation design involving applicants, traditional self-report measures may be subject to faking (i.e., intentional distortion of responses with the goal of presenting a positive image; Ziegler, MacCann, & Roberts, 2012). To the extent such faking-related variance is unrelated to the criterion of interest, it would attenuate validity estimates relative to those based on a concurrent validation

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

design where motivation to fake is less salient. As another example, giving a biodata instrument to current employees may yield erroneous inferences if scores on the biodata instrument reflect the experiences of the employee on the current job. The use of a concurrent strategy requires the inference that scores on the predictor have not been influenced by experience on the current job. This is because the goal is to use predictor–criterion relationships based on incumbent data to estimate these relationships among applicants. Thus, findings from predictive and concurrent designs cannot be generalized automatically to all situations and to other types of predictors and criteria.

Occasionally, a selection procedure is designed for predicting higher-level work than that for which candidates are initially selected. Such higher-level work may be considered a target job in a criterion-related study if a substantial number of individuals who remain employed and available for advancement progress to the higher level within a reasonable period of time. Regardless of the number who advance to the higher level, assessment of candidates for such work may still be acceptable if the validity study is conducted using criteria that reflect performance at both the level of work that the candidate will be hired to perform and the higher level. The same logic may apply to situations in which people are rotated among jobs.

For some jobs in some organizations, successful performance is more closely related to abilities that contribute broadly to organizational effectiveness. In such instances, the testing professional may accumulate evidence in support of the relationship between predictor constructs (e.g., flexibility, adaptability, team orientation, learning speed, and capacity) and organization-wide rather than job-specific criteria (e.g., working collaboratively across business units).

Criterion development

In general, if criteria are chosen to represent work-related activities, behaviors, or outcomes, then the results of an analysis of work are helpful in criterion construction. If the goal of a given study is the prediction of organizational criteria such as tenure, absenteeism, or other types of organization-wide criteria, an in-depth work analysis is usually not necessary, although an understanding of the work and its context may be beneficial. Some considerations in criterion development follow.

Criteria should be chosen on the basis of work relevance, freedom from contamination, and reliability rather than availability or convenience. This implies that the purposes of the validation study are (a) clearly stated, (b) supportive of the organization's needs and purposes, and (c) acceptable in the social and legal context of the organization. The testing professional should

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

not use criterion measures that are unrelated to the purposes of the study to achieve the appearance of broad coverage.

Criterion relevance. Criteria should represent important organizational, team, or individual outcomes such as work-related behaviors, outputs, attitudes, or performance in training as indicated by a review of information about the work. Criteria need not be all-inclusive, but there should be a clear rationale linking the criteria to the proposed uses of the selection procedure. Criteria can be measures of overall or task-specific work performance, work behaviors, or work outcomes. Depending on the work being studied and the purposes of the validation study, various criteria, such as a standard work sample, behavioral and performance ratings, success in work- relevant training, turnover, contextual performance/organizational citizenship, or rate of advancement may be appropriate. Regardless of the measure used as a criterion, it is necessary to ensure its relevance to the work.

Criterion contamination. A criterion measure is contaminated to the extent that it includes extraneous, systematic variance. Examples of possible contaminating factors include differences in the quality of machinery, unequal sales territories, raters' knowledge of predictor scores, job tenure, shift, location of the job, and attitudes of raters. Conditions of evaluation may be another source of contamination. Employees who know that a supervisor is formally evaluating them may exhibit job performance based on maximal levels of motivation rather than typical levels of motivation. Although avoiding completely (or even knowing) all sources of contamination is impossible, efforts should be made to minimize their effects. For instance, standardizing the administration of the criterion measure minimizes one source of possible contamination. Measurement of some contaminating variables might enable the testing professional to statistically account for them, but in other cases, special diligence in the construction of the measurement procedure and its use may be all that can be done.

Criterion deficiency. A criterion measure is deficient to the extent that it excludes relevant, systematic variance. For example, a criterion measure intended as a measure of overall work performance would be deficient if it did not include all work behaviors or outcomes critical to work performance.

One common form of deficiency arises in practice when one limits the criterion measure to only those elements of the work performance domain theoretically expected to relate to a partial set of KSAOs/competencies measured by the selection battery, yet one desires to make inferences regarding relations between the scores on the selection battery and the full domain of performance on the job of interest. Under these circumstances, given the breadth of inference desired, the criterion measure used in the validation study should provide representative coverage of the full performance domain. To the extent the criterion measure used in the validation study does

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

not provide such coverage, testing professionals should be clear about what elements are omitted and the implications this has for supporting the desired inferences.

Criterion bias. Criterion bias is systematic error resulting from criterion contamination or deficiency that can differentially affect the criterion performance of different subgroups. The presence or absence of criterion bias cannot be detected from knowledge of criterion scores alone. A difference in criterion scores of older and younger employees or day and night shift workers could reflect bias in raters or differences in equipment or conditions, or the difference might reflect genuine differences in performance (or a combination of these factors). The possibility of criterion bias must be anticipated. The testing professional should protect against bias insofar as is feasible and use professional judgment when evaluating the data.

Criterion reliability. Criterion measures should exhibit adequate levels of reliability. When planning and conducting a criterion-related validation study, one should identify the conditions of measurement (e.g., raters, items, or occasions) across which one wishes to generalize the criterion scores of interest. To the extent possible, one should adopt a study design that will allow for calculation of reliability estimates that evaluate whether scores generalize across those conditions. In the event it is not possible to gather such data as part of the measure development or criterion-related validation effort, results regarding the reliability of scores should be qualified accordingly

The most appropriate estimate(s) of criterion reliability in a given study will depend on the measurement design underlying one's criterion measures, the conditions of measurement one wishes to generalize scores across, and the way in which the criterion measure will be used (Hunter & Schmidt, 1996; Putka & Hoffman, 2014; Putka & Sackett, 2010). When reporting estimates of criterion reliability, one should clearly describe the measurement design used and clarify what sources of error are reflected in the reported indices of reliability (e.g., rater-specific, item-specific, or occasion-specific errors).

Ratings as criteria. Among the most commonly used and generally appropriate measures of performance are ratings. If raters (supervisors, peers, self, clients, or others) are expected to evaluate several different aspects of performance, then the development of rating factors is ordinarily guided by an analysis of the work. Further, raters should be sufficiently familiar with the relevant demands of the work, as well as the individual to be rated, to effectively evaluate performance and should be trained in the observation and evaluation of work performance. Research suggests that performance ratings collected for research purposes are preferable for use in validation studies compared to those routinely collected for administrative use (Jawahar & Williams, 1997).

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Archival data as criteria. The growing prevalence of human resource information systems (HRISs) and other organizational data systems now make drawing on archival data as a potential source of criteria for use in validation studies increasingly viable. These archival criteria may reflect a variety of variables, such as turnover, disciplinary incidents, absenteeism, sales, customer service metrics, or engagement. Prior to using such archival data for analysis, one should take extra precautions to ensure the data are appropriate for the intended use (e.g., aligned with the work analysis, free from contamination, and acceptably reliable). In particular, the testing professional should seek to understand why the dataset exists and, if possible, test the accuracy of the archival data. Unlike data directly gathered by the team conducting the validation study, the quality of archival data is not often readily apparent. Issues surrounding the consistency of variable and value definitions over time and data owner confidence in the data are a few examples of important factors to consider. In addition, testing professionals should take into account data privacy and other policies and regulations governing the use of different types of archival data and try to identify unintended consequences of use.

Choice of predictor

Many factors, including critical KSAOs or competencies identified through work analyses, professional judgment, and the proposed use of the selection procedure, influence the choice of the predictor.

Selecting predictors. Variables chosen as predictors should have a theoretical, logical, or empirical foundation. The rationale for a choice of predictor(s) should be specified. A predictor is more likely to provide evidence of validity if there is good reason or theory to suppose that a relationship exists between it and the behavior it is designed to predict. A clear understanding of the work (e.g., via results of a work analysis), the research literature, or the logic of predictor development provides this rationale. This principle is not intended to rule out the application of serendipitous findings, but such findings, especially if based on small research samples, should be verified through replication with an independent sample.

Preliminary choices among predictors should be based on data and/or information about the target job (e.g., job descriptions, work analysis results) and the testing professional's scientific knowledge without regard for personal bias or prejudice. Therefore, the testing professional's choice of specific predictors should be based on theory and the findings of relevant research rather than personal interest or mere familiarity. Finally, in selecting predictors, it is important that testing professionals recognize the criticality of the distinction between the predictor construct (what is measured [e.g., general mental ability]) and the predictor method (how it is measured

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

[e.g., the interview]). Otherwise, confounded comparisons of predictors and method/construct comparisons (e.g., the comparative meta-analytic estimates of criterion-related validity of interviews and general mental ability tests) that are fundamentally uninterpretable absent further specification (Arthur & Villado, 2008) may result. However, comparisons of a specific interview and specific test in the same context are informative.

Predictor contamination. As with criteria, a predictor measure is contaminated to the extent that it includes extraneous variance. A number of factors can contribute to predictor contamination, such as unstandardized administrative procedures, use of irrelevant content, and applicant cheating or faking. Some procedures, such as unstructured interviews and unproctored internet tests, may be more susceptible than others to predictor contamination. Testing professionals should take steps to identify, assess, and mitigate sources of predictor contamination.

Predictor deficiency. Again, as with criteria, a predictor measure can be deficient. Predictor deficiency may manifest in two ways. The first involves deficiency in measuring a specific construct of interest (e.g., a stated intent to measure conscientiousness, but using a measure that only taps the orderliness facet of conscientiousness without tapping the industriousness facet). The second stems from not including all possible job-relevant determinants of a criterion of interest in a predictor set. Whereas the former is an issue of the psychometric quality of the predictor, the latter is rarely feasible and is often dictated by local circumstances and context. When judging whether the second form of deficiency is problematic, professional judgment that takes into account both psychometric and practical considerations, including systematic bias against subgroups, is required.

Predictors and selection decision strategies. Selection decisions based on human judges should be recognized as predictors. Decision makers who interpret and act upon predictor data interject something of themselves into the interpretive or decision-making process. Judgments or decisions thus may become at least an additional predictor or, in some instances, the only predictor. For example, if the decision strategy uses judgment to combine the scores from multiple predictors (e.g., standardized tests, reference checks, interview results) into a final selection decision, the actual predictor is the judgment reached by the person who weights and summarizes all the information. Ideally, it is this judgment that should be the focus of the validation effort. If this is not feasible, validity evidence for the specific components may be the best evidence available, although it is suggestive, rather than definitive, evidence of the validity of the judgment.

Scores produced by algorithms based on structured inputs (e.g., closed-ended assessment items) or unstructured inputs (e.g., resumes, open-ended text responses, or oral responses to stimuli) that are used to make selection

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

decisions should also be recognized as predictors. In cases where scores from such algorithms are used as part of the selection process, the conceptual and methodological basis for that use should be sufficiently documented to establish a clear rationale for linking the resulting scores to the criterion constructs of interest. In addition, when some form of empirical keying is used, clear evidence of cross-validity should be provided prior to operational use to guard against empirically driven algorithms' propensity to capitalize on chance. As is the case for all predictors, it is also important that algorithms do not introduce systematic bias against relevant subgroups.

Predictor reliability. The scores obtained from predictor measures should exhibit adequate levels of reliability. The factors critical to addressing the issue of reliability of criterion measures that were discussed earlier apply to predictor measures as well (e.g., identifying the conditions of measurement across which one wishes to generalize the scores of interest; adopting a study design that will allow for calculation of reliability estimates that evaluate whether scores generalize to the said conditions). Once again, in the event it is not possible to gather such data as part of the predictor development or criterion-related validation effort, results regarding the reliability of predictor scores should be qualified accordingly.

The estimates of predictor score reliability that are most appropriate in a given study will depend on the measurement design underlying one's predictor measures, the conditions of measurement one wishes to generalize scores across (e.g., raters, items, or occasions), and the ways in which the predictor measure will be used (e.g., for rank ordering applicants, or for making pass-fail or hire-no hire decisions; Haertel, 2006; Hunter & Schmidt, 1996; Putka & Sackett, 2010). When reporting estimates of predictor reliability, one should clearly describe the measurement design underlying the collection of data on which indices of reliability are being estimated and clarify the sources of error that are reflected in the reported indices of reliability.

Choice of participants

Validation samples should be chosen to be aligned with the selection situations to which they are intended to generalize. In part, this means ensuring that the validation sample represents relevant characteristics, such as demographics, motivation, ability, and experience. Convenience samples are discouraged to the extent they are deficient in these characteristics.

It is not feasible to investigate the validity of a test for all possible subgroups in employment testing. When there is credible evidence of potential bias, and sufficient data are available for analysis, determining whether bias exists requires the proper statistical analysis (e.g., differential validity or differential prediction analysis), along with large enough subsamples to detect practically meaningful differences wherever they might exist (i.e., have

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

adequate statistical power and precision). No matter how important a subsample may be to the testing professional, when it is too small, it cannot be statistically compared with other subsamples in an appropriate manner until additional data are available.

Data analysis for criterion-related validity

The quality of the validation study depends as much on the appropriateness of the data analysis as on the data collected during the research. Testing professionals need to ensure that the statistics used are appropriate. Moreover, as with the choice of criterion or predictor variables, the testing professional should not choose a data analysis method simply because the computer package for it is readily available. Testing professionals who delegate data analyses to others retain responsibility for ensuring the suitability and accuracy of the analyses.

Strength of the predictor–criterion relationship. The analysis should provide information about effect sizes and the statistical significance associated with predictor–criterion relationships, along with standard errors or confidence intervals for those relationships. Effect size estimates are useful in making professional judgments about the strength of predictor–criterion relationships (Schmidt, 1996), and standard errors and confidence intervals provide key information on uncertainty in the estimated relationships. Although methods exist for testing the statistical significance of validity estimates and estimating standard errors or confidence intervals, the scientific literature is still evolving with regard to significance testing and estimates of uncertainty for validities, including those that have been corrected for statistical artifacts.

Research on the power of criterion-related validation studies and meta-analytic research suggests that achieving adequate power while simultaneously controlling Type I error rates can be problematic in a local validation study and may require sample sizes that are difficult to obtain. Testing professionals should give at least equal attention to the risks of Type II error.

Reports of any analysis should include the number of cases and the characteristics of distributions of predictor and criterion variables (e.g., central tendency, variance), as well as point estimates and standard errors or confidence intervals for interrelationships among all variables studied.

Adjustments to validity estimates. Testing professionals should obtain as unbiased an estimate as possible of the operational validity of the predictor in the population in which it is used. Observed validity coefficients may misestimate their respective predictor–criterion relationships due to the effects of range restriction and criterion unreliability. When range restriction distorts validity coefficients, a suitable bivariate or multi-variate adjustment should be made when the necessary information is available (e.g.,

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Beatty, Barratt, Berry, & Sackett, 2014; Sackett & Yang, 2000; Schmidt, Oh, & Le, 2006). Adjustment of the validity coefficient for criterion unreliability should be made if an appropriate estimate of criterion reliability can be obtained. Testing professionals should make sure that reliability estimates used in making corrections are appropriate to avoid under or overestimating validity coefficients. For example, in a study utilizing a criterion-related strategy in which the criteria are performance ratings, differences between raters and differences across time may be considered in estimating criterion reliability because internal consistency estimates, by themselves, may be inadequate.

In theory, criterion reliability places a ceiling on validity estimates. Thus, the effect of criterion unreliability is to underestimate criterion-related validity in the population of interest. In practice, particularly for ratings-based criterion measures, observed reliability may not necessarily limit observed validity in one's research sample. Specifically, corrections for attenuation are premised on the assumption that rater-specific variance (given the one or two raters that are typically available to rate each job incumbent) is uncorrelated with the predictor of interest. To the extent this assumption does not hold, then observed validity may not be limited by observed reliability, and it is possible for corrected validities to overestimate true validities. Given this uncertainty, and given open debate regarding this issue in the scientific literature (e.g., Murphy & DeShon, 2000; Putka, Hoffman, & Carter, 2014; Schmidt, Viswesvaran, & Ones, 2000), it is best to provide both corrected and uncorrected estimates of criterion-related validity.

If assumptions underlying adjustment procedures are met, the adjusted coefficient is generally the best point estimate of the population validity coefficient. However, testing professionals should be cautious about implying that corrected correlation coefficients are statistically significant because the usual tests of statistical significance and standard error or confidence intervals for unadjusted coefficients do not apply to adjusted coefficients such as those adjusted for restriction of range and/or criterion unreliability. Procedures for testing the significance of validity coefficients that have been corrected for direct range restriction and/or criterion unreliability, as well as providing standard errors and confidence intervals for them, are described in a variety of sources (e.g., Bobko, 1983; Raju & Brand, 2003). Procedures for establishing standard errors and confidence intervals for coefficients corrected for indirect range restriction have also started to emerge (e.g., Fife, Mendoza, & Terry, 2013; Li, Chan, & Cui, 2011). No adjustment of a validity coefficient for unreliability of the predictor should be made or reported unless it is clearly stated that the coefficient is theoretical and cannot be interpreted as reflecting the actual operational validity of the selection procedure.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Combining predictors and combining criteria. When predictors are used in combination, testing professionals should consider and document the method of combination. Predictors can be combined using weights derived from multiple methods, including a multiple regression analysis (or another appropriate multivariate technique), weights based on Pareto optimization (DeCorte, Lievens, & Sackett, 2011), unit weights (Bobko, Roth, & Buster, 2007), empirical weights not fully optimized against the criterion (e.g., rounded regression weights, correlation weights), or rational weights (e.g., determined from work-analytic procedures or based on professional judgment).

When combining scores, care must be taken to ensure that differences in the variances and covariances among different predictors do not lead to unintentional over- or underweighting of one or more predictors (Oswald, Putka, & Ock, 2015). When measures are combined, testing professionals should recognize that effective weights (i.e., the contributions of individual measures to the variance of the composite) are a function of the variances and covariances among variables in the composite and are unlikely to be the same as the nominal weights (i.e., the observed weight assigned to a given variable). Particular caution should be taken when predictors in one's validation study are differentially impacted by range restriction, as the predictor variances and covariances pertinent to weighting may differ greatly when unrestricted (Sackett, Lievens, Berry, & Landers, 2007).

In addition to being dependent on the weighting strategies noted above, both the validity of predictor information and the rank ordering of candidates based on a selection process involving multiple predictors will depend on whether predictor information is combined in a compensatory or noncompensatory manner (e.g., as part of a process involving different cutoff scores for individual predictors, or involving a multiple hurdle or staged selection process; DeCorte et al., 2011; Finch, Edwards, & Wallace, 2009). Testing professionals should be cognizant of the implications that weighting and sequencing choices have for the expected mean standing on the criterion of interest for those selected (e.g., expected mean job performance, expected mean turnover rate) and any anticipated subgroup differences on the predictor composite.

Regardless of whether a compensatory or noncompensatory combination of predictor measures is used, a clear rationale for the combination, ultimately used should be provided (e.g., meeting larger organizational goals or needs, administrative convenience, organizational values, reduced testing costs, or balancing potential tradeoffs between validity and subgroup differences).

Similarly, if the testing professional combines scores from several criteria into a composite, there should be a rationale to support the rules of

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

combination, and the rules of combination should be described. As was the case with predictors, testing professionals should recognize that the effective weights of each component of a criterion composite are a function of those components' variances and covariances, and they are not simply a function of the nominal weights assigned to the components by the testing professional.

Cross-validation. Testing professionals should guard against overestimates of validity resulting from capitalization on chance, especially when the research sample is small. Estimates of the validity of a composite battery developed on the basis of a regression equation should be adjusted using the appropriate shrinkage formula or be cross-validated on another sample (Schmitt & Ployhart, 1999). If the weights assigned to predictors are not based on regression analyses but are still informed by relations between predictors and criteria in the research sample (e.g., correlation-based weights), the resulting validities for the composite battery will be inflated, and cross-validity estimates should be provided. Additionally, if the final selection or scoring of items for a given predictor measure is based on items' observed relations with the criterion in the research sample, then the resulting validities for the predictor measure will be inflated, and cross-validity estimates should be provided. Rational or unit weights are both independent of the data set; therefore, assigning these kinds of weights to predictors does not result in shrinkage of validity estimates.

Interpreting validation analyses. Results obtained using a criterion-related strategy should be interpreted against the background of the relevant research literature. Cumulative research knowledge plays an important role in any validation effort. A large body of research regarding relationships between many predictors and work performance currently exists (e.g., Arthur, Day, McNelly, & Edens, 2003; Christian, Edwards, & Bradley, 2010; Huffcutt, Conway, Roth, & Stone, 2001; Sackett, & Walmsley, 2014; Schmidt & Hunter, 1998).

An extremely large sample or replication is required to give full credence to unusual findings. Such findings include, but are not limited to, suppressor or moderator effects, nonlinear regression results, and benefits of configural scoring. Post hoc hypotheses in multivariate studies and differential weightings of highly correlated predictors are particularly suspect and should be replicated before they are accepted and results implemented.

Evidence for Validity Based on Content

Evidence for validity based on content typically consists of a demonstration of a strong linkage between the content of the selection procedure and important work behaviors, activities, worker requirements, or outcomes on the job. This linkage also supports construct interpretation. When the selection

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

procedure is designed explicitly as a sample of important elements in the work domain, the validation study should provide evidence that the selection procedure samples the important work behaviors, activities, and/or worker KSAOs necessary for performance on the job, in job training, or on specified aspects of either. This provides the rationale for the generalization of the results from the validation study to prediction of work behaviors (Goldstein, Zedeck, & Schneider, 1993). Comments and a critical consideration of the usefulness of content evidence as part of the validation process are provided in an issue of *Industrial and Organizational Psychology* (see Murphy, 2009; Sackett, 2009b). Stelly and Goldstein (2007) have also considered the importance of test content examinations as indicators that a measure represents a theoretical construct.

The selection procedures discussed here are those designed as samples of important work behaviors, activities, and/or worker KSAOs drawn from the work domain and defined by the analysis of work; these selection procedures are labeled "content-based predictors." The content of the selection procedure includes the questions; tasks; themes, format, wording, and meaning of items; response formats; instructions; and guidelines regarding the administration and scoring of the selection procedure. The following provides guidance for the development or choice of procedures based primarily on content.

Feasibility of a content-based validation study

A number of issues may affect the feasibility of a content-based validation study and should be evaluated before beginning such a study. Among these issues are the stability of the work and the worker requirements, interference of irrelevant content, availability of qualified and unbiased SMEs, and cost and time constraints.

The testing professional should consider whether the work and the worker requirements are reasonably stable and take appropriate steps to define them when a question arises. When feasible, a content-based selection procedure should remove or minimize content that is irrelevant to the domain sampled. Virtually any content-based procedure includes some elements that are not part of the work domain (e.g., standardization of the selection procedure or use of response formats that are not part of the job content, such as multiple-choice formats or written responses when the job does not require writing).

The success of a content-based validation study is closely related to the qualifications of the SMEs. SMEs define the work domain; participate in the analysis of work by identifying the important work behaviors, activities, and worker KSAOs; and establish the relationship between the selection procedures and the work behaviors or worker requirements. The experts should

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

be competent to perform the task set before them. For example, those who evaluate the job or the worker requirements should have thorough knowledge of the work behaviors and activities, responsibilities of the job incumbents, and/or the KSAOs prerequisite to effective performance on the job. For the task of defining work behaviors, the SMEs should include persons who are fully knowledgeable about relevant organizational characteristics such as shift, location, type of equipment used, software and hardware, and so forth. A method for translating SME judgments into the selection procedure should be selected or developed and documented. If SME ratings are used to evaluate the match of the content-based procedure to the work and worker requirements, then procedures and criteria for rating each aspect should be standardized and delineated.

Cost and time constraints can affect the feasibility and the fidelity of some content-based procedures. In some situations, designing and implementing a simulation that replicates the work setting or type of work may be too costly. Even when a content-based procedure is feasible, cost and time constraints may affect the fidelity of the procedure. In these instances, the testing professional must use judgment to determine whether the fidelity of the selection procedure is sufficient for the organization's purposes.

Design and conduct of content-based strategies

The content-based validation study specifically demonstrates that the content of the selection procedure represents an adequate sample of the important work behaviors, activities, and/or worker KSAOs defined by the analysis of work. In addition to choosing appropriate SMEs, other steps in this process include defining the content to be included in the selection procedure, developing the selection procedure, collecting SME judgments about the link between the selection procedure and the requirements of the job, establishing the guidelines for administration and scoring, and evaluating the effectiveness of the validation effort.

Defining the content domain

The characterization of the work domain should be based on accurate and thorough information about the work, including analysis of work behaviors and activities, responsibilities of the job incumbents, and/or the KSAOs prerequisite to effective performance on the job. In addition, definition of the content to be included in the domain is based on an understanding of the work and may consider organizational needs, labor markets, and other factors that are relevant to personnel specifications and relevant to the organization's purposes. The domain need not include everything that is done on the job. The testing professional should indicate what important work behaviors, activities, and worker KSAOs are included in the domain, describe how the

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

content of the work domain is linked to the selection procedure, and explain why certain parts of the domain were or were not included in the selection procedure.

The fact that the construct assessed by a selection procedure is labeled an ability or personality characteristic does not per se preclude the reliance on a content-oriented strategy. When selection procedure content is linked to job content, content-oriented strategies are useful. When selection procedure content is less clearly linked to job content, other sources of validity evidence take precedence.

The selection procedure content should be based on an analysis of work that specifies whether the employee is expected to be able to perform all of the important work behaviors and activities and/or to possess all of the relevant KSAOs before selection into the job, or whether basic or advanced training will be provided to employees after selection to develop additional performance capabilities and KSAOs. If the intended purpose of the selection procedure is to hire or promote individuals into jobs for which no advanced training is provided, the testing professional should define the selection procedure in terms of the work behaviors, activities, and/or KSAOs an employee is expected to have before placement on the job. If the intent of the content-based procedure is to select individuals for a training program, the work behaviors, activities, and/or worker KSAOs should include those needed to succeed in the training program. Because the intended purpose is to hire or promote individuals who are able to perform the prerequisite work behaviors and activities and/or who possess KSAOs to learn the work as well as to perform the work, the selection procedure should be based on an analysis of work that defines the balance between the work behaviors, activities, and/or KSAOs the applicant is expected to have before placement on the job and the amount of training the organization will provide. For example, the fact that an employee will be taught to interpret company technical manuals may mean that the job applicant should be evaluated for reading ability. A selection procedure that assesses the individual's ability to read at a level required for understanding the technical manuals would likely be predictive of work performance that is dependent upon interpreting company technical manuals.

A content-based selection procedure may also include evidence of specific prior training, experience, or achievement. This evidence is judged on the basis of the relationship between the content of the experience and the content of the work requiring that experience. To justify such relationships, more than a superficial resemblance between the content of the experience variables and the content of the work is required (Buster, Roth, & Bobko, 2005). For example, course titles and job titles may not give an adequate

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

indication of the content of the course or the job or the level of proficiency an applicant has developed in some important area. What should be evaluated is the similarity between the behaviors, activities, processes performed, or the KSAOs required by the work.

Developing or choosing the selection procedure

The content of a content-based selection procedure is usually restricted to important or frequent behaviors and activities or to prerequisite KSAOs. The selection procedure should reflect adequate coverage of work behaviors and activities and/or worker requirements from this restricted domain to provide sufficient evidence to support the validity of the inference. The fidelity of the selection procedure content to important work behaviors forms the basis for the inference.

Sampling the content domain. The process of constructing or choosing the selection procedure requires sampling the work content domain. Not every element of the work domain needs to be assessed. Rather, a sample of the work behaviors, activities, and worker KSAOs can provide a good estimate of the predicted work performance. Sampling should have a rationale based on the professional judgment of the testing professional and an analysis of work that details important work behaviors and activities, important components of the work context, and KSAOs needed to perform the work. Random sampling of the content of the work domain is usually not feasible or appropriate. Instead, the selection procedure might measure the most important work behaviors or KSAOs or a few that are prerequisite to others or a smaller set of KSAOs used to predict a subset of critical work out- comes (e.g., accidents, turnover). The rationale underlying the sampling should be documented in a test plan specifying which KSAOs are to be measured by which assessment methods.

Describing the level of specificity. In defining the work content domain, the degree of specificity needed in a work analysis and a selection procedure should be described in advance. The more fidelity a selection procedure has with exact job components, the more likely it is that a satisfactory level of content-based evidence will be demonstrated. However, when the work changes and fidelity drops, the selection procedure is less likely to remain appropriate. Thus, considering the extent to which the work is likely to change is important. If changes are likely to be frequent, then the testing professional may wish to develop a selection procedure that has less specificity. For example, in developing a selection procedure for a job involving the preparation of electronic documents, the procedure may exclude content such as demonstrating proficiency with a particular software program and instead include content that is less specific, such as demonstrating proficiency with software program principles and techniques.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

The degree to which the results of content-based validation studies can be generalized depends in part on the specificity of the selection procedure and its applicability across settings, time, and jobs. Although general measures may be more resilient to work changes and more transferable to other, similar work, they also may be subject to more scrutiny because the correspondence between the measure and the work content is less detailed. At times, a reanalysis of the work in the new setting may be useful in determining whether the selection tools link to the contemporary work content domains.

Competency modeling. Many organizations use a competency model to organize and integrate various aspects of their human resource efforts (e.g., training, selection, compensation). These models can be useful in several ways, such as allowing the organization to standardize language and effort across processes and organizational units and to express aspirational human capability goals. The competency model may also direct effort toward the KSAOs that ought to be considered in a valid selection program. A rigorous competency modeling study could be the foundation for content-oriented selection procedure research, just as a rigorous traditional work analysis project could be the foundation for content-oriented selection procedure research. The developer of the selection procedure must determine if the competency model is detailed and rigorous enough to serve as the foundation for a content validation study. See Campion, Fink, Ruggeberg, Carr, Phillips, and Odman (2011) for best practices in competency modeling.

Procedural considerations

The testing professional needs to establish the guidelines for administering and scoring the content-based procedure. Typically, defining the administration and scoring guidelines for a paper-based procedure that measures job-related knowledge or cognitive skills is relatively uncomplicated; however, a content-based selection procedure that includes work behaviors or activities may pose administration and scoring challenges, which should be evaluated in advance. Generally, the more closely a selection procedure replicates a work behavior, the more accurate the content-based inference. At the same time, the more closely a selection procedure replicates a work behavior, the more difficult the procedure may be to administer and score.

For example, troubleshooting multistep computer problems may be an important part of a technical support person's work. It may be difficult, however, to develop and score a multistep troubleshooting simulation or work sample, because examinees may not use the same steps or strategy when attempting to solve the problem. A lower–fidelity alternative such as single-step problems could be used so that important aspects of the work domain are still included in the selection procedure. In all cases, the testing

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

professional should ensure that the procedures are measuring skills and knowledge that are important in the work, rather than irrelevant content.

Evaluating content-related evidence

Evidence for validity based on content rests on demonstrating that the selection procedure adequately samples and is linked to the important work behaviors, activities, and/or worker KSAOs defined by the analysis of work. The documented methods used in developing the selection procedure constitute the primary evidence for the inference that scores from the selection procedure can be generalized to the work behaviors and can be interpreted in terms of predicted work performance. The sufficiency of the match between selection procedure and work domain is a matter of professional judgment based on evidence collected in the validation effort (Goldstein et al., 1993).

Reliability of performance on content-based selection procedures should be determined when feasible. The type of reliability estimate reported should reflect consideration of the measurement design underlying one's selection procedure, the generalizations one wishes to make regarding the resulting scores, and how the predictor measure will be used (e.g., for rank ordering applicants, or for making pass–fail or hire–no hire decisions; cf. Predictor reliability).

Evidence of Validity Based on Internal Structure

Information about the internal structure of any selection procedure can also support validation arguments. Internal structure evidence alone is not sufficient evidence to establish the usefulness of a selection procedure in predicting future work performance; however, internal structure is important in planning the development of a selection procedure. The specific analyses that are relevant depend on the conceptual framework of the selection procedure, which in turn is typically established by the proposed use of the procedure.

When evidence of validity is based on internal structure, the testing professional may consider the relationships among items, components of the selection procedures, or scales measuring constructs. Inclusion of items in a selection procedure should be based primarily on their relevance to a construct or content domain and secondarily on their intercorrelations. Well-constructed components or scales that have near-zero correlations with other components or scales, or a total score, should not necessarily be eliminated. For example, if the selection procedure purposely contains components relevant to different construct or content domains (e.g., a selection battery composed of a reading test, an in-basket, and an interview), the scores on these components may not be highly correlated.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

However, if the conceptual framework posits a single dimension or construct, one should offer evidence that covariances among components are accounted for by a strong single factor. If the intent of the conceptual framework requires more complex internal structure (e.g., hypothesized multifactor measure), a careful examination of the degree of dimensionality should be undertaken. In the latter case, overall internal consistency might not be an appropriate measure. For example, a lengthy multi-item measure that actually reflects several dimensions may have a high degree of internal consistency as measured by coefficient alpha simply because of the number of items (Cortina, 1993), or a performance rating form with several theoretically unrelated scales may display a high degree of internal consistency because of halo effect.

GENERALIZING VALIDITY EVIDENCE

Depending on the context and purpose of a selection procedure, sufficient accumulated validity evidence may be available to justify the appropriateness of applying a selection system in a new setting without conducting a local validation research study. In these instances, use of the selection procedure may be based on a demonstration of the generalized validity inferences from that selection procedure, coupled with a compelling argument for its direct applicability to the current selection situation. Although neither mutually exclusive nor exhaustive, several strategies for generalizing validity evidence have been delineated in the organizational research literature (Hoffman & McPhail, 1998): (a) transportability, (b) synthetic validity/job component validity, and (c) meta-analytic validity generalization.

Transportability

One approach to generalizing the validity of inferences from scores on a selection procedure involves the use of a specific selection procedure in a new situation, based on results of a validation research study conducted elsewhere. When these research findings are determined to be applicable to a current selection situation due to a preponderance of key observable and/or underlying similarities with other validity evidence, this is referred to as demonstrating the "transportability" of that evidence. When evaluating whether to "transport" the use of a specific selection procedure, a careful review of the original validation study is warranted to ensure the technical soundness of that study and to determine its conceptual and empirical relevance to the new situation. At a broad level, comparability in terms of job content or job requirements, job context, and job applicant group (if feasible) should be considered when determining the appropriateness of transportability in a particular situation (Hoffman, Rashovsky, & D'Egidio, 2007; Johnson, 2007).

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Synthetic Validity/Job Component Validity

A second approach to establishing generalization of the validity of inferences based on scores from a selection procedure is referred to either as synthetic validity or job component validity. Although some testing professionals distinguish these terms, others do not, and in either case several variations on each exist.

A defining feature of synthetic validity/job component validity is the justification of the use of a selection procedure based upon the demonstrated validity of inferences from scores on the selection procedure with respect to one or more domains of work (job components). If this relationship is well established, then the validity of the selection procedure for that job component, when coupled with other relevant information, may lead to the professional judgment that the selection procedure is generalizable and therefore applicable to other selection settings in which the job components are comparable.

The validity of a selection procedure may be established with respect to a range of relevant components of work, then "synthesized" (empirically combined) for use for a given job or job family based on those particular components of work that are deemed relevant through a job analysis (see Johnson & Carter, 2010; Steel, Huffcutt, & Kammeyer-Mueller, 2006). In some instances, this process may involve conducting a research study designed to demonstrate evidence for the generalized validity of inferences from scores on a set of selection procedures and then using various subsets of these procedures for selection into both jobs or job families in the original study, as well as into other jobs or job families. In other cases, it may involve generalizing the validity of inferences based on scores on selection procedures examined in one or more research studies conducted elsewhere to the new situation. In both cases, detailed analysis of the work (e.g., a job analysis or work analysis) is required for the use of this strategy of generalizing validity evidence. When many jobs share common job components, the synthetic validity approach may provide a source of validity evidence that is not feasible in each criterion-related validity study conducted for each job; and synthetic validity may help reduce burdensome data collection efforts that impede many local validation efforts. However, under the synthetic validity approach, those job requirements specific to a job that necessitate unique KSAOs for performance may not be sufficiently evaluated and thus require other sources of evidence (Johnson, Steel, Scherbaum, Hoffman, & Jeanneret, 2010; Sackett, Putka, & McCloy, 2012).

Meta-Analysis

Meta-analysis is a third procedure and strategy that can be used to determine the degree to which predictor–criterion relationships are specific to

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

the situations in which the validity data have been gathered or are generalizable to other situations, as well as to determine various factors that predict cross-situation variability. Meta- analysis requires the accumulation of empirical findings across an appropriately determined set of validity studies to determine the most accurate summary estimates of the predictor–criterion relationship for the kinds of work domains and settings included in the studies.

Meta-analysis is a strategy that is applied in cases where multiple original studies relied upon criterion-related evidence of validity**.** The question to be answered using a meta-analytic strategy is the extent to which valid inferences about work behavior or job performance can be drawn from predictor scores across given jobs or job families in different settings. (Note that the focus here is on using meta-analysis to examine predictor–criterion relationships. Meta-analysis also can be used to examine other issues relevant to selection, such as convergence among instruments intended to measure the same construct or mean differences between subgroups.)

Meta-analysis is the basis for the technique that is often referred to as "validity generalization." In general, research has shown that meaningful amounts of variation in observed differences in obtained validity coefficients in different situations can be attributed to sampling error variance, direct or incidental range restriction, and other statistical artifacts (Ackerman & Humphreys, 1990; Barrick & Mount, 1991; Callender & Osburn, 1980; 1981; Hartigan & Wigdor, 1989; Hunter & Hunter, 1984; Schmidt & Hunter, 2015; Schmidt, Hunter, & Pearlman, 1981). These findings are particularly well-established for cognitive ability tests, and research results also are accruing that indicate the generalizability of predictor–criterion relationships involving noncognitive constructs in employment settings (Berry, Ones, & Sackett, 2007; Chiaburu, Oh, Berry, Li, & Gardener, 2011; Hurtz & Donovan, 2000; Judge, Rodell, Klinger, Simon, & Crawford, 2013). Professional judgment in interpreting and applying the results of meta-analytic research is important. Testing professionals should consider the meta-analytic methods used and their underlying assumptions, the tenability of the assumptions, and statistical artifacts that may influence or bias the results (Bobko & Stone-Romero, 1998; Raju, Anselmi, Goodman, & Thomas, 1998; Raju et al. 1991; Raju, Pappas, & Williams, 1989). In evaluating meta-analytic evidence, the testing professional should be concerned with potential moderators to the extent that such moderators would affect conclusions about the presence and generalizability of validity (Aguinis & Pierce, 1998). Whenever a given meta-analysis has investigated a substantive moderator of interest, testing professionals should consider both statistical power to detect the moderator effect and the precision of the reported effects (Aguinis, Sturman, & Pierce, 2008; Oswald & Johnson, 1998; Steel & Kammeyer-Mueller, 2002).

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Reporting all critical aspects of the conduct of a meta-analysis is important, just as it is with the conduct of individual studies. Reports and results contributing to the meta-analysis should be clearly identified and should be available whenever possible so that any critical consumer can determine their appropriateness. Testing professionals should consult the relevant literature to ensure that the meta-analytic strategies used are current, sound, and properly applied; the appropriate procedures for estimating predictor–criterion relationships on the basis of cumulative evidence are followed; the conditions for the application of meta-analytic results are met; and the application of meta-analytic conclusions is appropriate for the work and settings studied. The rules by which the testing professionals categorize the work and jobs studied, the selection procedures used, the job performance criteria used, and other study characteristics that are hypothesized to impact the study results should be fully reported (Appelbaum et al., 2018; Aytug, Rothstein, Zhou, & Kern, 2012; Guion, 2011). Experts who meta-analyze the same domain of studies can reach somewhat different results and interpretations (see Nieminen, Nicklin, McClure, & Chakrabarti, 2011, who compare different experts' meta-analyses in a domain; see Van Iddekinge, Roth, Raymark, & Odle-Dusseau, 2012, who summarize an exchange on their meta-analysis in the integrity domain). Conversely, missing or unreported information relevant to a meta-analysis will compromise the quality and integrity of the results and, therefore, the inferences that can be made from them (e.g., when effect sizes are unobtainable from a testing professional, or when effect sizes are available, but critical study information may not be reported due to proprietary issues).

Note that sole reliance upon available cumulative evidence may not be sufficient to meet specific employer operational needs, such as for the placement of employees or for the optimal combination of procedures within a broader employment system that includes recruitment, selection, placement, and training. Consequently, additional studies, including evidence from meta-analytic studies and cooperative studies across organizations, may also be informative to meet these specific operational needs.

Meta-analytic methods for demonstrating generalized validity are still evolving (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cheung, 2015; Schmidt & Hunter, 2015; Tipton & Pustejovsky, 2015). Testing professionals should be aware of continuing research and critiques that may provide further refinement of meta-analytic techniques as well as a broader range of predictor–criterion relationships to which meta-analysis has been applied.

Generalizing validity evidence from meta-analytic results can often be more useful than making similar generalizations from a single study. However, if important conditions in the operational setting are not represented in the meta-analysis (e.g., the local setting involves a managerial job and the

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

meta-analytic database is limited to entry-level jobs), a local individual study may be more relevant than the average predictor–criterion relationship reported in a meta-analytic study. Specifically, a competently conducted study, with a large and organizationally relevant sample that uses the same test, for the same kind of work activities may be more accurate, informative, and useful than an accumulation of small validation studies that are highly heterogeneous, homogeneous but markedly deficient, or otherwise not representative of the setting to which one seeks to generalize validity. A Bayesian approach to meta-analysis balances the validity information from a meta-analysis with locally estimated validity coefficients in a statistically defined manner (Newman, Jacobs, & Bartram, 2007).

Reliance on meta-analytic results is more straightforward when results are organized around relevant predictor and criterion constructs. When different measures of predictors are correlated with different measures of criteria in a meta-analysis, findings are meaningful to the extent that predictor and criterion measures correlate highly within their respective constructs (e.g., predictor measures correlate highly with other measures of the same purported construct) and are generally higher than the criterion-related validities. The particular predictor and criterion measures involved in the meta-analysis cannot be assumed to be the same as other measures that happen to use the same construct labels without additional rational and empirical evidence that those other measures indeed reflect the same construct.

When studies are cumulated on the basis of common measurement methods (e.g., interviews, biodata, situational judgment tests) or mode (e.g., web-based, paper-and-pencil-based, video-based) instead of predictor and criterion constructs, a unique set of interpretational difficulties arises (Arthur & Villado, 2008). Generalization can be relatively straightforward when, for example, an empirical biodata scale has been developed for a specific occupation, multiple validity studies have been conducted using that scale in that occupation, and the intent is to generalize to another setting that employs individuals in that same occupation. By contrast, testing professionals may have great difficulty generalizing broadly about biodata, interviews, situational judgment tests, or any other method. Because methods such as the interview can be designed to assess widely varying constructs (e.g., job knowledge, integrity, teamwork), generalizing from cumulative findings is only possible if the features of the method that result in method–criterion relationships 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.

Meta-analyses vary in the degree to which the studies included specify the content and scoring of the procedure, the extent of the structure, the

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

setting in which the selection procedure is used, and so on. Generalizing from meta-analytic results based on one set of procedures and one setting to a new situation in which different selection procedures or settings are used but not specified is not warranted. For example, if all studies in the database involve interviews that are focused on technical knowledge, then any results from meta-analysis about validity do not support generalization to interviews that are focused on interpersonal skills. In contrast, generalization could be supported by a cumulative database that codes and meta-analytically compares interviews based on their technical and interpersonal content and their structure and scoring, so long as inferences are to interviews that meet the same specifications.

FAIRNESS AND BIAS

Fairness

Fairness is a social rather than a psychometric concept. Its definition depends on what one considers to be fair. Fairness has no single meaning and, therefore, no single definition, whether statistical, psychometric, or social. The *Standards* notes a number of possible meanings of "fairness."

One meaning views fairness as requiring equal group outcomes (e.g., equal passing rates for subgroups of interest). The *Standards* rejects this definition and notes that, although group differences should trigger heightened scrutiny for possible sources of bias (i.e., construct underrepresentation or construct irrelevant components that differentially affect the performance of different groups of test takers), outcome differences in and of themselves do not indicate bias.

Another meaning views fairness in terms of the equitable treatment of all examinees during the selection process. Equitable treatment in terms of testing conditions, access to practice materials, performance feedback, retest opportunities, and other features of test administration, including providing reasonable accommodation for test takers with disabilities when appropriate, all exemplify important aspects of fairness under this perspective. Conditions related to mode of administration may be particularly important to consider given recent technological advances (e.g., testing via computers, laptops, tablets, and other mobile devices such as smartphones).

A third meaning views fairness as requiring that examinees have comparable access to the constructs measured by a selection procedure. Accessible testing situations enable all test takers to show their status on a construct without being unduly advantaged or disadvantaged by other individual characteristics. Under this view, it may be particularly important to consider whether factors such as age, race, ethnicity, gender, socioeconomic status, cultural background, disability, and language proficiency restrict accessibility and affect measurement of the construct of interest.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Another meaning views fairness as a lack of bias. One form of bias is measurement bias, which is discussed below. In the employment context, research generally focuses on evaluating predictive bias, and this approach views predictor use as fair if a common regression line can be used to describe the predictor–criterion relationship for all subgroups of interest. Subgroup differences in regression slopes or intercepts may signal predictive bias. There is broad scientific agreement on this definition of bias, but there is no similar broad agreement that the lack of bias can be equated with fairness. For example, a selection system might exhibit no predictive bias by race or sex but still be viewed as unfair if equitable treatment (e.g., access to practice materials) was not provided to all examinees.

In summary, there are multiple perspectives on fairness. There is agreement that issues of equitable treatment, access, bias, and scrutiny for possible bias when subgroup differences are observed are important concerns in personnel selection. Most organizations strive for a diverse and inclusive workforce and equitable treatment of cultural and linguistic minorities. There is not, however, agreement that the term "fairness" can be uniquely defined in terms of any of these issues.

Bias

The *Standards* notes that *bias* refers to systematic error in a test score that differentially affects the performance of different groups of test takers. The effect of irrelevant sources of variance on scores on a given variable is referred to as *measurement bias*, whereas the effects of irrelevant sources of variance on predictor–criterion relationships, such that slope or intercepts of the regression line relating the predictor to the criterion are different for one group than for another, is referred to as *predictive bias*. Both forms of bias are discussed below.

Predictive bias

Although fairness has no single accepted meaning, there is agreement as to the meaning of predictive bias. There is also agreement on the importance of testing for and avoiding predictive bias against subgroups of interest in employee selection. Predictive bias is found when, for a given subgroup, systematic nonzero errors of prediction are made for members of the subgroup (Cleary, 1968; Humphreys, 1952). Another term used to describe this phenomenon is differential prediction. The term "differential prediction" is sometimes used in the classification and placement literature to refer to differences in predicted performance when an individual is classified into one condition rather than into another; this usage should not be confused with the use of the term here to refer to predictive bias. Although other definitions of bias have been introduced, such models have been critiqued and found

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

wanting on grounds such as lack of internal consistency (Petersen & Novick, 1976).

Testing for predictive bias involves using moderated multiple regression, where the criterion measure is regressed on the predictor score, subgroup membership, and an interaction term between the two. Slope and/or intercept differences between subgroups indicate predictive bias (Berry & Zhao, 2015). The *Standards* notes that the moderated multiple regression approach is more appropriate than the use of separate subgroup correlation coefficients in evaluating predictive bias hypotheses, which is generally consistent with research recommendations (Berry, Clark, & McClure, 2011). In predictive bias analyses, it is useful to consider effect sizes as well as statistical significance. See Nye and Sackett (2017) and Dahlke and Sackett (2017) for treatment of effect sizes in predictive bias analysis.

The definition above views any difference in slopes or intercepts as evidence of predictive bias. It is not uncommon, however, to frame the question as "is the use of a given predictor biased against members of a specified group?" In such cases, simply knowing that slopes or intercepts differ does not answer the question. Instead the focus is on whether the performance of the group in question is underpredicted; only a finding of underprediction signals bias against the group of interest.

Predictive bias has been examined extensively in the cognitive ability domain in the U.S. For White–African American and White– Hispanic comparisons, slope differences are rarely found. Although intercept differences are not uncommon, they typically take the form of overprediction of minority subgroup performance (Schmidt, Pearlman, & Hunter, 1980). Aguinis, Culpepper, and Pierce (2010) challenged the overprediction finding from previous research, indicating that the intercept test was biased toward overestimating the size of intercept differences and that earlier analyses used measures of observed validity as opposed to operational validity. In a paper that corrected these two problems, Berry and Zhao (2015) reported evidence that the performance of African-Americans generally remains overpredicted when using cognitive ability tests, and that underprediction occurred in only very specific and relatively uncommon circumstances. This result was found regardless of whether subgroup regression slopes differed or not. (Similar results were found by Mattern & Patterson [2013] in the college admissions context.)

Based on research using the same dataset as Mattern and Patterson (2013), Aguinis, Culpepper, and Pierce (2016) questioned whether differential prediction findings generalize across contexts. Interpretation of their findings, however, is clouded by the inclusion of multiple highly correlated tests in one prediction model, leading to instability in regression weights

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

across samples. Future research is needed to understand when and why various forms of differential prediction may exist.

There has been little published research on predictive bias associated with other predictor constructs and for other subgroup comparisons, although some work on male–female comparisons and on personality constructs has appeared. Saad and Sackett (2002) report findings parallel to those in the ability domain in examining predictive bias by sex using personality measures (i.e., little evidence of slope differences and intercept differences in the form of overprediction of female performance). Keiser, Sackett, Kuncel, and Brothen (2016) report that college admissions scores do not under-predict women's cognitive performance but do underpredict female performance on less cognitive, discretionary components of academic performance. In the international context, a variety of sub-groups may be of interest in predictive bias research (Myors et al., 2008). Given the limited research to date, broad conclusions about the prevalence of predictive bias for many constructs and subgroup comparisons are premature at this time.

Several important technical concerns with the analysis of predictive bias are noted here. First, analysis of predictive bias is appropriately conducted on predictors as operationally used. If, for example, selection will be conducted using the composite of multiple tests, analyses of predictive bias should be done using the composite, rather than using each test separately (Sackett, Laczo, & Lippe, 2003). Second, predictive bias requires an unbiased criterion. Confidence in the criterion measure is a prerequisite for an analysis of predictive bias. The third is the issue of statistical power to detect slope and intercept differences. Small total or subgroup sample sizes, unequal subgroup sample sizes, range restriction, and predictor unreliability are factors that can contribute to low power (Aguinis et al., 2010). A fourth is the assumption of homogeneity of error variances (Aguinis, Peterson, & Pierce, 1999); alternative statistical tests may be preferable when this assumption is violated (Oswald, Saad, & Sackett, 2000). Fifth is the need to use an unbiased estimate of the intercept difference and operational validity parameters instead of observed parameters (Berry & Zhao, 2015).

Some perspectives view the analysis of predictive bias as an activity contingent on a finding of mean subgroup differences. However, subgroup differences and predictive bias can exist independently of one another. Thus, whether or not subgroup differences on the predictor are found, predictive bias analysis should be undertaken when there are compelling reasons to question whether a predictor and a criterion are related in a comparable fashion for specific subgroups, given the availability of appropriate data. In domains where relevant research exists, generalized evidence can be appropriate for examining predictive bias (e.g., Berry & Zhao, 2015).

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Measurement bias

Measurement bias refers to sources of irrelevant variance that result in systematically higher or lower scores for members of particular groups, and it is a potential concern for all variables, both predictors and criteria. Determining whether measurement bias is present is often difficult, as this requires comparing an observed score to a true score.

Linked to the idea of measurement bias in terms of conducting analysis at the item level is the concept of an item sensitivity review (Golubovich, Grand, Ryan, & Schmitt, 2014), in which items are reviewed by individuals with diverse perspectives for language or content that might have differing meaning for members of various subgroups or could be demeaning or offensive to members of various subgroups. Instructions to candidates and to scorers or assessors may also be reviewed in a similar manner. The value of such analyses will vary by selection procedure content, and the need for and use of such information is a matter of the testing professional's judgment in a given situation.

One approach to examining measurement bias in the domain of multiitem selection procedures is to perform a differential item functioning (DIF) analysis. DIF refers to analyses that identify items for which members of different subgroups with identical total test scores (or identical estimated true scores in item response theory [IRT] models) have differing item performance.

A number of points related to DIF research are worth noting. First, these analyses require samples of sufficient sizes to produce stable results. Second, empirical research in domains where DIF analyses are common has rarely found sizable and replicable DIF effects (Sackett, Schmitt, Ellingson, & Kabin, 2001). Third, research has suggested that for cognitive tests it is common to find roughly equal numbers of differentially functioning items favoring each sub- group, resulting in no systematic bias at the test level (Chernyshenko & Drasgow, 2004). As a result of these factors, DIF findings should be viewed with caution. DIF analyses are not a routine or expected part of the selection procedure development and validation process in employment settings; however, testing professionals may choose to explore DIF when data sets appropriate for such analysis are available. Such analyses may be particularly useful when selection procedures are used in cross-cultural settings and test takers differ linguistically.

OPERATIONAL CONSIDERATIONS IN PERSONNEL SELECTION

This section of the *Principles* describes operational issues associated with the development or choice of a selection procedure, the conduct or accumulation of research to support the validity inferences made, documentation of the research effort in technical reports and administration manuals, and

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

subsequent implementation and use. The need for sound professional judgment based on the scientific literature and the testing professional's own experience will be required at every step of the process. In addition, all aspects of the research described in the *Principles* should be performed in compliance with the ethical standards of the American Psychological Association (2017a) as endorsed by SIOP.

Topics are introduced in an order that generally corresponds to the temporal progression of the validation effort. For example, the section on understanding work and worker requirements precedes decisions regarding the selection procedure. In other cases, the placement is based on the logical relationship among the topics. The order in which steps are taken in practice is ultimately a matter of professional and scientific judgment based on the given situation. It is recognized that in some instances a selection procedure may be implemented at the same time the validation process is underway.

Initiating a Validation Effort

The testing professional works collaboratively with representatives of the organization to define its needs and objectives, identify organizational constraints, plan the research, and communicate with major stakeholders regarding aspects of the process that will involve or affect them.

Defining the organization's needs, objectives, and constraints

Testing professionals use their expertise and experience to assist the organization in refining its goals and objectives. Different units of the organization may have different and sometimes competing and conflicting objectives. For instance, one unit may prefer rigorous selectionstandards even though they create hardships for another unit responsible for recruiting qualified applicants.

Organizations often consider costs (price, time, administrative effort) when choosing among selection procedures. These costs should be weighed against the benefits of the proposed selection system through a cost–benefit analysis.

The testing professional is encouraged to work with all units (e.g., human resources, internal or outsourced recruiting, labor relations, legal, compliance, information technology) that may have an effect on or be affected by the selection procedure and with other relevant stakeholders (e.g., internal or external individuals, and groups such as labor organizations, work councils, advocacy groups, customers). The testing professional provides accurate information regarding the benefits and limitations of various strategies in meeting the organization's goals based on past experience and the relevant body of scientific research. In all situations, the testing professional

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

and the organization's representatives should factor in the desires of the various stakeholders and determine the relative levels of consideration to be given to each point of view.

Climate and culture. Testing professionals face the challenge of ensuring high quality selection procedures in the context of the organization's history and current environment regarding employment-related strategies and practices, as well as the cultural setting in which it operates. Organizations operate in complex environments that sometimes place extreme and conflicting pressures on the management team. Testing professionals must consider the attitudes and commitments of organization leaders and employees who are faced with intense competition, mergers, stakeholder demands, and other corporate events that may influence the perceived relative importance of selection research. Testing professionals also may need to take into account the legal and labor environment when deciding on validation approaches or selection instruments. In addition, many HR functions are interrelated, with actions in one area affecting other areas. For example, changes in the selection standards often impact the level and extent of training required. Global selection systems should also take into consideration locally accepted practices and the organization's ability to execute the selection procedure accurately and reliably, regardless of location.

Workforce size and availability. The number of individuals who currently perform the work and their similarity to the applicant population can be important considerations when designing the validation strategy. The number of workers may shape the validation strategy pursued (e.g., validity generalization, synthetic validation, content-oriented strategy) as well as affect the feasibility and method for pilot testing procedures. Even when the number of workers is sufficient to conduct a local validation study, their availability and willingness to participate in a study may be limited. For example, organizational needs may require that a core group of workers be present on the job at all times, labor organizations may influence the number and type of persons willing to participate in the research, or workers who have experienced organizational restructuring may be skeptical about the purpose of the research and its effect on their own positions. Careful consideration should also be given to the timing of the data collection for the validation study. For example, attempting to collect assessment data or manager ratings during the unit's busy season or when the organization is downsizing can affect the quality of the data.

Large discrepancies in the capabilities of incumbents and the available applicant pool also present challenges, particularly in establishing norms and setting cutoff scores. For example, organizations that base cutoff scores on the performance of incumbents may find that those cutoff scores are too high, and thus inappropriate, if the organization's workforce is more

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

capable than the applicant pool. Similarly, organizations seeking to upgrade the skills of their current workforce may need other sources of information for setting cutoff scores.

Sources of information. Sources of information needed for validation and implementation efforts may include, but are not limited to, the workers themselves, managers, supervisors, trainers, customers, archival records, business performance metrics, and research conducted internal and external to the organization (including meta-analyses and sources such as O*NET). Based on the complexity of the work, the climate, and organizational constraints, some sources of information may be preferred over others. In some situations, the preferred source of information may not be available. Depending on the organizational constraints, alternatives to the testing professional's preferred source of information may be required. Alternative sources also may be used to supplement information gathered from the preferred source. Sources of information must be complete enough and directly relevant to support validation efforts.

Acceptability of selection procedures. Most organizations desire selection procedures that are predictive of important outcomes, easy and quick to administer, cost effective, and legally defensible. However, there are often additional considerations. For example, an organization's past experiences with respect to certain types of selection procedures may influence its decisions. Selection procedures that have been legally challenged in the past may not be acceptable to organizations, particularly if the organization was not successful in defending them. In addition, selection procedures that are viewed as controversial by individuals, labor organizations, or other stakeholders may not be acceptable. Some organizations find certain types of selection procedure questions unacceptable. For example, some biodata (e.g., childhood experiences) and personality inventory items may be viewed as an invasion of privacy, even if they can be shown to be related conceptually and empirically to the criterion measures or the requirements of the job. Cultures also differ in the acceptability of different kinds of selection procedures, so candidates' willingness to complete the assessment should be taken into consideration.

Some organizations prefer selection procedures that provide information regarding the strengths and developmental needs of the test taker. In such cases, procedures that measure knowledge or content that can be learned (e.g., software) may be preferred over procedures that elicit information concerning previous life experiences or stable personality traits. Procedures that appear more relevant or face valid to the organization may be more acceptable to stakeholders than other procedures that relate to a less obvious construct, regardless of any empirical evidence of validity. However, face validity is not an acceptable substitute for other forms of validity

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

evidence as treated in the *Principles*. Although acceptability is important, it is just one of many factors to consider when selecting or designing an effective selection procedure. Nevertheless, the testing professional should explain to decision makers issues underlying selection procedure acceptability as part of the initial planning effort.

Communicating the validation plan

Both management and workers need to understand in general terms the purpose of the research, the plan for conducting the research, and their respective roles in the development and implementation of the selection procedure. The testing professional must use professional judgment in determining the appropriate information to provide and the communication format and style that will be most effective. Testing professionals encourage organizations to consider the effects of participation in the validation effort on employees, managers, and business/organizational units. For example, organizations typically decide that data from a concurrent validation or selection system development study will be kept confidential and not be used for subsequent employment-related decisions. Organizations may also limit the number of performance ratings a manager is asked to make to minimize the demands on the manager's time and promote high quality ratings.

Understanding Work and Worker Requirements

In many businesses and industries, the nature of work changes rapidly. Factors such as changes in technology, mission, security context, strategy, organizational structure, the applicant pool, or customer demands result in substantive and frequent changes in work behaviors and requirements. A new work analysis should be conducted when test developers or users have reason to believe that the nature of the work performed has changed meaningfully since any prior analysis was conducted.

Strategies for analyzing the work domain and defining worker requirements

The approach, method, and analyses used in a specific study of work is a function of the nature of the work itself, those who perform the work, and the organizational setting in which the work is accomplished. There is no single strategy that must be carried out, and multiple strategies may be appropriate.

There are situations in which the importance or relevance of a criterion indicator or construct is self-evident and does not require extensive work analysis. For example, absenteeism and turnover and their underlying constructs may be relevant to all jobs and all work activities in an organization. Therefore, demonstration of their relevance is not typically necessary.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Considerations in specifying the sampling plan

The sampling plan for data collection should take into account a variety of factors, including the number of workers, their work locations, their demographic characteristics, their performance-related characteristics (e.g., amount of experience, training, proficiency), shift or other work cycles, and other variables that might influence the work analysis. Inclusion of a broad sample of incumbents (or other SMEs) is likely to increase the representativeness of the results.

Documentation of the results

The methodology, data collection methods, analyses, results, and implications of the work analysis for the validation effort should be documented. Frequently, this documentation will include a description of the major work activities, important worker requirements and their relationships to selection procedure content, and scoring when appropriate. See Technical Validation Report section for more information about documenting results.

Selecting Assessment Procedures for the Validation Effort

The testing professional should exercise professional judgment to determine those selection procedures that should be included in the validation effort and take into consideration the organizational needs as well as the issues discussed in this section. The result of this step is often the test plan, which describes the predictor constructs that will be measured with each assessment procedure. An example of this might be a KSAO-by-test matrix, often included in a technical report.

Review of research literature and the organization's objectives

Testing professionals should become familiar with not only the organization's objectives for the selection system but also research relevant to the constructs to be measured. The research literature can be used to inform choices about selection procedures and the validation strategy to be employed.

Psychometric considerations

When selecting one or more predictors, a number of psychometric characteristics should be considered for each instrument. Some of the more important psychometric considerations include reliability, evidence supporting the validity of the intended inferences, and differences among subgroups.

Scoring considerations

The testing professional must ensure that administration and scoring tasks can be completed accurately and consistently across candidates and locations. For all testing modalities, regardless of format, medium, or platform, test professionals should ensure that scoring rubrics are standardized,

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

reliable, and appropriate in order to allow test users to make score-based inferences consistent with the content and intent of the assessment.

Format and medium

Format refers to the design of response requirements for selection procedure items (e.g., multiple-choice, essay). The choice of format may be influenced by the resources available to administer and score the selection procedure. For example, objectively scored items with established correct responses may be administered and scored in less time than selection procedures that require the individual to respond in more complex ways or that use rater-evaluated individual responses.

Medium refers to the method of delivery of the selection procedure content. For example, a measure of cognitive ability could be presented via paper-and-pencil, computer, video, or orally. There are advantages and disadvantages in selecting a particular medium. For example, computer-administered procedures may reduce the demands on administrators and enhance standardization.

Testing professionals may choose to use multiple media in test administration; however, changing the medium may affect the construct being measured and threaten the equivalency of scores across media. For example, converting a paper-and-pencil situational judgment test to a video in which the situations will be acted out will reduce the reading component of the test. Also, administering speeded tests of cognitive ability on computer rather than paper-and-pencil may alter the construct being measured (Mead & Drasgow, 1993). A number of considerations are important when evaluating different format and medium options, depending upon the visibility and impact of the job and the use of the examination scores. In prescreening testing, cost and efficiency of operation, as well as breadth of recruiting and accessibility of testing opportunities, may be the primary concern to the organization. In testing applications for high visibility/high impact roles (e.g., public safety), security, standardization of testing conditions, and candidate authentication are more important concerns. Organizational decision makers may find that unproctored internet-based assessments allow for developing a larger applicant pool. An alternative is a proctored assessment approach, which may be more costly and require applicants to travel to a test site, although it more readily allows for standardization of measurement, verification of applicant identity, and verification of applicant performance. Increasingly, technology is enabling remote forms of proctoring, and the testing professional considering this form of proctoring should carefully consider its pros and cons (Weiner & Hurtz, 2017). Professional judgment must be used in determining the appropriate medium for test administration in light of the organization's goals.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

In addition to understanding that scores from the same test delivered via different media or using different response formats may be noncomparable, developers of selection systems should be cognizant that format and medium can affect mean score differences among subgroups (Hough, Oswald, & Ployhart, 2001). Over time, assessment systems may demonstrate changes in scores or even validity if test material becomes compromised.

Acceptability to the candidate

In addition to the organization's needs and objectives, testing professionals should also consider the acceptability of the selection procedure to candidates. A number of factors influence candidates' reactions to a selection procedure, including individual characteristics (e.g., work experiences, demographics, and cultural backgrounds), the role of the individual (e.g., applicant, incumbent, manager), the extent to which the content of the selection procedure resembles the work, the individual's capability with respect to the constructs measured, length of the process, the modality of the online assessment, and the perceived passing or selection rate. Generally, the greater the similarity between the selection procedure and the work performed, the greater the acceptability to candidates, management, and other stakeholders. However, selection procedures that too closely resemble the work may be perceived as obsolete when the work changes and may assess KSAOs that are not needed at entry because they are learned during on-the-job training. Some selection procedures may appear less face valid than other procedures. For example, the value of information collected on biodata forms and personality inventories in predicting job performance may not be obvious to some, despite demonstrated validity. Communications regarding the selection procedure, the constructs measured, and the role of incumbents and managers in developing the procedure may improve understanding and acceptance of a selection procedure.

There are times when some candidates refuse to participate in certain types of selection procedures. It may be useful to consider whether desirable candidates remove themselves from consideration because of factors in the selection process. In addition, recruiters sometimes resist or attempt to circumvent the use of selection procedures because it increases the need for additional candidates. Testing professionals should consider approaches designed to minimize negative perceptions of a selection procedure.

Alternate forms

Alternate forms of a selection procedure (including item banks for adaptive tests and/or unproctored tests) may be needed to reduce practice effects and enhance security. Alternate forms may help the organization to mitigate the effects of a security breach and continue assessment

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

after a security breach. Testing professionals can provide information to organizations to help them balance these advantages with the increased costs for development and validation of alternate forms. If alternate forms are developed, care must be taken to ensure that candidate scores are comparable across forms. If alternate forms are used, establishing the equivalence of scores on the different forms is usually necessary. The statistical procedures used in equating studies typically take into account the size and relevant characteristics of the samples, the use of an anchor test or linking-test items, and the feasibility of determining equating functions within subgroups. Monitoring score distribution qualities across multiple test forms for parallel structure is important.

Selecting the Validation Strategy

Once testing professionals have worked with the organization to define its objectives for developing a selection procedure, understand the requirements of the work, and reach agreement on the type of selection procedure(s), testing professionals must decide what validation strategy or strategies will be pursued to accumulate evidence to support the intended inference(s). Clearly, the strategy selected must be feasible in the organizational context, and it must meet the project goals within the constraints imposed by the situation.

Fit to objectives, constraints, and selection procedures

In choosing a validation strategy, the testing professional should consider the fit of the strategy to the organization's objectives and constraints, as well as its fit to the selection procedures planned and the criterion measures. Three examples are provided below to describe possible ways in which validation strategies may be matched to organizational objectives and constraints.

In the first scenario, an organization wanting to assemble validity evidence for a small population position may rely upon a validity generalization strategy because extensive cumulative evidence exists for the predictor–criterion relationship in similar situations. In a second scenario, another organization wanting to extend a selection procedure from one business unit to another may use a transportability study to establish the validity of the employee selection procedure in another business unit with the same job. In a third scenario, neither option may be available when a position is unique to the organization, and in this case, the organization may use a content-based validity strategy.

Individual assessments

Individual assessment refers to one-on-one evaluations on the basis of a wide range of cognitive and noncognitive measures that are integrated by the

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

assessor, often resulting in a recommendation rather than a selection decision or prediction of a specific level of job performance (Silzer & Jeanneret, 2011). The assessor should have a rationale for the determination and use of the selection procedures. In such instances, the validity of the assessor's clinical judgments is most important to the evaluation of the assessment process. If there are multiple assessors, the consistency of their assessment findings can be valuable to understanding validity and making accurate judgments about the relevant KSAOs or competencies. Validation research studies of clinical judgments are clearly an exception rather than the rule (Church & Rotolo, 2011; Kwaske, 2008; Ryan & Sackett, 1998; Silzer & Jeanneret, 2011). However, both validity generalization and content-oriented validation strategies may be appropriate when designing an individual assessment strategy. For example, there may be a wide range of generalizable evidence that has been accumulated by a test publisher or the assessing psychologist demonstrating that a personality scale (e.g., conscientiousness) is predictive of successful managerial performance (e.g., Morris, Daisley, Wheeler, & Boyer, 2015) and would, therefore, be appropriate for use in an executive assessment protocol. An example of a content-oriented validation approach would be demonstrating the relationship of an in-basket selection procedure that measures planning capability to the planning requirements of an executive position.

Selecting Criterion Measures

When the source of validity evidence is the relationships between predictor and criterion scores, the nature of those criteria is deter- mined by the outcomes from the work analysis, including worker requirements (e.g., KSAO or competency model) and proposed uses of the selection procedures. Professional judgment should be exercised in selecting appropriate criteria given known organizational constraints and climate.

Performance-oriented criteria

Criteria that are representative of work activities, behaviors, or outcomes usually focus on the job performance of incumbents. Supervisory ratings are the most frequently used criteria, and often they are designed specifically for use in the research study, as opposed to operational performance management measures used for administrative purposes. Other performance information may also be useful (e.g., training program scores, sales, error rates, customer ratings, and productivity indices). Attention to avoiding bias against demographic groups is important when selecting criteria, and consideration should be given to psychometric characteristics of all criteria whenever feasible.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Other indices

Depending on the objective of the validation effort, indices other than those directly related to task performance may be appropriate. Examples include absenteeism, turnover, and other organizational citizenship behaviors. The testing professional should be cautious about deficiencies or contaminating factors in all indices.

Relevance and psychometric considerations

Criteria are typically expected to represent some organizationally relevant construct (e.g., work performance, citizenship behavior, counterproductive behavior), and the quality of that representation should be established. For example, the fidelity of a work sample used as a criterion should be documented on the basis of the work analysis. Supervisory ratings should be defined and scaled in terms of relevant work activities or situations. All criteria should be representative of important work behaviors, outcomes, or relevant organizational expectations regarding individual employee behavior or team performance.

Although criteria should demonstrate adequate levels of reliability, calculation of an appropriate reliability estimate may be influenced by the data available for the study and organizational constraints. For example, one can typically calculate some form of criterion reliability estimate in any local validation study. However, depending upon the inferences the testing professional desires to make regarding the criterion scores, the reliability estimate that the testing professional is able to calculate (based on the data on hand) may not appropriately reflect the types of measurement error of interest. When reporting criterion reliability, the testing professional should describe the type of reliability estimate and sources of error that are reflected in (and ignored by) the reliability index.

Data Collection

The collection of both predictor and criterion data in a validation study requires careful planning and organizing to ensure complete and accurate data. The standardized conditions under which the validation research is conducted are normally replicated to the extent possible during actual use of the selection procedure. In order to collect accurate and complete information, the test user should consider the following activities.

Communications

Relevant information about the data collection effort should be communicated to all those affected by the effort, including management, those who take the test for research purposes, those who provide criterion data, those who will use the test, and other appropriate stakeholders. Appropriate

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

communications will facilitate the data collection and encourage all involved to provide accurate and complete information. The kind of information shared depends on the needs of the organization and the individuals involved. For example, participants in the validation research will want to know how their test results will be used, who will have access to the results, and how security of test and criterion data will be maintained over time. Supervisors who provide criterion ratings and others who provide archival criterion data will want to know the logistics of data collection, ultimate use, provisions for confidentiality, and data security protections. End users, such as the staffing organization or the client organization employing individuals who were screened with the selection procedures, should have an overview of the study. When feasible, anticipated uses of work analysis, test, and criterion data should be shared with those who generated them. Periodic updates to stakeholders on project status, go-live dates, responsibilities, and process, as well as a final briefing on the project results, are recommended.

Pilot testing

The testing professional should determine the extent to which pilot testing is feasible, necessary, or useful to ensure that data collection will go smoothly. Previous experience with specific selection procedures may reduce or eliminate this need. Availability of test takers and opportunities to conduct pilot testing may be influenced by various organizational constraints, such as strained union–management relationships and security concerns.

Match between data collection and implementation expectations

Selection procedures should be administered in the same way during the validation research that they will be administered in actual use. For example, if interviewers are provided face-to-face training in the validation study, similar training should be provided in actual use. Instructions and answers to candidate questions should be as similar as possible during validation and implementation.

Confidentiality

Confidentiality is an ethical responsibility of the testing professional. It is also a major concern to all those involved in the research. Those who provide information, performance ratings, or content validity linkages may be more willing to provide accurate information if they are assured of the confidentiality of their individual contributions. Participants in validation research studies should be given confidentiality unless there are persuasive reasons to proceed otherwise.

The testing professional should carefully decide what level of anonymity or confidentiality is required by relevant privacy laws and can be established,

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

communicate it to participants, and maintain it thereafter. The testing professional provides the maximum confidentiality feasible in the collection and storage of data, recognizing that identifying information of some type is often required to link data stored in different databases, collected at different times, or collected by different methods. Online data collection presents additional confidentiality challenges, such as insuring the security of the data collected.

Quality control and security

The test user should employ data collection techniques that are designed to enhance the accuracy and security of the data and test materials. Public disclosure of the content and scoring of most selection procedures should be recognized as a potentially serious threat to their reliability, validity, and subsequent use. All data should be retained at a level of security that permits access only for those with a need to know.

Issues of quality control and test security become particularly salient in unproctored internet testing (UIT) or remotely proctored internet testing (RPIT) environments. In these contexts, mechanisms and procedures should be adopted that diminish the chances of the assessment content being compromised, reduce the opportunity for cheating on the assessment, and help ensure positive identification of the individual completing the assessment. Test users considering the use of UIT or RPIT should be familiar with the advantages and disadvantages of these assessment options, as well as evolving technology for RPIT and emerging best practices in these areas (e.g., International Test Commission, 2006; Sackett, 2009a).

Data Analyses

A wide variety of data may be collected and analyzed throughout the course of a validation study. The responsibilities and supervision of the people who conduct data analyses should be commensurate with their capabilities and relevant experience.

Data accuracy and management

As part of the data collection process, measures and procedures should be included to facilitate later analyses of the quality of data provided by validation study participants. For example, the testing professional should consider including content or mechanisms to help identify careless or insufficient effort responding (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Meade & Craig, 2012).

Raters of job performance should be asked about factors that could influence their ability to provide quality performance ratings, such as their level of familiarity with the ratee's performance, opportunities to observe the

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

ratee's performance, and length of supervision of the ratee. All decision rules used when preparing the data for analyses should be clearly documented and appropriately justified.

Although becoming less common given advances in technology, a double-entry process should be considered to help ensure accurate entry when data are manually entered. Regardless of whether data are manually entered or captured through technology, values for all variables in the resulting data set should be checked for out-of-range values or, in the case of a technology-enabled data collection, missing data that may be indicative of a technology glitch. Data should also be checked for logical inconsistences that often arise when extracting data for a validation study from multiple sources, for example, checking that demographic information (e.g., sex, race, age, tenure) obtained from archival and self-report data match. Clear decision rules for handling any inconsistencies should be documented.

If archival data are included in the validation study, extra precautions should be taken prior to analyzing such data. Issues involving data privacy, data integrity, and consistency of variable naming and definitions over time are all critical factors to consider.

Missing data and outliers

Often, one or more data points are missing and/or outliers exist in the data set. The testing professional must examine each situation on its own merits and follow a strategy for handling missing data and/or outliers based on professional judgment informed by best practices cited in the literature on handling missing data and outliers.

When analyzing data collected for validation studies, two commonly recommended strategies are full information maximum likelihood (FIML) and multiple imputation (MI) approaches (Enders, 2010). Default options for handling missing data in common statistical packages (i.e., listwise and pairwise deletion, mean imputation) are often poor choices (Wilkinson & APA's Taskforce on Statistical Inference, 1999; see Enders, 2010; Graham, 2009; Little & Rubin, 2002; and Newman, 2014, for methods for dealing with missing data). When there are missing data, the testing professional should provide (a) a summary of missing data patterns and the nature of the missingness (e.g., missing at random, missing completely at random, missing not at random) and (b) justification for the missing data technique adopted for analyses.

Testing professionals should also check their data for both univariate and multivariate outliers (Aguinis, Gottfredson, & Joo, 2013). Documentation should include how outliers were defined and identified. If clear outliers are found, sensitivity analyses should be performed to evaluate the effects of including and excluding outliers on the validation study results, or robust

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

estimation/analysis techniques should be used that account for the presence of outliers. Orr, Sackett, and DuBois (1991) report that most testing professionals oppose dropping outliers unless there is evidence that the data point is erroneous. Dropping outliers to obtain more favorable results is not appropriate.

Descriptive statistics

Most data analyses will begin with descriptive statistics for predictor and criterion variables that present analyses of frequencies, central tendencies, and variances. Such descriptions should be provided for the total group and for relevant subgroups if they are large enough to yield reasonably reliable estimates.

Appropriate analyses

Data analyses should be appropriate for the method or strategy undertaken. Data are frequently collected as part of the analysis of work during the pilot or field testing of predictor/criterion measures and during the validation effort itself. Data analytic methods used should be appropriate for the nature of the data (e.g., nominal, ordinal, interval, ratio), sample sizes, and other considerations that will lead to correct inferences. For example, the presence of nonindependence (clustering of individuals) in the predictor–criterion data being analyzed can affect the accuracy/quality of inferences and should be considered. (For a review of nonindependence issues and their potential effects on evaluating predictor–criterion relations, see Bliese & Hanges, 2004; and LaHuis & Avis, 2007.)

Differential prediction

Organizations vary in their goals, and competing interests within an organization are not unusual. Efforts to reduce differences for one subgroup may increase differences for another. Given the difficulty in reconciling different interests in the case of substantial over- or underprediction, testing professionals often consider the effects of the prediction errors and their relationship to organizational goals.

A finding of predictive bias does not necessarily prevent the operational use of a selection procedure. For example, if the study is based upon an extremely large sample, a finding of a small but statistically significant differential prediction may have little practical effect. In general, the finding of concern would be evidence of substantial underprediction of performance in the subgroup of interest. Such a finding would generally preclude operational use of the predictor and would likely lead to additional research and considerations of modifying or replacing the selection procedure for all groups.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Absent a finding of substantial underprediction, a reasonable course of action for some organizations would be to recommend uniform operational use of the predictor for all groups. However, a substantial amount of overprediction may also lead to a consideration of dropping the predictor for all groups and/or investigating alternate selection procedures for all groups.

Combining selection procedures into a selection system

As noted earlier, the testing professional must exercise professional judgment regarding the outcomes of the overall selection system to determine those predictors that should be included in the final selection system and the method of combination and sequencing that will meet the goals of the organization (cf. Combining predictors and combining criteria). The methods used for combining and sequencing predictors should be clearly documented and justified. When combining predictors to form an overall score or make an overall decision, organizations may have different goals and values. For example, some organizations may put more emphasis on maximizing validity relative to minimizing subgroup differences. In contrast, other organizations may put more emphasis on minimizing subgroup differences relative to maximizing validity or may desire striking a balance between maximizing validity and minimizing sub- group differences.

Multiple hurdles versus compensatory models

Taking into account the purpose of the assessment and the outcomes of the validity study, the testing professional must decide whether candidates are required to score above a specific level on each of several assessments (multiple hurdles) or achieve a specific total score across all assessments (compensatory model). There are no absolutes regarding which model should be implemented, and, at times, hybrid approaches are possible (e.g., a hurdle may be most appropriate for one predictor, while a compensatory model may be best for other predictors within the overall selection procedure). The rationale and supporting evidence should be presented for the model recommended for assessment scoring and interpretation. Testing professionals should be aware that the method of combining test scores might influence the overall reliability of the entire selection process and the subgroup passing rates (Sackett & Roth, 1996).

Cutoff scores versus rank orders

Two frequently implemented selection decision strategies are the use of (a) a cutoff score and (b) rank order/top-down selection. A cutoff score defines the point on a selection procedure score distribution below which candidates are rejected. There is no single method for establishing cutoff scores; several potentially viable options exist (Mueller, Norris, & Oppler, 2007). For

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

example, cutoff scores may be criterion referenced when the predictor score can be linked to a meaningful performance threshold. If based on valid predictors demonstrating linearity or monotonicity throughout the range of prediction, cutoff scores may be set as high or as low as needed to meet the requirements of the organization. When there is an indication of nonmonotonicity in predictor-criterion relationships, this finding should be taken into consideration in determining how to use those scores for making personnel decisions (e.g., Converse & Oswald, 2014).

When data are not locally available to evaluate linearity and monotonicity in predictor-criterion relations, testing professionals should consider findings from past research and their implications for proposed use of scores. For example, though research evidence suggests relations between measures of cognitive ability and job performance are linear (e.g., Arneson, Sackett, & Beatty, 2011; Coward & Sackett, 1990), findings with regard to linearity of relations between other types of predictors (e.g., personality measures) and job performance have been mixed and is an open area of research (e.g., Carter et al., 2014).

Beyond the factors noted above, professional judgment is necessary when setting any cutoff score and when deciding between use of cutoff scores, top-down selection, or score bands (addressed in the next section). These decisions are typically driven by the goals of the organization and may be based on factors such as the estimated cost–benefit ratio, the number of vacancies and the selection ratio, the labor market, expectancy of success versus failure, the consequences of failure on the job, other consequences of selection decision errors, the relative emphasis on the performance and diversity goals of the organization, judgments as to the level a KSAO/competency or performance required by the work, and the utility of the selection procedure. Whatever the decision, the testing professional should document the rationale for it. The goals of the organization may favor a particular alternative. For example, some organizations decide to use a cutoff score rather than rank ordering to increase workforce diversity, recognizing that a reduction also may occur in job performance and utility. Whatever the decision, the researcher should document the rationale for it.

When evaluating or recommending cutoff scores for selection procedures, it may be useful to consider conditional standard errors for the selection measure/composite on which the cutoff scores are being set in the vicinity of those cutoff scores. Documentation should indicate the model used to compute the conditional standard errors (Brennan, 1998; Qualls-Payne, 1992; Raju, Price, Oshima, & Nering, 2007). One might also provide estimates of the percentage of applicants who would be classified in the same way (i.e., pass/fail) on two or more replications of the selection procedure at the given cutoff score (Haertel, 2006).

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Bands

Bands are ranges of selection procedure scores in which candidates are treated alike. The implementation of a banding procedure makes use of cut-off scores (i.e., to delineate predictor score ranges that define the bands), and there are a variety of methods for determining bands (Campion et al., 2001; Cascio, Outtz, Zedeck, & Goldstein, 1991).

Bands may be created for a variety of administrative or organizational purposes; they also may be formed to take into account the imprecision of selection procedure scores and their inferences. However, because bands group candidates who have different selection procedure scores, predictions of expected criterion outcomes are less precise. Thus, banding will generally yield lower expected criterion outcomes and selection utility (with regard to the criterion outcomes predicted by the selection procedure) than will top-down, rank order selection. On the other hand, the lowered expected criterion outcomes and selection utility may be balanced by benefits such as administrative ease and the possibility of increased workforce diversity, depending on how within-band selection decisions are made. If a banding procedure is implemented, the basis for its development and the decision rules to be followed in its administration should be clearly documented.

Norms

Normative information relevant to the applicant pool and the incumbent population should be presented when appropriate. The normative group should be described in terms of its relevant demographic and occupational characteristics. The time frame in which the normative results were established should be stated.

Communicating the effectiveness of selection procedures

Two potentially effective methods for communicating the effectiveness of selection procedures are expectancy analyses and utility estimates.

Expectancies and practical value. Expectancy charts may assist in understanding the relationship between a selection procedure score and work performance. Further, information in the Taylor-Russell tables (Taylor & Russell, 1939) identifies what proportions of hired candidates will be successful under different combinations of test validity (expressed as correlation coefficients), selection ratios, and percentages of current employees that are satisfactory performers.

Utility. Projected productivity gains or utility estimates for each employee and the organization due to use of the selection procedure may be relevant in assessing its practical value. Utility estimates also may be used to compare the relative value of alternative selection procedures. The literature regarding the impact of selection tests on employee productivity has

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

provided several means to estimate utility (Brogden, 1949; Cascio, 2000; Cronbach & Gleser, 1965; Hunter, Schmidt, & Judiesch, 1990; Naylor & Shine, 1965; Raju, Burke, & Normand, 1990; Schmidt, Hunter, McKenzie, & Muldrow, 1979). Some of these utility estimates express utility in terms of reductions in some outcome of interest (e.g., reduction in accidents, reduction in person hours needed to accomplish a body of work). Others express utility in dollar terms, with the dollar value obtained via a regression equation incorporating a number of parameters, such as the increment in validity over current practices and the dollar value of a standard deviation of performance. Still others express utility in terms of percentage increases in output due to improved selection. The values for terms in these models are often estimated with some uncertainty, and thus the result is a projection of gains to be realized if all of the model assumptions hold true. For various reasons, including feasibility, testing professionals often do not conduct follow-up studies to determine whether projected gains are, in fact, realized. Under such circumstances, the results of utility analyses should be identified as estimates based on a set of assumptions, and minimal and maximal point estimates of utility should be presented, when appropriate, to reflect the uncertainty in estimating various parameters in the utility model.

Appropriate Use of Selection Procedures

Inferences from selection procedure scores are validated for use in a prescribed manner for specific purposes. To the extent that a use deviates from either the prescribed procedures or the intended purpose, the inference of validity for the selection procedure is likely to be affected.

Combining selection procedures

Personnel decisions are often made on the basis of information from a combination of selection procedures. The individual components as well as the combination should be based upon evidence of validity. Changes in the components or the mix of components typically require the accumulation of additional evidence to support the validity of inferences for the altered procedure. When a compensatory approach is used, the addition or deletion of a selection procedure component can fundamentally change the inferences that might be supported. Under these circumstances, the original validation evidence might not be sufficient to support the altered selection procedure.

Using selection procedures for other purposes

The selection procedure should be used only for the purposes for which there is validity evidence. For example, diagnostic use of a selection procedure that

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

has not been validated in a way to yield such information should be avoided. Likewise, the use of a selection procedure designed for an educational environment cannot be justified for the purpose of predicting success in employment settings unless the education tasks and the work performed in the validation research or their underlying requirements are closely related or unless the relevant research literature supports this generalization.

Recommendations

The recommendations based on the results of a validation effort should be consistent with the objectives of the research, the data analyses performed, and the testing professional's professional judgment and ethical responsibilities. The recommended use should be consistent with the procedures used in and the outcomes from the validation research, including the validity evidence for each selection procedure or composite score and the integration of information from multiple sources. In addition, the testing professional typically considers the cost, labor market, effects on protected groups as well as workforce diversity, and performance expectations of the organization, particularly when choosing a strategy to determine who is selected by the procedure.

Technical Validation Report

Reports of validation efforts should include enough detail to enable a testing professional competent in personnel selection to know what was done, draw independent conclusions in evaluating the research, replicate the study, and make recommendations regarding the use of the selection procedure. The reports must accurately portray the findings, as well as the interpretations of and decisions based on the results. Research findings that qualify the conclusions or support the generalizability of results should be reported. The following information should be included:

Identifying information

The report should include the authors, their credentials, their affiliations, dates of the study, and other information that would permit another testing professional to understand who conducted the original research.

Statement of purpose

The purpose of the validation research should be stated in the report.

Analysis of work

The report should contain a description of the analysis of work, the characteristics of the participants in the process, any judgments made by SMEs, instructions that were provided to participants in data collection efforts

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

for their specific tasks, and data analyses and results including reliability/precision. If any of the documents used in the analysis of work were translated, then a description of the translation and adaptation procedures should be included.

Search for alternative selection procedures

The report should document any search for selection procedures (including alternate combinations of the procedures) that show substantially equal or greater validity for the given selection situation with an accompanying reduction in subgroup differences.

Identification or development of selection procedures

Names, editions, and forms of selection procedures purchased from publishers should be provided, as well as technical descriptions and, if appropriate, sample item content. When proprietary selection tools are developed, the testing professional should include a description of the content, including the construct(s) measured by the content, and the process by which the content was developed, if appropriate. Typically, content and scoring algorithms should not be included in technical reports or administration manuals in order to protect the confidentiality of operational items. However, detailed documentation of the scoring procedures will help to ensure accurate and consistent scoring.

The rationale for the use of each statistical procedure and results of relevant analyses performed should be included. If raters are an integral part of the selection procedure, as in some work samples, then the reliability and agreement of their ratings should be determined and documented.

Establishing validity

The report should provide a description of the validation studies conducted such that another testing professional could reproduce the analyses and results. The report should also include methods used by the testing professional to determine that the selection procedure is statistically and practically significantly related to a criterion measure and/or representative of a job content domain. Establishing the relationship of a selection procedure to job content and KSAOs is particularly important when conducting a job content validation study, both to justify the use of the selection procedure and to provide substantive support for its validity.

Criterion validation studies, when conducted, should report the following in detail: a description of the criterion measures; the rationale for their use; the data collection procedures; and a discussion of the measures' relevance, reliability, possible deficiencies, possible sources of contamination, and freedom from or control of biasing sources of variance. If the testing

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

professional developed the criterion measure, then the report should include the rationale and steps taken to develop it, so it can be well understood and, if needed, replicated in future validation studies.

Research sample

The sampling procedure and the characteristics of the research sample relative to the appropriate interpretation of the results should be described. The description should include a definition of the population that the sample is designed to represent, sampling biases that may detract from the representativeness of the sample, the significance of any deviations from representativeness for the interpretation of the results, and any statistical power analysis results. Data informing the potential restriction in the range of scores on predictors or criterion measures are especially important. When a transportability study is conducted to support the use of a particular selection procedure, the relationship between the original validation research sample and the population for which the use of the selection procedure is proposed should be included in the technical report. Test developers should make clear whether psychometrics in the technical report refer to candidates or incumbents, and results for concurrent validation studies should not be represented as the results for predictive validation studies.

Results

All summary statistics that relate to the conclusions drawn by the testing professional and the recommendations for use should be included. Complete statistical results related to the development and validation, not just statistically significant or supportive results, should be presented and clearly labeled. Both uncorrected and corrected values should be presented when corrections are made for statistical artifacts such as restriction of range or unreliability of the criterion.

Scoring and transformation of raw scores

Methods and algorithms used to score content should be fully described. For example, when weighted scores, derived scales, or composite or categorical scores are used, rationale should be pro- vided in detail. When performance tasks, work samples, or other methods requiring some element of judgment are used, a description of the type of rater training conducted and scoring criteria should be provided.

Normative information

Parameters for normative data provide testing professionals and users with information that guides relevant interpretations. Such parameters often include demographic and occupational characteristics of the normative sample, time frame of the data collection, and status of test takers (e.g.,

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

candidates, incumbents, students). When normative information is presented, it should include measures of central tendency and variability (and skewness when appropriate), and it should clearly describe the normative data (e.g., percentiles, standard scores). Normative tables usually report the percent passing at specific scores and may be useful in determining the effects of a cutoff score. Expectancy tables indicate the proportion of a specific sample of candidates who reach a specified level of success and are often used to guide implementation decisions.

Recommendations

The recommendations for implementation of selection procedures and the rationale supporting the recommendations (e.g., the use of rank ordering, score bands, or cutoff scores, and the means of combining information in making personnel decisions) should be provided. Some implementation rules may change over time (e.g., those applied to cutoff scores), and subsequent modifications should be documented and placed in an addendum to the research report or administration manual.

Caution regarding interpretations

Research reports or administration manuals should help readers make appropriate interpretations of data and should warn them against common misuses of information.

Technology–enabled selection procedures

If the selection procedure is technology enabled, the researcher should document the technology requirements and any technology-based accommodations that can be provided by the administrator for test takers with disabilities.

References

There should be complete references for all published literature and technical reports cited in the report. Technical reports completed for private organizations are often considered proprietary and confidential, and the testing professional may not violate the limitations imposed by the organization. Consequently, some technical reports that may have been used by the testing professional may not be generally available.

Administration Information

Individuals with test administration responsibilities include those responsible for day-to-day activities such as scheduling testing sessions, administering the selection procedure, scoring the procedure, maintaining the databases, and reporting scores or results. Those who have responsibilities related to the technology supporting administration, such as

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

programming scoring algorithms, maintaining web interfaces for testing and score reporting, and ensuring that updated or expanded test content is incorporated into the existing testing system, should be considered as part of a test administration team. The accuracy of their work is the responsibility of the test administration lead.

Those with day-to-day administration responsibilities should be aware of any personal limitations (physical, perceptual, cognitive) that might affect their ability to administer and/or score a test fairly and accurately, and they should not administer assessments when they cannot meet the demands of their roles or if there are barriers to their effective delivery of responsibilities.

Complete documentation should be available with regard to administering the selection procedure, scoring it, and interpreting the score, regardless of the mode of assessment delivery (e.g., paper-and-pencil, computerized, internet/web based). Although this documentation is sometimes a part of a technical report, it is often separate so that confidential information in the validation study is protected, and administrators are provided with only the information necessary and appropriate to administer the selection procedure. In other situations, the test user in the organization will develop some of the administration information and procedures, because the testing professional may not know the organization's specific policies or the details of its implementation strategies. In deciding whether separate documents are needed, the testing professional should consider who has access to each document, the sensitivity of the information to be included, the purpose of each document, and the intended audiences.

Administration information developed by a publisher is often supplemented with addenda that cover local decisions made by the user organization. Consequently, not all the information listed below will be found in administration documentation from a publisher or vendor. However, the testing professional in the user organization should try to provide answers or guidance for the issues raised.

The information developed for users or examinees should be clear, accurate, and complete for its purposes. Communications regarding selection procedures should be stated as clearly and accurately as possible so that readers know how to carry out administrative responsibilities competently. The writing style of all informational material should be appropriate to address the understanding and needs of the likely audience. When a test is to be administered in multiple countries and in multiple languages, documentation and supporting materials required for administration may need to undergo appropriate translation procedures. Normally, the following information should be included as administration documentation:

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Introduction and overview

This section of the documentation should inform the reader of the purpose of the assessment procedure and provide an overview of the empirical research that supports the use of the procedure. The introduction should explain why the organization uses formal, validated selection procedures, the benefits of professionally developed selection procedures, the importance of assessment security, and the degree of consistency required in administration. Care must be taken in preparing such documents to avoid giving the reader an impression that an assessment program is more useful or applicable than is really the case.

Contact information

The administration documentation should provide information about whom to contact in case questions or unanticipated problems associated with the selection procedure arise.

Selection procedures

The selection procedures should be thoroughly described. Names, editions, and forms of published procedures as well as information for ordering materials and ensuring their security should be provided. Although entire tests are not usually included in administration documentation for security reasons, providing sample items that represent all relevant aspects of the test can be very helpful. When proprietary tests are developed, the testing professional should include a description of the items, the construct(s) that are measured, and sample items.

Applicability

The description of the selection procedure should indicate to whom the procedure is applicable (e.g., job candidates for a specified job) and state any exceptions to test requirements (e.g., exemptions for job incumbents). Information on applicability to testing individuals with disabilities and individuals from different cultural and linguistic groups should be included. If the organization has rules about when tests are administered, these rules must be clearly stated in the administration documentation used by the organization. For example, some organizations only administer a selection procedure when there is a job vacancy, whereas other organizations may administer selection procedures periodically in order to build pools of qualified candidates.

Administration responsibilities

The administration documentation should state the necessary qualifications of those with different administrative responsibilities (e.g., for maintaining a

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

scoring algorithm, for handling retesting requests) and the training required to administer selection procedures in general, as well as training for the specific selection procedure of interest. Training should emphasize that failures in following the standardized protocols may render any research results and the meaning and interpretation of operational scores irrelevant to some degree. Testing professionals should document the nature of and the need for standardized administration of tests or other procedures. Periodic training may be needed to maintain understanding and compliance to the administration rules, especially when the people who are involved in administration change. Observational checks or other quality control mechanisms should be built into the test administration system to ensure accurate and consistent administration. Pass rates or mean scores of the assessment should be reviewed periodically to look for spikes, which may indicate the scoring key has been compromised, or dips, which might indicate problems or inconsistencies in test administration, such as not following test time limits or administration procedures.

Information provided to candidates

Many organizations provide information to candidates about the employee selection process (via brochures, web pages, emails, videos), and such information should be clear, pertinent, and timely. Depending on the test, the population of test takers, and the circumstances, the administrator should consider what information about the selection procedure to provide candidates. For example, information about the intended test use, administrative procedures, test format and interface, test completion strategies (e.g., opportunity to go back and change item responses), time parameters, feedback and access to scores (e.g., who will have access and how long data will be retained), confidentiality protections and conditions under which records may be released, processes for requesting accommodation for disability, warnings about improper candidate behavior and responsibility to respect copyright laws, retesting policies, and other relevant user policies as appropriate might be provided. Administrators might also convey whether and how test takers may review and correct their personal information, as well as how to appeal when test scores are cancelled or withheld (as in credential and licensure test settings), or when allegations of misconduct occur. Regardless of what information is provided to candidates, it should be clear and consistent. Both the content and the process for orienting candidates should be standardized whenever possible. The administration documentation should describe these materials and indicate how they are provided to candidates (e.g., via open website or email). The rules for distribution should be explicitly stated in order to facilitate consistent treatment of candidates.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Guidelines for administration of selection procedures

The testing professional should use the administration documentation as an opportunity to convey the organization's requirements for selection procedure administration. In addition to detailed instructions regarding the actual administration of the selection procedure, the documentation may include rules and tips for providing an appropriate testing environment as well as ensuring the candidate's identity. Some technology-enabled tests may require that test takers receive instruction and practice prior to administration. Test administrators are responsible for ensuring that any such instruction and practice are provided. When the test taker is responsible for his/her own testing environment (e.g., unproctored internet testing), the administrator still has the responsibility of informing the test taker of environmental factors likely to affect performance and of the characteristics of an appropriate testing environment. Further, those with test administration responsibilities also are responsible for informing test takers of any instructions regarding security (e.g., identification verification, setting up web cams, verification codes) and the consequences for the test taker of not following test security procedures. Appeals processes when testing irregularities have been detected should be conveyed.

Reasonable effort should be made to ensure the integrity of test scores (e.g., verifying identities of test takers). When appropriate, test administrators should be trained on how to take precautions against cheating, how to detect and prevent opportunities to cheat, and how to monitor and detect cheating as it occurs. Those who use technologies designed to detect irregularities (e.g., particular answer patterns or erasure patterns, plagiarism) are responsible for their appropriate use. Administrators should monitor the administration to control possible disruptions, protect the security of test materials, and prevent collaborative efforts by candidates. Although older versions of tests are sometimes made available by the test user for practice purposes, in general, tests should not be made available to the public or resold to unqualified test users. The security provisions, like other aspects of the *Principles,* apply equally to computer and internet-administered sessions.

Administration environment

There are a number of factors that potentially affect test administration. Examples include (but are not limited to) an appropriate workspace; adequate lighting; a quiet, comfortable setting, free of distractions; and the extent to which the test is technology enabled and the corresponding effects of requirements such as browser, monitor size, and touch screen. The testing professional should consider these conditions and their potential effects on test performance. At a minimum, selection procedure administration should

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

be in an environment that is responsive to candidates' concerns about the selection procedures and maintains their dignity. When effects of the environment on test performance are known, test takers should be informed which specific test-taking conditions may have consequences (e.g., potential lowered performance). Administrators should inform test takers on the general environmental conditions conducive for test taking when individuals are responsible for their own testing environments (e.g., unproctored internet testing).

Scoring instructions and interpretation guidelines

Testing professionals should provide the selection procedure administrators or users with details on how the selection procedure is to be scored and how results should be interpreted. Note that the documentation provided in commercially available test manuals may not provide sufficient or complete documentation with regard to the proper application of the selection procedure. Administration documentation should provide objective information regarding any role of the test administrator in the intended interpretation of test scores, the positive and negative consequences of test use, and protecting the security of test content and the privacy of test takers. The administration documentation should therefore help readers make appropriate interpretations of scores and related information and warn them against common misuses.

Processes should be followed to ensure accuracy in scoring, checking, and recording results. This principle applies to the testing professional and to any agent to whom this responsibility has been delegated. The responsibility cannot be ignored or substituted by purchasing services from an outside scoring service. Quality control checks and routine monitoring should be implemented to ensure accurate scoring and recording. Procedures for rescoring of tests when mistakes are suspected should be clear to administrators.

Instructions for scoring by the user should be presented in the administration documentation in detail to reduce clerical errors in scoring and to increase the reliability of any required judgments. Distinctions among measured constructs should be described to support the accuracy of scoring judgments. Scoring keys should not be included in technical reports or administration manuals and should be made available only to persons who score or scale responses.

If computer-based test scoring and interpretation procedures (e.g., automated feedback reports) are used to process responses to a selection procedure and generate reports, the testing professional should provide detailed instructions on how they are to be used in decision making. When relevant to the interpretation of test scores, the conditions under which the test

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

was administered (e.g., unproctored setting, accommodated test conditions) should be shared with the test user.

Test score databases

Organizations should decide what records of assessment administrations and scores are to be maintained and should provide detailed information regarding record keeping and databases (or should reference that detailed information). In addition, policies on the retention of records (e.g., duration, security, accessibility) and the use of archival data over time should be established and communicated as appropriate. As testing professionals establish data retention policies, they should keep in mind federal, state, and local guidelines; industry best practices; and recent court rulings on data retention for additional guidance on data collection, record keeping, and maintenance. Raw item data and scores should be retained, because data reported in derived scales may limit further research. When personally identifying information is included in research databases, the testing professional must ensure those data are secure and accessible only by those with a need to know. Databases should be maintained for sufficient time periods to support periodic audits of the selection process an ongoing evaluation of operational selection systems.

Reporting and using selection procedure scores

Documentation provided by the testing professional must communicate how selection procedure scores are to be reported and used. Results should be reported in language likely to be interpreted correctly by persons who receive them. The administration documentation should also indicate who has access to selection procedure scores.

Administrators should be cautioned about using selection procedure information for uses other than those intended. For example, although selection procedure data may have some validity in determining later retention decisions, more potentially relevant measures such as performance ratings may be available. Furthermore, if the pattern of selection procedure scores is used to make differential assignments to jobs or job groupings, evidence is required to support those assignments, such as by demonstrating that the scores are linked to, or predictive of, different performance levels across those jobs or job groupings.

Candidate feedback

In addition to reporting selection procedure scores to others within the organization, the testing professional should include information on how to provide feedback to candidates, if such feedback is feasible and appropriate. Feedback should be provided in clear language that is understandable

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

by candidates receiving the feedback, and feedback information should not violate the security of the test or its scoring.

Nonstandard administrations

The administration documentation should cover nonstandard selection procedure administrations. Such administrations encompass not only accommodated selection procedure sessions but also sessions that were disrupted (e.g., power failures, local emergency, and illness of a candidate), involved errors (e.g., questions and answer sheet did not match, timing mistake), or were nonstandard in some other way. Note that in some cases, the reporting of a nonstandard administration may be left to the test taker (e.g., a dropped internet connection), and whenever such reporting is invoked, the reporting and procedures for doing so must be clearly explained to the test taker.

The administration documentation should establish a clear process to document and explain any changes to selection procedures, disruptions in administration, or any other deviation from established procedures in the administration, scoring, or handling of scores. Although it is impossible to predict all possible occurrences, the testing professional should communicate general principles for how deviations from normal procedures are to be handled.

Reassessing candidates

Generally, employers should provide opportunities for reassessment and reconsidering candidates whenever technically and administratively feasible. In some situations, as in one-time examinations, reassessment may not be a viable option. To facilitate consistency of candidate treatment, the administration documentation should clearly explain whether candidates may be reassessed and how reassessment will take place. In some organizations, specific time intervals must elapse before reassessment occurs. In other organizations, significant developmental activities must have occurred prior to reassessment.

Corrective reassessment

Users in conjunction with testing professionals should consider when corrective reassessment is appropriate. Critical errors on the part of the administrator (e.g., timing mistakes, use of nonmatching selection procedure booklet and answer sheet) and extraordinary disturbances (e.g., fire alarm, acutely ill test taker) usually justify reassessment. The administration documentation should cover procedures and guidelines for granting corrective reassessment and documenting all requests. When test takers are remote from any administrative personnel (e.g., unproctored internet

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

testing), it is important that the test taker be informed of the conditions under which he/she might ask for a corrective reassessment, the required timing of such requests, the documentation required, and the procedure for doing so.

Security of the selection procedure

Selection procedure content that is widely known to job candidates in an organization (through study, coaching, internet resources, or other means) is usually less effective in differentiating among job candidates on relevant constructs. Maintenance of test security therefore is required, which necessarily limits the type and amount of test feedback provided to candidates. The more detail on candidate responses that is provided, the greater the security risk. The administration documentation should emphasize the importance of safeguarding the content, scoring, and validity of the selection procedure as well as monitoring for overexposure of the content.

Selection procedures usually represent a significant investment on the part of the organization for development and validation. The administration documentation should point out the value of the selection procedure itself and the cost of compromised selection procedures in terms of the additional research required and the possibility and risk of less capable candidates being hired.

It is important to communicate, exercise, and enforce practices that protect the security of selection procedure documents (e.g., verification codes for test access, rotation of content) and the security of selection procedure scoring. Procedures for the security of testing administrator training materials and previous test editions should be documented.

Selection procedure scores must be kept secure and should be released only to those who have a need to know and who are qualified to interpret them. International laws regarding data privacy change often and should be consulted in making these determinations. Special practices may be required to protect confidential materials and selection procedure information that exist in electronic forms. Although security practices may be difficult to apply in the case of employment interviews, the importance of security as a means of preserving their content, standardization, and validity should be considered. Organizations are encouraged to develop policies that specify the length of time confidential information is to be retained. When confidential information is destroyed, the user should consider ways of maintaining its security, such as having selection personnel supervise the destruction of the documents.

When other documents are mentioned, they should be referenced fully. When the documents are internal publications, the means of acquiring those documents should be described.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Other Circumstances Regarding the Validation Effort and Use of Selection Procedures

Influence of changes in organizational demands

Because organizations and their workforces are dynamic in nature, changes in organizational functioning may occur, and subsequent selection procedure modifications may be necessary. Changing work requirements may lead to the introduction of a new assessment or adjustments in cutoff scores for existing ones; both would require further study of the existing selection procedure. If advised of such circumstances, the testing professional should examine each situation on its own merits and make recommendations to the organization regarding the impact of organizational change on the validation and use of any selection procedure.

Review of validation and need for updating the validation effort

Testing professionals should develop strategies to anticipate that the validity of inferences for a selection procedure used in a particular situation may change over time. Such changes may occur because of changes in the work itself, worker requirements, or work setting, or the emergence of new jobs. Users of a selection procedure (either on their own or with testing professional assistance) should periodically review the operational use of the assessment instrument using the available data (including timeliness of normative data if appropriate) to determine whether additional research is needed to support the continued use of the selection procedure. When needed, the research should be brought up to date and reported. There is also a possible need for evidence that score interpretations continue to be appropriate when there is a change in test format, mode of administration, instructions, or language used in administering a test; the greater the changes, the more likely the need. The technical or administration documentation should be revised accordingly (or an addendum added) if changes in research data or use of procedures make any statement or instruction incorrect or misleading.

Assessing Candidates with Disabilities

Assessing candidates with disabilities may require special accommodations that deviate from standardized procedures in order to remove construct-irrelevant barriers that otherwise interfere with test takers' ability to demonstrate their standing on job-relevant constructs. Accommodations are made to minimize the impact of a known disability that is not relevant to the construct being assessed. For example, an individual's upper extremity motor impairment may lower a score on a measure of cognitive ability because of the candidate's difficulty taking the test, even though the motor impairment is not related to the individual's cognitive ability. Accommodations, which

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

typically do not affect the construct being measured, may include, but are not limited to, modifications to the environment (e.g., high desks), the testing medium (e.g., Braille, text reader), and the testing time limit. Adaptations to the test content, which often do change the construct being measured, are relatively rare in employment testing. Combinations of accommodations may be required to make valid inferences regarding the candidate's standing on the construct(s) of interest. The appropriate accommodation for a specific test taker must be determined by the facts of the test taker's situation; however, rules for determining who is eligible for an accommodation, how test takers can request and accommodation, and how accommodation requests will be evaluated should be as standardized as feasible. Test users should document these procedures and are responsible for monitoring their appropriate implementation. Test takers should be informed of the process and requirements for obtaining any needed accommodation and the confidentiality provisions regarding their disability status.

Professional judgment is required on the part of the user and the developer regarding the type or types of accommodations that have the least negative impact on the validity of the inferences made from the selection procedure scores. Empirical research is usually lacking on the effect of given accommodations on selection procedure performance for candidates with different disabilities or with varying magnitudes of the same disability. Note that a test may be modified so that it no longer assesses the same construct but still provides useful information; if a test is modified, such information should be documented. For example, an individual with dyscalculia may need a calculator for certain items on a broader mathematics problem-solving assessment, rendering the test modified but still useful for indicating something about the individual's skills. If a test no longer assesses the same construct in the same way as the original, these test scores can no longer be directly compared with scores from the unmodified test. When score reports are made, it is appropriate to indicate deviation from standard administration procedures and discuss how such deviation may affect results and interpretation to the extent permitted by law.

Responsibilities of the selection procedure developers, testing professionals, and users related to accommodation

Testing professionals and individuals charged with approving accommodations should be knowledgeable about the availability of accommodated forms of the selection procedure, psychometric theory, and the likely effect of the disability on selection procedure performance. In many employee selection contexts, empirical research to demonstrate comparability between the original procedure and the altered procedure will not be feasible. When changes mean the test no longer assesses the

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

same construct, this is considered a modification of the test itself. Users may choose to alter the original selection procedure, develop an altered procedure for candidates with disabilities, or waive the selection procedure altogether and use other information regarding the candidate's job-related KSAOs or competencies. Implications of these latter changes should be considered seriously, because they create potential challenges in terms of standardization, fairness, job relevance, and other issues.

Selection procedure accommodation and modification. The test user should take steps to ensure that a candidate's score on the selection procedure accurately reflects the candidate's ability rather than construct-irrelevant disabilities. One of these steps is a dialog with the candidate with the disability about possible accommodations. In some cases, the construct cannot be assessed without reasonably accommodating the disability. Other times, the disability does not affect performance on the selection procedure, and therefore no accommodation is necessary. Components of a selection procedure battery should be considered separately when determining appropriate accommodations. To the extent possible, standardized features of administration should be retained in order to maximize comparability among scores. Approval of prespecified, commonly used accommodations that are irrelevant to selection procedure scores and their psychometric interpretation (e.g., adjusting table height) may be delegated to administrators.

Development and validation. Although most employers have too few cases of accommodated tests for extensive research, the principles set forth in the *Principles* in the preparation of altered selection procedures for candidates with disabilities should be followed to the extent possible. Altered

procedures should be pilot tested when possible and feasible; at the very least, this provides practical experience in ensuring the altered procedure can be made operational and run smoothly. Practical limitations, such as small sample size, often restrict the ability of the testing professional to statistically equate data from accommodated versions of the selection procedure to data from the original form, thereby challenging the strict comparability of scores. These considerations also limit efforts to establish the reliability of the accommodated scores and the validity of the inferences made from these scores. Nevertheless, the reliability of accommodated selection procedure scores and the validity of inferences based on these scores should be determined whenever possible. In the rare case when it is possible and appropriate, the effects of administration of the original form of the selection procedure to candidates with disabilities also should be examined.

Documentation and communications regarding accommodations and modifications. Descriptions of the changes made, the psychometric characteristics of the accommodated or modified selection procedures, and, when sufficient volume of test takers makes it feasible, statistics summarizing the

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

performance of candidates with disabilities on the accommodated or modified forms of the procedure, and the original forms if available should be included in the documentation. Legal considerations may prohibit giving decision-makers information on whether a candidate's score was earned under a selection procedure accommodation and the nature of the accommodation. However, test users may designate those scores earned with an accommodation in such a way to permit special handling in data analysis.

Maintaining consistency with assessment use in the organization. The selection procedures used when assessing candidates with disabilities should resemble as closely as possible the selection procedures used for other candidates. To be clear, selection procedures are developed for the purpose of making selection decisions, not for the purpose of assessing the existence or extent of a candidate's disability. The addition of a procedure designed to assess the existence or degree of a disability is inappropriate as a selection tool and unlawful in many situations.

Candidate Linguistic and Cultural Background

In addition to identifying candidates with the KSAOs necessary to perform the job, creating and maintaining a diverse workforce is usually a corporate goal. Thus, the test developer must carefully consider the language requirements of the job to determine the languages in which the test will be offered. Developers should also ensure that the content and reading level are appropriate and equivalent across test forms administered in different languages (e.g., through translation and adaptation procedures; through relevant psychometric and statistical comparisons between groups). When appropriate, test administrators should inform test takers of linguistic options.

The cultural backgrounds of test takers can also introduce construct-irrelevant barriers to test performance. Again, the test developer must consider the test content and format and take steps to minimize these barriers to ensure the test is consistent with the requirements of the job.

REFERENCES

Ackerman, P. J., & Humphreys, L. G. (1990). Individual differences theory in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. *1*, pp. 233–282). Palo Alto, CA: Consulting Psychologists Press.

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95*, 648–680. doi:10.1037/a0018714

Aguinis, H., Culpepper, S.A., & Pierce, C. A. (2016). Differential prediction generalization in college admissions testing. *Journal of Educational Psychology, 108*, 1045–1059.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16*, 270–301.

Aguinis, H., Petersen, S. A., & Pierce, C. A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organizational Research Methods, 2*, 315–339.

Aguinis, H., & Pierce, C. A. (1998). Testing moderator variable hypotheses meta-analytically. *Journal of Management, 24*, 577–592.

Aguinis, H., Sturman, M. C., & Pierce, C. A. (2008). Comparison of three meta-analytic procedures for estimating moderating effects of categorical variables. *Organizational Research Methods, 11*, 9–34.

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

American Psychological Association. (2017a). *Ethical principles of psychologists and code of conduct* (2002, Amended June 1, 2010 and January 1, 2017). Retrieved from http://www.apa.org/ethics/code/ index.aspx.

American Psychological Association. (2017b). *Multicultural guidelines: An ecological approach to context, identity, and intersectionality*. Retrieved from http://www.apa.org/about/policy/multicultural- guidelines.pdf.

American Psychological Association. (2018). Professional practice guidelines for occupationally mandated psychological evaluations. *American Psychologist, 73*, 186–197.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*, 3–25.

Arneson, J. J., Sackett, P. R., & Beatty, A. S. (2011). Ability-performance relationships in education and employment settings: Critical tests of the more-is-better and the good-enough hypotheses. *Psychological Science, 22*, 1336–1342.

Arthur, W. Jr., & Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125–153.

Arthur, W. Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435–442.

Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods, 15*, 103–133.

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.

Beatty, A. S., Barratt, C. L., Berry, C. M., & Sackett, P. R. (2014). Testing the generalizability of indirect range restriction corrections. *Journal of Applied Psychology, 99*, 587–598.

Bemis, S. E. (1968). Occupational validity of the General Aptitude Test Battery. *Journal of Applied Psychology, 52*, 240–249.

Berry, C. M., Clark, M. A., & McClure, T. K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology, 96*, 881–906.

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology, 92*, 410–424.

Berry, C. M., & Zhao, P. (2015). Addressing criticisms of existing predictive bias research: Cognitive ability test scores still overpredict African Americans' job performance. *Journal of Applied Psychology, 100*, 162–179.

Bliese, P. D., & Hanges, P. J. (2004). Being both too liberal and too conservative: The perils of treating grouped data as though they were independent. *Organizational Research Methods, 7*, 400–417.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Bobko, P. (1983). An analysis of correlations corrected for attenuation and range restriction. *Journal of Applied Psychology, 68*, 584–589.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods, 10*, 689–709.

Bobko, P., & Stone-Romero, E. (1998). Meta-analysis is another useful research tool but it is not a panacea. In G. Ferris (Ed.), *Research in personnel and human resources management* (Vol. *16*, pp. 359–397). Greenwich, CT: JAI Press.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, UK: Wiley.

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement, 22*, 307–331.

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology, 2*, 171–183.

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

Callender, J. C., & Osburn, H. G. (1980). Development and testing of a new model of validity generalization. *Journal of Applied Psychology, 65*, 543–558.

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.

Campion, M. A., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). Doing competencies well: Best practices in competency modeling. *Personnel Psychology, 64*, 225–262.

Campion, M. A., Outtz, J. L., Zedeck, S., Schmidt, F. L., Kehoe, J. F., Murphy, K. R., & Guion, R. M. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology, 54*, 149–185.

Carter, N. T., Dalal, D. K., Boyce, A. S., O'Connell, M. S., Kung, M.-C., & Delgado, K. (2014). Uncovering curvilinear relationships between conscientiousness and job performance: How theoretically appropriate measurement makes an empirical difference. *Journal of Applied Psychology, 99*, 564–586.

Cascio, W. F. (2000). *Costing human resources: The financial impact of behavior in organizations* (4th ed.). Cincinnati, OH: Southwestern.

Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4*, 233–264.

Cheung, M. W. L. (2015). *Meta-analysis: A structural equation modeling approach*. Chichester, UK: Wiley.

Chiaburu, D. S., Oh, I.-S., Berry, C. M., Li, N., & Gardener, R. G. (2011). The five factor model of personality traits and organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 96*, 1140–1166.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion related validities. *Personnel Psychology, 63*, 83–117.

Church, A. H., & Rotolo, C. T. (2011). How are top companies assessing their high-potentials and senior executives? A talent management benchmark study. *Consulting Psychology Journal: Practice and Research, 65*, 199–223.

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

Converse, P. D., & Oswald, F. L. (2014). Thinking ahead: Assuming linear versus nonlinear personality-criterion relationships in personnel selection. *Human Performance, 27*, 61–79.

Coward, W. M., & Sackett, P. R. (1990). Linearity of ability-performance relationships: A reconfirmation. *Journal of Applied Psychology, 75*, 297–300.

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

Dahlke, J.A., and Sackett, P. R. (2017). Refinements to the *dMod* class of categorical-moderation effect sizes. *Organizational Research Methods, 21*, 226–234.

DeShon, R. P., & Alexander, R. A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal. *Psychological Methods, 1*, 261–277.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Fife, D. A., Mendoza, J. L., & Terry, R. (2013). Revisiting Case IV: A reassessment of bias and standard errors of Case IV under range restriction. *British Journal of Mathematical and Statistical Psychology, 66*, 521–542.

Finch, D. M., Edwards, B. D., & Wallace, J. C. (2009). Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology, 94*, 318–340.

Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job-analysis-content validity process. In N. Schmitt & W. Borman (Eds.), *Personnel selection in organizations* (pp. 3–34). San Francisco, CA: Jossey-Bass.

Golubovich, J., Grand, J. A., Ryan, A. M., & Schmitt, N. (2014). An examination of common sensitivity review practices in test development. *International Journal of Selection and Assessment, 22*, 1–11.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions* (2nd ed.). Mahwah, NJ: Erlbaum.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council of Education and Praeger Publishers.

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

Hoffman, C. C., & McPhail, S. M. (1998). Exploring options for supporting test use in situations precluding local validation. *Personnel Psychology, 51*, 987–1003.

Hoffman, C. C., Rashovsky, 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: Wiley.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9*, 1–42.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R.P. (2012). Detecting and deterring insufficient effort respond to surveys. *Journal of Business and Psychology, 27*, 99–114.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86*, 897−913.

Humphreys, L. G. (1952). Individual differences. *Annual Review of Psychology, 3*, 131–150.

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. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.

Hunter, J. E., Schmidt, F. L., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75*, 28–42.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological and statistical issues in the study of bias in mental testing. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 41–99). New York, NY: Plenum.

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

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

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

International Test Commission. (2006). International guidelines on computer-based testing and internet-delivered testing. *International Journal of Testing, 6*, 143–172.

International Test Commission. (2013). *ITC guidelines on test use*. Retrieved from https://www.intestcom.org/files/guideline_test_ use.pdf.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50*, 905–925.

Jeanneret, R., & Silzer, R. (Eds.). (1998). *Individual psychological assessment*. San Francisco, CA: Jossey-Bass.

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: Wiley.

Johnson, J. W., & Carter, G. (2010). Validating synthetic validation: Comparing traditional and synthetic validity coefficients. *Personnel Psychology, 63*, 755–795.

Johnson, J. W., Steel, P., Scherbaum, C. A., Hoffman, C. C., Jeanneret, P. R., & Foster, J. (2010). Validation is like motor oil: Synthetic is better. *Industrial and Organizational Psychology*, *3*, 305–328.

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E.R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology, 98*, 875–925.

Keiser, H. N., Sackett, P. R., Kuncel, N. R., & Brothen, T. (2016). Why women perform better in college than admission scores would predict: Exploring the roles of conscientiousness and course-taking patterns. *Journal of Applied Psychology, 101*, 569–581.

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology, 100*, 128–161.

Kwaske, I. H. (2008). Individual assessments for personnel selection: An update on a rarely researched but avidly practiced practice. *Consulting Psychology Journal: Practice and Research, 56*, 186–195.

LaHuis, D. M., & Avis, J. M. (2007). Using multilevel random coefficient modeling to investigate rater effects in performance ratings. *Organizational Research Methods, 10*, 97–107. doi:10.1177/1094428106289394

Levine, J. D., & Oswald, F. L. (2012). O*NET: The Occupational Information Network. In M. A. Wilson, W. Bennett Jr., S. G. Gibson, & G. M. Alliger (Eds.), *The handbook of work analysis in organizations: Methods, systems, applications, and science of work measurement in organizations* (pp. 281–301). New York, NY: Routledge/Psychology Press.

Li, J. C-H., Chan, W., Cui, Y. (2011). Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction. *British Journal of Mathematical and Statistical Psychology, 64*, 367–387.

Little, R J. A., & Rubin, D. B., (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.

Mattern, K. D., & Patterson, B. F. (2013). Test of slope and intercept bias in college admissions: A response to Aguinis, Culpepper, and Pierce (2010). *Journal of Applied Psychology*, *98*, 134–147. doi:10.1037/a0030610

McDaniel, M. A. (2007). Validity generalization as a test validation process. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 159–180). San Francisco, CA: Wiley.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–458.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455.

Morris, S. B., Daisley, R. L., Wheeler, M., & Boyer, P. (2015). A meta- analysis of the relationship between individual assessments and job performance. *Journal of Applied Psychology, 100*, 5–20.

Mueller, L., Norris, D., & Oppler, S. (2007). Implementation based on alternate validation procedures: Ranking, cuts scores, banding, and compensatory models. In S. M. McPhail (Ed.), *Alternative*

validation strategies: Developing new and leveraging existing validity evidence (pp. 349–405). San Francisco, CA: Jossey-Bass.

Murphy, K. R. (2009). Content validation is useful for many things, but validity isn't one of them. *Industrial and Organizational Psychology, 2*(4), 453–464.

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873–900.

Myors, B., Lievens, F., Schollaert, E., Van Hoye, G., Cronshaw, S. F., Mladinic, A., & Sackett, P. R. (2008). International perspectives on the legal environment for selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 206–246.

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

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods, 17*, 372–411.

Newman, D. A., Jacobs, R. R., & Bartram, D. (2007). Choosing the best method for local validity estimation: Relative accuracy of meta-analysis versus a local study versus Bayes-analysis. *Journal of Applied Psychology, 92*, 1394–1413.

Nieminen, L. R. G., Nicklin, J. M., McClure, T. K., & Chakrabarti, M. (2011). Meta-analytic decisions and reliability: A serendipitous case of three independent telecommuting meta-analyses. *Journal of Business and Psychology, 26*, 105–121.

Nye, C. D., and Sackett, P. R. (2017). New effect sizes for tests of categorical moderation and differential prediction. *Organizational Research Methods, 20*, 639–664.

Orr, J. M., Sackett, P. R., and DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44*, 473–486.

Oswald, F. L., & Johnson, J. W. (1998). On the robustness, bias, and stability of statistics from meta-analysis of correlation coefficients: Some initial Monte Carlo findings. *Journal of Applied Psychology, 83*, 164–178.

Oswald, F. L., Putka, D. J., & Ock, J. (2015). Weight a minute, what you see in a weighted composite is probably not what you get. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 187–205). New York, NY: Taylor & Francis.

Oswald, F., Saad, S., & Sackett, P. R. (2000). The homogeneity assumption in differential prediction analysis: Does it really matter? *Journal of Applied Psychology, 85*, 536–541.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65*, 373–406.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13*, 3–29.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (Eds.). (1999). *An occupational information system for the 21st century: The development of O*NET*. Washington, DC: American Psychological Association.

Putka, D. J., & Hoffman, B. J. (2014). The reliability of job performance ratings equals 0.52. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 247–275). New York, NY: Taylor & Francis.

Putka, D. J., Hoffman, B. J., & Carter, N. T. (2014) Correcting the correction: When individual raters offer distinct but valid perspectives. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*, 543–548.

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

Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement, 29*, 213–255.

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences*, *114*, 10870–10875.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Raju, N. S., Anselmi, T. V., Goodman, J. S., & Thomas, A. (1998). The effect of correlated artifacts and true validity on the accuracy of parameter estimation in validity generalization. *Personnel Psychology, 51*, 453–465.

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement, 27*, 52–71.

Raju, N. S., Burke, M. J., & Normand, J. (1990). A new approach for utility analysis. *Journal of Applied Psychology, 75*, 3–12.

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology, 76*, 432–446.

Raju, N. S., Pappas, S., & Williams, C. P. (1989). An empirical Monte Carlo test of the accuracy of the correlation, covariance, and regression slope models for assessing validity generalization. *Journal of Applied Psychology, 74*, 901–911.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement, 31*, 169–180.

Roth, P. L., Purvis, K. L., & Bobko, P. (2012). A meta-analysis of gender group differences for measures of job performance in field studies. *Journal of Management*, *38*, 719–739.

Ryan, A. M., & Sackett, P. R. (1998). Individual assessment: The research base. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment* (pp. 54–87). San Francisco, CA: Jossey- Bass.

Saad, S., & Sackett, P. R. (2002). Examining differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology, 87*, 667–674.

Sackett, P. R. (Ed.). (2009a). *Industrial and Organizational Psychology, 2*(1).

Sackett, P. R. (Ed.). (2009b). *Industrial and Organizational Psychology, 2*(4).

Sackett, P. R., Laczo, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology, 88*, 1046–1056.

Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology, 92*, 538–544.

Sackett, P. R., Putka, D. J., & McCloy, R. A. (2012). The concept of validity and the process of validation. In N. Schmitt (Ed.) *The Oxford handbook of personnel assessment and selection* (pp. 91–118). New York, NY: 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: Prospects in a post-affirmative action world. *American Psychologist, 56*, 302–318.

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

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*, 112–118.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115–129.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmidt, F. L., & Hunter, J. E., (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

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.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology, 66*,166–185.

Schmidt, F. L., Oh, I.-S., Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology, 59*, 281–305.

Schmidt, F. L., Pearlman, K., & Hunter, J. E., (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology, 33*, 705–724.
Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*, 901–912.
Schmitt, N., & Ployhart, R. E. (1999). Estimates of cross-validity for stepwise regression and with predictor selection. *Journal of Applied Psychology, 84*, 50–57.
Schneider, B., & Konz, A. (1989). Strategic job analysis. *Human Resources Management, 28*, 51–63.
Silzer, R., & Jeanneret, R., (2011). Individual psychological assessment: A practice and science in search of common ground. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 270–296.
Society for Industrial and Organizational Psychology. (1987). *Principles for the validation and use of personnel selection procedures*. (3rd ed.). College Park, MD: Author.
Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.
Steel, P. D. G., Huffcutt, A. I., & Kammeyer-Mueller, J. (2006). From the work, one knows the worker: A systematic review of the challenges, solutions, and steps to creating synthetic validity. *International Journal of Selection and Assessment, 14*, 16–36.
Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta- analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology, 87*, 96–111.
Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23*, 565–578.
Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustment for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics, 40*, 604–634.
Weiner, J. & Hurtz, G. (2017). A comparative study of online remote proctored versus onsite proctored high-stakes exams. *Journal of Applied Testing Technology*, *18*, 13–20.
Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.
Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: a critical review and recommendations for practice. *Personnel Psychology, 61*, 871–925.
Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The critical role of the research question, inclusion criteria, and transparency in meta-analyses of integrity test research: A reply to Harris et al. (2012) and Ones, Viswesvaran, and Schmidt (2012). *Journal of Applied Psychology, 97*, 543–549.
Ziegler, M., MacCann, C., & Roberts, R. D. (2012). *New perspectives on faking in personality assessment*. New York, NY: Oxford University Press.

GLOSSARY OF TERMS

Ability

A defined domain of cognitive, perceptual, psychomotor, or physical functioning.

Accommodation

A change in the content, format, and/or administration of a selection procedure made to eliminate an irrelevant source of score variance resulting from a test taker's disability.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Adjusted validity/reliability coefficient

A validity or reliability coefficient—most often a product-moment correlation—that has been adjusted to offset effects of differences in score variability, criterion variability, or unreliability of test and/or criterion. See Restriction of range or variability.

Alternate forms

Two or more versions of a selection procedure that are considered interchangeable in that they measure the same constructs in the same ways, are intended for the same purposes, and are administered using the same directions. Alternate forms is a generic term used to refer to either parallel forms or equivalent forms. Parallel forms have equal raw score means, equal standard deviations, equal error structures, and equal correlations with other measures for any given population. Equivalent forms do not have the statistical similarity of parallel forms, but the dissimilarities in raw score statistics are compensated for in the conversions to derived scores or in form-specific norm tables.

Analysis of work

Any method used to gain an understanding of the work behaviors and activities required, or the worker requirements (e.g., knowledge, skills, abilities, and other personal characteristics), and the context or environment in which an organization and individual may operate. This term subsumes what has earlier and variously been referred to as work and job analysis, and competency modeling.

Assessment

Any systematic method of obtaining information from tests and other sources used to draw inferences about characteristics of people.

Band

A range of scores treated as equivalent. Bands may be developed on an ad hoc basis (e.g., converting scores to categories, such as "high," "medium," and "low") or on the basis of psychometric information (e.g., bands defined by the standard error of measurement).

Battery

A set of selection procedures administered as a unit.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Bayesian approach
Statistical approach in which conclusions are formed by combining prior evidence (e.g., meta-analytic evidence) with new evidence (e.g., evidence from a local validity study).

Bias
In a statistical context, a systematic error in a score. In discussing fairness, bias refers to variance due to contamination or deficiency that differentially affects the scores of different groups of individuals.

Compensatory model
Two or more individual selection procedure component scores (often individual test scores) combined into a composite selection procedure according to some specified formula (including simple summation of scores, unit weighting, and regression weights). As a consequence of combining scores, some compensation for one or more of the constructs measured may occur due to differential performance on the individual selection procedures (i.e., a higher score on one test compensating for a lower score on another test).

Competency
An individual attribute (e.g., knowledge, skill, ability, or other characteristic) relevant to successful performance in a job or job family. A compilation of competencies for a job, job family, or organization is referred to as a competency model.

Composite score
A score that combines scores from several individual selection procedures according to a specified formula.

Concurrent validity evidence
Demonstration of the relationship between a criterion measure, such as job performance and other work outcomes, and scores on selection procedures obtained at approximately the same time.

Confidence interval
An interval between two values on a score scale within which, with specified probability, a score or parameter of interest is expected to lie.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Configural scoring rule (configural scoring)
A rule for scoring a set of two or more elements (such as items or subtests) in which the score depends on a particular pattern of responses to the elements.

Consequence-based evidence
Evidence that consequences of selection procedure use are consistent with the intended meaning or interpretation of the selection procedure.

Construct
A concept or characteristic of individuals inferred from empirical evidence and theory.

Construct irrelevance
The extent to which scores on a predictor are influenced by factors that are irrelevant to the construct. Such extraneous factors dis- tort the meaning of scores from what is implied in the proposed interpretation.

Contamination
Systematic variance that is irrelevant to the intended meaning of the measure.

Content domain
The set of behaviors, knowledge, skills, abilities, attitudes, or other characteristics to be measured by a test, represented in detailed test specifications, and often organized into categories by which items are classified.

Content-based validity evidence
Demonstration of the extent to which content on a selection procedure is a representative sample of work-related personal characteristics, work performance, or other work activities or outcomes.

Convergent evidence
Evidence based on the relationship between test scores and other measures of the same or related construct.

Correlation
The degree to which two sets of measures vary together.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Criterion
A measure of work performance or behavior, such as productivity, accident rate, absenteeism, tenure, reject rate, training score, and supervisory ratings of job relevant behaviors, tasks, or activities.

Criterion-related validity evidence
Demonstration of a statistical relationship between scores on a predictor and scores on a criterion measure.

Criterion relevance
The extent to which a criterion measure reflects important work performance dimensions or other work outcomes.

Critical score
A specified point in a distribution of scores at or above which candidates are considered successful in the selection process. The critical score differs from cutoff score in that a critical score is by definition criterion referenced (i.e., the critical score is related to a minimally acceptable criterion) and is the same for all applicant groups.

Cross-validation
The application of a scoring system or set of weights empirically derived in one sample to a different sample from the same population to investigate the stability of relationships based on the original weights.

Cutoff score
A score at or above which applicants are selected for further consideration in the selection process. The cutoff score may be established on the basis of a number of 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 based on their needs.

Deficiency
Failure of an operational predictor or criterion measure to fully rep- resent that conceptual predictor or criterion domain intended.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Derived score
A score that results from a numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores) of the original selection procedure score.

Differential item functioning
A statistical property of a test item in which different groups of test takers who have the same standing on the construct of measurement have different average item scores or, in some cases, different rates of endorsing various item options. Also known as DIF.

Differential prediction
The case in which use of a common regression equation results in systematic nonzero errors of prediction for subgroups.

Discriminant evidence
Evidence indicating whether two tests interpreted as measures of different constructs are sufficiently independent (uncorrelated) to be considered two distinct constructs.

Fairness
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; there is not, however, agreement that the term "fairness" can be uniquely defined in terms of any of these issues.

Generalized evidence of validity
Evidence of validity that generalizes to setting(s) other than the set- ting(s) in which the original validation evidence was documented. Generalized evidence of validity is accumulated through such strategies as transportability, synthetic validity/job component validity, and meta-analysis.

Imputation
A process for inferring values for missing variables. Modern imputation methods are widely seen as preferable to dropping cases with missing values from analysis.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Individual assessment
An integrative process in which multiple predictors (commonly tests, work samples, and an interview) are administered to an individual, with the results integrated judgmentally or mechanically by an assessor, commonly resulting in a narrative report about the individual.

Internal consistency reliability
An indicator of the reliability of a score derived from the statistical interrelationships of responses among item responses or scores on different parts of an assessment.

Internal structure validity evidence
Demonstration of the degree to which psychometric and statistical relationships among items, scales, or other components within a selection procedure are consistent with the intended meaning of scores on the selection procedure.

Interrater agreement
The consistency with which two or more judges rate the work or performance of examinees.

Item
A statement, question, exercise, or task on a selection procedure for which the test taker is to select or construct a response, or perform a task.

Item response theory (IRT)
A mathematical model of the relationship between performance on a test item and the test taker's standing on a scale of the construct of measurement, usually denoted as θ. In the case of items scored 0/1 (incorrect/correct response) the model describes the relation- ship between θ and the item mean score (P) for test takers at level θ, over the range of permissible values of θ. In most applications, the mathematical function relating P to θ is assumed to be a logistic function that closely resembles the cumulative normal distribution.

Job analysis
See Analysis of work.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Job component validity
See Synthetic validity evidence.

Job description
A statement of the work behaviors and activities required or the worker requirements (e.g., knowledge, skills, abilities, and other personal characteristics).

Job knowledge
Information (often technical in nature) needed to perform the work required by the job.

Job relatedness
The inference that scores on a selection procedure are relevant to performance or other behavior on the job; job relatedness may be demonstrated by appropriate criterion-related validity coefficients or by gathering evidence of the job relevance of the content of the selection instrument, or of the construct measured.

KSAOs
Knowledge, skills, abilities, and other personal characteristics required in completing work in the context or environment in which an organization and individual may operate.

Local evidence
Evidence (usually related to reliability or validity) collected in a single organization or at a specific location.

Local study (local setting)
See Local evidence.

Measurement bias
See Bias.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Meta-analysis
A statistical method of research in which the results from sever- al independent studies of comparable phenomena are combined to estimate a parameter or the degree of relationship between variables.

Moderator variable
A variable that affects the strength, form, or direction of a predictor– criterion relationship.

Modification/modified tests
A change in test content, format (including response formats), and/ or administration conditions that is made to increase accessibility for some individuals but that also affects the construct measured and, consequently, results in scores that differ in meaning from scores from the unmodified assessment.

Multiple-hurdle model
The implementation of a selection process whereby two or more separate procedures must be passed sequentially by the job applicant.

Normative
Pertaining to norm groups or the sample on which descriptive statistics (e.g., mean, standard deviation) or score interpretations (e.g., percentile, expectancy) are based.

Norms
Statistics or tabular data (often raw and percentile scores) that summarize performance of a defined group on a selection procedure.

Objective
Pertaining to scores obtained in a way that minimizes bias or error due to variation in sources deemed to be irrelevant (e.g., observers, scorers, settings).

Operational setting
The specific organization, work context, applicants, and employees to which a selection procedure is applied.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Outlier
A data point in a predictor or criterion distribution substantially removed from other data points (i.e., an extreme score). Outliers can have undue influence on a summary statistic of interest (e.g., a correlation); they merit careful scrutiny to ascertain whether they are erroneous and need to be removed or replaced (e.g., a misscored test, an equipment failure).

Pareto-optimization
A method used in settings where one is pursuing two or more objectives (e.g., validity maximization, group difference minimization, cost minimization) to identify the highest level of one objective attainable at a given level of another objective.

Personal characteristics
Traits, dispositions, or other features that describe individuals.

Population
The universe of cases from which a sample is drawn and to which the sample results may be projected or generalized.

Power
The probability that a statistical test will yield statistically significant results if an effect of specified magnitude exists in the population.

Predictive bias
The systematic under- or overprediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to criterion performance.

Predictive validity evidence
Demonstration of the relationship between selection procedure scores and some future work behavior or work outcomes.

Predictor
A measure used to predict criterion performance.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Predictor–criterion relationship
The relationship between a predictor and external criteria (e.g., job performance, tenure) or other predictors and measures of the same construct.

Professional judgment
Evaluations and decisions that are informed by and representative of the profession's commonly accepted empirical, methodological, and experiential knowledge base.

Psychometric
Pertaining to the measurement of psychological characteristics such as aptitudes, personality traits, achievement, skill, and knowledge.

Reliability
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) in the application of a measurement procedure.

Reliability estimate
An indicator that reflects the degree to which scores are free of measurement error variance.

Response process
A component, usually hypothetical, of a cognitive account of some behavior, such as making an item response.

Restriction of range or variability
Reduction in the observed score variance of a sample, compared to the variance of an entire population, as a consequence of constraints on the process of sampling.

Sample
A selection of a specified number of entities called sampling units (test takers, items, etc.) from a large specified set of possible entities, called the population. A random sample is a selection according to a random process, with the selection of each entity in no way dependent on the selection of other entities. A stratified random sample is a set of random samples, each of a

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

specified size, from several different sets, which are viewed as strata of the population.

Sampling bias
The extent to which a sampling process introduces systematic misrepresentation of the intended population.

Score
A number describing the assessment of an individual; a generic term applied for convenience to such diverse kinds of measurements as tests, production counts, absence records, course grades, ratings, or other selection procedures or criterion measures.

Selection procedure
An assessment instrument used to inform a personnel decision such as hiring, promotion, or placement.

Selection procedure (test) user
The individual(s) or organization that selects, administers, and scores selection procedures (tests) and usually interprets scores that are obtained for a specified purpose in a defined organizational context.

Sensitivity review
A process of reviewing test items to identify content that might be interpreted differently or be offensive to members of various groups of test takers.

Shrinkage formula
An adjustment to the multiple correlation coefficient for the fact that the beta weights in a prediction equation cannot be expected to fit a second sample as well as the original.

Skill
Level of proficiency on a specific task or group of tasks.

Standardization
(a) In test construction, the development of scoring norms or protocols based on the test performance of a sample of individuals selected to be representative of the candidates who will take the test for some defined use; (b) in

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

selection procedure administration, the uniform administration and scoring of a selection procedure in a manner that is the same for all candidates.

Standard score
A derived score resulting in a distribution of scores for a specified population with specified values for the mean and standard deviation. The term is sometimes used to describe a distribution with a mean of 0.0 and a standard deviation of 1.0.

Statistical power
See Power.

Statistical significance
The finding that statistical estimates are inconsistent with a null hypothesis at some specified probability level.

Subject matter experts (SMEs)
Individuals who have thorough knowledge of the work behaviors, activities, or responsibilities of job incumbents and the KSAOs needed for effective performance on the job.

Synthetic validity evidence
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); also referred to as "job component validity evidence."

Systematic error
A consistent score component (often observed indirectly) not related to the intended construct of measurement.

Test
A measure or procedure in which a sample of an examinee's behavior in a specified domain is obtained, evaluated, and scored using a standardized process.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Test development
Process through which a test or other predictor is planned, constructed, evaluated, and modified, including consideration of content, format, administration, scoring, item properties, scaling, and technical quality for its intended purpose.

Test specifications
Documentation of the purpose and intended uses of a test as well as of the test's content, format, length, psychometric characteristics (of the items and test overall), delivery mode, administration, scoring, and score reporting.

Trait
An enduring characteristic of a person that is common to a number of that person's activities.

Transportability
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.

Type I and Type II errors
Errors in hypothesis testing; Type I error involves concluding that a significant relationship exists when it does not; Type II error involves concluding that no significant relationship exists when it does.

Validation
The process by which evidence of validity is gathered, analyzed, and summarized.

Validity
The degree to which accumulated evidence and theory support spe- cific interpretations of scores from a selection procedure entailed by the proposed uses of that selection procedure.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press

Validity argument
An explicit scientific rationale for the conclusion that accumulated evidence and theory support the proposed interpretation(s) of selection procedure scores entailed by the proposed uses.

Validity coefficient
A measured coefficient reflecting the relationship between a selection procedure and a criterion that provides evidence about the validity of the selection variable.

Validity evidence
Any research or theoretical evidence that pertains to the interpretation of predictor scores, or the rationale for the relevance of the interpretations, to the proposed use.

Validity generalization
Justification for the use of a selection procedure or battery in a new setting without conducting a local validation research study. See generalized evidence of validity.

https://doi.org/10.1017/iop.2018.195 Published online by Cambridge University Press