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 that supports those inferences. Three sources of evidence will be described: namely, evidence of validity based on relationships with measures of other variables, evidence based on content, and evidence based on the internal structure of the selection procedure. 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 concept a selection procedure is intended to measure. At times the construct is not fully understood or well articulated. However, relationships among variables reflect their underlying constructs. For example, a predictor generally cannot correlate with a criterion unless to some extent one or more of the same constructs underlie both variables. Consequently, validation efforts based on constructs apply to all investigations of validity.

Principles for using a criterion-related strategy to accumulate validity evidence in employment settings are elaborated below. While not explicitly discussed, the following principles also apply to research using variables other than job performance criteria (e.g., convergent and discriminant evidence). 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

Personnel selection procedures are used to predict future performance or other work behavior. Evidence for criterion-related validity typically consists of a demonstration of a relationship (via statistical significance testing or establishing confidence intervals) between the results of 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 test use, job information, and existing knowledge regarding test validity.

A standardized procedure is one that presents and uses consistent directions and procedures for administration, scoring, and interpretation. Standardized predictors and 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 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) must be obtained or developed. Of these characteristics, the most important is relevance. A relevant criterion is one that reflects the relative standing of employees with respect to important work behavior(s) or outcome measure(s). If such a 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 work 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.

A number of factors related to statistical power can influence the feasibility of a criterion-related study. Among these factors are the degree (and type) of range restriction in the predictor or the criterion, reliability of the criterion, and statistical power. Sample size, the statistic computed, the probability level chosen for the confidence interval, and the size of the predictor-criterion relationship determine the confidence interval around the validity estimate. In practice, these threats and factors occur in varying levels that, in combination, affect power and the precision of estimation. Therefore, statistical power and precision of estimation should be carefully considered before undertaking a criterion-related validity study, and if a study is conducted, the report should include information relevant to power estimation.

Design and Conduct of Criterion-Related Studies

If a criterion-related strategy is feasible, attention is then directed to the design and conduct 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 survival 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 collected, usually 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 in the time of predictor data collection relative to a selection decision or the time at which employees start in a job—before, simultaneously, shortly after, or after a substantial time period in the job. 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 may depend upon the predictor construct. For tests of cognitive abilities, 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). Findings regarding the comparability of 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 or work 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. Where employees do not advance to the higher level in sufficient numbers, assessment of candidates for such work still may 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.

In some organizations, work changes so rapidly or is so fluid that validation with regard to performance in one or more "target" job(s) is impossible; successful performance is more closely related to abilities that contribute broadly to organizational effectiveness. In such instances, the researcher may accumulate evidence in support of the relationship between predictor constructs (e.g., flexibility, adaptability, team orientation, learning speed, and capacity) and organization-wide criteria (such as working effectively under very tight deadlines).

Criterion Development

In general, if criteria are chosen to represent work-related activities, behaviors or outcomes, 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 analysis is usually not necessary, though an understanding of the work and its context is 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. 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 researcher should 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, and 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 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 upon 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 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. While 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 researcher to control statistically for them; in other cases, special diligence in the construction of the measurement procedure and in 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 work behaviors or outcomes critical to job performance.

Criterion bias. Criterion bias is systematic error resulting from criterion contamination or deficiency that differentially affects 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. The possibility of criterion bias must be anticipated. The researcher should protect against bias insofar as is feasible and use professional judgment when evaluating the data.

Criterion reliability. When estimated by appropriate measures, criterion measures should exhibit reliability. For examples of appropriate and inappropriate uses of a variety of reliability estimates see Hunter and Schmidt (1996). 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.

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, 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 can be preferable for use in validation studies to those routinely collected for administrative use (Jawahar & Williams, 1997).

Choice of Predictor

Many factors, including professional judgment and the proposed use of the selection procedure, influence the choice of the predictor(s).

Selecting predictors. Variables chosen as predictors should have an empirical, logical, or theoretical 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, 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 the researcher's scientific knowledge without regard for personal bias or prejudice. Therefore, the researcher's choice of specific predictors should be based on theory and the findings of relevant research rather than personal interest or mere familiarity.

Predictor contamination. As with criteria, a predictor measure is contaminated to the extent that it includes extraneous, systematic variance. A number of factors can contribute to predictor contamination including unstandardized administrative procedures and irrelevant content. Some procedures, such as unstructured interviews, may be more susceptible than others to predictor contamination. Efforts should be made to minimize predictor contamination.

Predictors and selection decision strategies. Outcomes of decision strategies 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 multiple predictors (e.g., 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, support for the judgment should be based on validity evidence for the specific components.

Predictor reliability. Predictor reliability, like criterion reliability, should be estimated whenever feasible. Predictor reliability should be estimated through appropriate methods and should be sufficiently high to warrant use. Predictor, like criterion, reliability places a ceiling on any validity estimate.

Choice of Participants

Samples should be chosen with the intent to generalize to the selection situations of interest. The impact of characteristics such as demographics, motivation, ability, and experience on predictor-criterion relationships, and hence on this generalization, is an empirical question. No variable should be assumed to moderate validity coefficients in the absence of explicit evidence for such an effect.

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. Researchers need to ensure that the statistics used are appropriate. Moreover, as with the choice of criterion or predictor variables, the researcher should not choose a data analysis method simply because the computer package for it is readily available. Researchers 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 or confidence associated with predictor-criterion relationships. Effect size estimates and confidence intervals can be useful in making professional judgments about the strength of predictor-criterion relationships (Schmidt, 1996). Other approaches such as expectancy tables are also useful in many situations, particularly if the assumptions of a correlational analysis are not met.

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. Researchers should give at least equal attention to the risks of Type II error.

Reports of any analysis should provide information about the nature of the predictor-criterion relationship and how it might be used in prediction. The information should include number of cases, measures of central tendency, characteristics of distributions, and variability for both predictor and criterion variables, as well as the interrelationships among all variables studied.

Adjustments to validity estimates. Researchers should obtain as unbiased an estimate as possible of the validity of the predictor in the population in which it is used. Observed validity coefficients may underestimate the predictor-criterion relationship due to the effects of range restriction and unreliability in the predictors or criteria. When range restriction causes underestimation of the validity coefficient, a suitable bivariate or multivariate adjustment should be made when the necessary information is available. Adjustment of the validity coefficient for criterion unreliability should be made if an appropriate estimate of criterion reliability can be obtained. Researchers 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.

When adjustments are made, both unadjusted and adjusted validity coefficients should be reported. Researchers should be aware that the usual tests of statistical significance do not apply to adjusted coefficients such as those adjusted for restriction of range and/or criterion unreliability (Bobko & Riecke, 1980; Raju & Brand, in press; Raju, Burke, Normand, & Langlois, 1991). The adjusted coefficient is generally the best point estimate of the population validity coefficient; confidence intervals around it should be computed. 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.

Combining predictors and criteria. Where predictors are used in combination, researchers should consider and document the method of combination. Predictors can be combined using weights derived from a multiple regression analysis (or another appropriate multivariate technique), unit weights, unequal weights that approximate regression weights, weights that are determined from work-analytic procedures, or weights based on professional judgment. Generally, after cross-validation, the more complex weighting procedures offer no or only a slight improvement over simple weighting techniques (Aamodt & Kimbrough, 1985). When combining scores, care must be taken to ensure that differences in the variability of different predictors do not lead to over- or underweighting of one or more predictors.

Selection procedures that have linear relationships with work performance can be combined for use in either a linear manner (e.g., by summing scores on different selection procedures) or in a configural manner (e.g., by using multiple cutoffs). The researcher should be aware of the administrative, legal, and other implications of each choice. When configural selection rules are used, a clear rationale for their use should be provided (e.g., meeting larger organizational goals or needs, administrative convenience, or reduced testing costs).

Similarly, if the researcher combines scores from several criteria into a composite score, there should be a rationale to support the rules of combination and the rules of combination should be described. Usually, it is better to assign unit or equal weights to the several criterion components than to attempt to develop precise empirical weights. When measures are combined, researchers should recognize that effective weights (i.e., the contributions of the various components to the variance of the composite) are a function of a variable's standard deviation and are unlikely to be the same as the nominal weights.

Cross-validation. Researchers 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. The assignment of either rational or unit weights to predictors does not result in shrinkage in the usual sense. Where a smaller number of predictors is selected for use based on sample validity coefficients from a larger number included in the study, shrinkage formulas can be used only if the larger number is entered into the formula as the number of predictors, though this will produce a slightly conservative estimate of the cross-validated multiple correlation.

Documenting and interpreting validation analyses. The 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 (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, 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 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).

The content-based selection procedures discussed here are those designed as representative samples of the most important work behaviors, activities, and/or worker KSAOs drawn from the work domain and defined by the analysis of work. The content of the selection procedure includes the questions, tasks, themes, format, wording, and meaning of items, response formats, 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, the interference of irrelevant content, the availability of qualified and unbiased subject matter experts, and cost and time constraints.

The researcher should consider whether the work and the worker requirements are reasonably stable. 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 the content-based validation study is closely related to the qualifications of the subject matter experts (SMEs). SMEs define the work domain and participate in the analysis of work by identifying the important work behaviors, activities, and worker KSAOs. The experts should have thorough knowledge of the work behaviors and activities, responsibilities of the job incumbents, and the KSAOs prerequisite to effective performance on the job. The SMEs should include persons who are fully knowledgeable about relevant organizational characteristics such as shift, location, type of equipment used, and so forth. A method for translating subject matter expert 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, procedures and criteria for rating each aspect should be standardized and delineated.

Cost and time constraints can affect the feasibility 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. In others, developing and assessing the reliability of the procedure may take too long because samples are too small or the behavior is not easily measured using this strategy.

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. This involves choosing subject matter experts, defining the content to be included in the selection procedure, developing the selection procedure, 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 researcher should indicate what important work behaviors, activities, and worker KSAOs are included in the domain, describe how the 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 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 have all the important work behaviors, activities, and/or KSAOs before selection into the job or whether basic or advanced training will be provided after selection. If the intended purpose of the selection procedure is to hire or promote individuals into jobs for which no advanced training is provided, the researcher 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 would be those needed to succeed in a training program. Because the intended purpose is to hire or promote individuals who have the prerequisite work behaviors, activities, and/or 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.

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. For example, course titles and job titles may not give an adequate 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.

Choosing the Selection Procedure

The development or choice of a selection procedure usually is restricted to important or frequent behaviors and activities or to prerequisite KSAOs. The researcher should have 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 researcher 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. The rationale underlying the sampling should be documented.

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 a selection procedure has fidelity to exact job components, the more likely it is that the 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, the researcher may wish to develop a selection procedure that has less specificity. For example, in developing a selection procedure for the job of word processor, the procedure may exclude content such as "demonstrates proficiency with a particular word processing program" and instead include content that is less specific, such as "demonstrates proficiency with word processing principles and techniques."

The degree to which the results of validation studies can be generalized depends in part on the specificity of the selection procedure and its applicability across settings, time, and jobs. While 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.

Procedural Considerations

The researcher 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. On the other hand, including a work behavior or activity in the content-based selection procedure may introduce 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. Yet, 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 researcher 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. If ratings from more than one rater are used to evaluate performance on a simulation or work sample, the researcher should evaluate inter-rater agreement in operational use.

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 researcher 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.

However, if the conceptual framework posits a single dimension or construct, one should strive for a high level of homogeneity among the components, which can be evaluated in terms of various internal consistency estimates of reliability. If the intent of the conceptual framework requires a more complex internal structure, overall internal consistency might not be an appropriate measure. For example, the internal consistency reliability estimate for a performance rating form involving several supposedly unrelated scales might only represent halo effect.

When scoring involves a high level of judgment on the part of those doing the scoring, indices of inter-rater or scorer consistency, such as generalizability coefficients or measures of inter-rater agreement, may be more appropriate than internal consistency estimates.