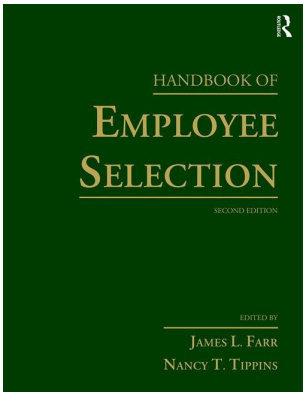

Handbook of Employee Selection

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

Choosing a Psychological Assessment

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-18
Michael J. Zickar, Jose M. Cortina, Nathan T. Carter
Published online on: 22 Mar 2017

18

Choosing a Psychological Assessment

Reliability, Validity, and More

MICHAEL J. ZICKAR, JOSE M. CORTINA, AND NATHAN T. CARTER

Consumers of psychological tests have a large number of tests to choose from these days and often have little factual information that can be used to pick a particular test. Googling *tests for hiring employees* results in roughly 48,200,000 hits, *personality tests* has 10,300,000 hits, *integrity tests* 7,220,000 hits, and *tests to hire salespeople* 693,000 hits. Clicking on some of these sites, we found claims such as "Never hire a bad salesperson again"; "Our sales assessment validity is backed by brain research. No other sales assessment is"; and "You can start testing your job candidates today—it's that quick and easy!" These quick-and-easy fool-proof solutions might seem attractive to employers who need a hiring solution but have little expertise to choose among tests and vendors. Fortunately, industrial-organizational psychologists have conducted more than 100 years of research and practice that can help people choose tests appropriate for a particular job. In this chapter, we review some of the key concepts underlying the science of testing, particularly reliability and validity. Then we discuss how employers can use these concepts, as well as relevant information that should be provided by any reputable test developer (but which often is not!) to choose a particular test best suited for particular needs.

RELIABILITY

Even though reliability theory is one of the first topics covered in graduate measurement courses, it is one of the most misunderstood topics. Most students learn about reliability in the context of classical test theory and are deceived by the simple formula $X = T + E$, where an observed score is mysteriously parsed into a true score, T, and error, E. Students who delve a little deeper into reliability theory realize that there is little "true" about the true score, and often what they think is error is not. What is often lost with novice researchers is that the source of error that is identified in a particular measure is dictated by the type of reliability coefficient calculated. In this section, we focus on three common types of error that are often present in psychological measures: error associated with different items, error associated with different raters, and error due to issues related to momentary, time-limited phenomena. As a test consumer, you will want to pay keen attention to the level of reliability reported as well as the type of coefficients presented. Also, as we will discuss, the level of reliability needed will be dictated partially by how you plan to use the test.

Error Due to Items

When one of the authors [Zickar] took the GRE Psychology Subject exam, there was an item that asked something like "What was the name of the first computerized learning system?" He got that item correct, not because he knew a lot about psychology, but because he had been an undergraduate student at the University of Illinois where nearly every freshman had to use the computerized system PLATO to learn chemistry, mathematics, or economics. In a sense, Zickar got one extra item correct because of the unique content of one item that was biased in his favor. Students from other universities across the country and world were not so lucky.

Internal consistency measures of reliability, such as the popular coefficient alpha, are largely a function of inter-item covariances. As items relate more strongly with each other, holding all else equal, internal consistency reliability increases. Tests that have a large percentage of items that are dominated by unique variance will be more susceptible to error due to individual items and, therefore, have a lower internal consistency reliability. In addition, all else being equal, scales with few items are more susceptible to the unique influence of individual items. For example, if the GRE Psychology test had only three items and one of them was related to the PLATO learning system, Zickar's score would have been greatly inflated. As it was, the small increase that he got by having "inside information" on that one item probably made little difference on his overall test score, given the large number of items on the subject test.

Although it might be tempting to eliminate error due to the uniqueness of individual items by administering a scale consisting of items that ask the same item in slightly different ways, this approach runs the risk of compromising measure sufficiency. Research has also shown that, although asking the same item in slightly different ways may result in a high internal consistency index, the resulting narrowness of the scale may result in reduced validity (see Roznowski & Hanisch, 1990). A better way to minimize the error associated with unique item content is to increase the number of items, while making sure that individual items do not share construct-irrelevant components (i.e., are contaminated). As a test consumer, if measurement precision is of key importance, make sure to avoid tests that report high reliabilities but are extremely short.

Error Due to Raters

The classic Japanese movie *Rashomon* is a good way to understand the nature of rater error. In that movie, several observers witness the same crime, though when they retell what they observe, their retellings are vastly different. When observing behavior or coding written behavior, observers interpret information differently. Some raters are more lenient, but others are more stringent. Some interviewers might give preference to blondes, whereas others may unconsciously give high ratings to people who wear blue ties. Differences in rater behavior can sometimes be reduced by providing training, though given the different ways in which individuals view the world, these differences are unlikely to be completely eliminated.

Most tests that will be considered for selection will not have this source of error given that most pre-employment tests rely on objectively scored items that require no individual rater to make a judgment. Tests that involve projective items as well as work samples and standardized interviews, however, both require individual raters to interpret test behaviors, thus potentially introducing this type of error. When raters are involved in judging job-related variables, research has shown that this type of error can be significant. For example, Woehr, Sheehan, and Bennett (2005) found that unique, idiosyncratic source-specific factors were responsible for two-thirds of the variance in performance ratings. Employment interview researchers have also demonstrated the inter-rater reliability of interviewees is typically fairly low (see Conway, Jako, and Goodman, 1995).

There are many ways to reduce the amount of error related to raters. If at all possible, it is important to standardize the nature of information that different raters observe. In addition, providing frame-of-reference training (e.g., Conway et al., 1995) that attempts to provide common standards of comparison might help improve inter-rater reliability. Computerized scoring

algorithms are used by large-scale testing companies to interpret and score written essays in the GRE and other certification tests, thereby eliminating the possibility of rater unreliability. If you cannot reduce error by standardizing information, the best way to reduce it is to increase the number of raters, thereby reducing the amount of error through aggregation in the same way that increasing the number of items reduces internal inconsistency. Taking an average of a large number of raters will cancel out the positive and negative errors associated with individual raters. In terms of choosing tests that require raters (e.g., projective tests, standardized oral interviews), make sure that you find out how many raters are needed to ensure reasonable reliability. See Greguras and Robie (1998) for procedures on how to determine the appropriate number of raters. For some tests, the demands needed to achieve acceptable reliability may be prohibitive or too costly.

Error Due to Momentary Time-Limited Factors

There are lots of reasons that scores on tests may vary from one testing administration to another. Weird things can happen in testing administrations. For example, in an entrance testing session, one of our students witnessed another student vomiting (perhaps because of nervousness) in the vicinity of other students. It is possible that the students who were near the projectile vomiter would score lower on that particular administration compared to administrations at other times. Although that is a bizarre, rare event, many time-limited errors can be due to test administrators, the testing environment, or temporary issues related to the test taker.

Test administrators can give too much time or not enough time. They can be unnecessarily harsh and intimidating, thus increasing test anxiety, or they can be so welcoming and pleasant that test takers do much better than normal. Administrators can give erroneous instructions or mishandle timing devices or they can inadvertently give away correct answers for difficult items.

Related to the testing environment, the heating or air conditioning system can fail. A picture in the testing room of the previous school principal might remind a single test taker of a mean uncle who used to taunt him about how he would be a failure for his whole life, thus prompting that student to do poorly. Or that student may be given a comfortable chair that fits him just right. In an unproctored Internet testing environment, the test takers can choose where they take their test, further adding to standardization problems (see Tippins et al., 2006).

Test takers can have unique things happen to them on one testing occasion that might not happen to them on another testing occasion. Test takers can be hungover or sick with the flu. They could have just been dumped by a fiancée. They may have had an especially good night's sleep or an especially poor one.

Regardless of the source of time-limited momentary effects, these events are unlikely to happen if the test taker were to take the test at a different time. Events that are predictable and are expected to occur *every time* a respondent takes a test would not be considered error even if they were distinct from the construct that the test is measuring. For example, test anxiety would not be considered error in the context of test-retest reliability if the test taker experienced the same level of anxiety each time s/he took a math test, even though test anxiety is clearly a different construct than mathematics ability. Although it would be impossible to eliminate all sources of time-limited error, it is possible to minimize the effects of error due to administration and environment by having standardized instructions and environments for test takers.

Measures of reliability sensitive to time-limited factors, such as test-retest reliability, rest on the assumption that all score differences across two separate testing administrations are due to momentary time-limited errors. Of course, differences in scores across two administrations can be due not only to time-limited errors such as the ones mentioned but also to true change in the underlying construct. For example, dramatic changes in vocabulary test scores given across six months may be due to true growth in vocabulary rather than momentary, time-limited errors (see Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007). As a test user, you have a lot of control over minimizing this source of error. Standardizing test administration for all examinees is an important step to ensure that error under your control is minimized. Some test

administrators have found that providing video-based test instructions and introductions are helpful so that everyone has precisely the same instructions. Also, making sure applicants are isolated from outside distractions is especially important. Make sure to follow administration instructions as dictated by test manuals. One can never eliminate this type of error, but standardized testing experiences for all test takers helps minimize this error. In unproctored Internet testing, instructions should include making sure the testing environment is free from distractions and that no outside help is used to solve items.

Conclusions on Sources of Error

Error can contaminate our measures and compromise measurement. This can wreak havoc in work contexts such as top-down selection, where small differences in test scores might have significant consequences. Test developers are often unaware of the amount of error that their tests are likely to generate because they have used a single operationalization of reliability, generally internal consistency, which is sensitive to one source of error but ignores other sources of error. One way to calculate the effects of multiple sources of error is through application of Generalizability Theory (GT), which is an ANOVA-based approach that can be used to determine the magnitude of various sources of error simultaneously. GT approaches to error estimation are used less frequently than traditional approaches to reliability because they require more extensive data collections (especially compared to internal consistency analyses). For readers more interested in GT, we refer them to Chapter 1 in this Handbook (Putka), as well as to Shavelson and Webb (1991). Test developers rarely report GT coefficients, however, so as a test user, you are likely forced to rely on individual reliability coefficients such as test-retest and internal consistency coefficients.

How to Use Reliability Information in Choosing a Test

First, all reputable testing firms should be able to provide reliability information about the tests they are selling. Most tests report only a single reliability coefficient, typically coefficient alpha, that is sensitive to error due to items but ignores error due to time or raters. That type of reliability coefficient may be useful for certain purposes, though less useful for others. Any test that requires subjective scoring (e.g., structured interviews and projective tests) by a rater should report consistency across raters. In general, you will want to see multiple types of reliability presented.

In addition to the types of reliability reported, desired levels of reliability may differ depending on the way the test is used. Remember that reliability is related to the uncertainty of a test score, which is often best quantified by the standard error of measurement (SEM). Tests that play an important part in determining whether somebody is hired or promoted need to have higher levels of reliability than tests that might be given little weight or used as a rough screening device perhaps early in the process. Therefore, if you are using a single test to hire your next CEO, that test should have extremely high reliability (and validity!), but if you were using a battery of tests to screen out the bottom 20% of candidates for an entry-level position, lower levels of reliability might be tolerated. In addition, tests that have less significant consequences, such as tests used for staff development, can have lower levels of reliability. Finally, tests with somewhat lower reliabilities that are averaged across a group of individuals might be tolerated. For example, if you are using a cognitive ability test to determine whether applicants from a particular region score higher than another region, lower levels of reliability can be tolerated given that errors within individuals may cancel out.

As noted here, the target level of reliability depends on the particular usage of a test; therefore, it is difficult to give a single value of reliability needed to use a test. One generalization that is safe to make, though, is that any test publisher who does not make appropriate reliability information available should be avoided!

VALIDITY

Our review of validity focuses on sufficiency and contamination, two concepts that are deemed critical for demonstrating evidence of content validity. Several types of evidence can be collected to support test validation, including evidence from criterion-related validity, content-oriented validity, and construct validity. We do believe that all forms of validity are related to each other and the concepts of sufficiency and contamination, although most often used in discussion of content validity, are relevant to all forms of validity (see Landy, 1986). For example, the SIOP *Principles* (SIOP, 2003) discuss contamination in the context of content validity, criterion-related validity, and item bias. We believe that understanding these fundamental issues related to test validity is important for test consumers in order to make better choices about which tests to use.

Sufficiency

In discussions of validity, it is often asked whether the test in question covers all of the ground that it should. For example, measures of job performance have been expanded to accommodate dimensions that have been added to models of job performance (e.g., adaptive performance; Pulakos, Arad, Donovan, & Plamondon, 2000), and measures of intelligence have been expanded to accommodate dimensions that have been added to models of intelligence (e.g., practical intelligence; Sternberg, Wagner, Williams, & Horvath, 1995).

The criticism to which these expansions responded was that prior selection measures (both predictors and criteria) often failed to capture the entire domain of the construct being measured, (i.e., they were insufficient). Consider the example of adaptive performance. Pulakos et al. (2000) argued that employees often engaged in various work behaviors that contributed to organizational effectiveness but were not recognized by existing models and measures of job performance. Specifically, they suggested that categories of work behavior such as Handling Crises and Cultural Adaptability were crucial to effectiveness in some organizations but were conspicuously absent from existing measures of job performance.

One might conclude from this that existing measures were insufficient. It would be more appropriate, however, to say that existing *models* of performance were insufficient, and that the measures merely reflected the inferior models on which they were based. If we assume that a measure is unidimensional, then insufficiency can only indicate factorial complexity at the model level. It seems more parsimonious, then, to stipulate that sufficiency is a property of conceptual models rather than one of measures. Once a model has been deemed to cover the full breadth of its domain (e.g., a performance model that consists of technical performance, contextual/citizenship performance, adaptive performance, interpersonal performance, etc.), then unidimensional scales measuring each factor can be developed. Reliability then reflects proportion of true score variance, and validity represents lack of contamination (i.e., the introduction of construct-irrelevant variance into a measure).

This position may seem at odds with the Standards for Educational and Psychological Testing. In the section on content-related evidence, it is stated that

> construct underrepresentation . . . may give an unfair advantage or disadvantage to one or more subgroups. Careful review of the construct and test content domain by a diverse panel of experts may point to potential sources of irrelevant difficulty (or easiness) that require further investigation.
>
> (AERA et al., 1999, p.12)

There are several observations to be made about this passage. The first is that sufficiency is inextricably intertwined with content-related validity evidence. Evidence of insufficiency comes from a comparison of test content to the "content domain." Omissions suggest insufficiency. Second, the solution that is offered in the passage has to do with contamination rather than sufficiency. This may have been incidental, but it may also have been due to an inability to refer to insufficiency without also referring to deficiencies in the definitions of the construct of interest and of the domain of items that apply to it. Third, this passage is representative of the

Standards as a whole in that nowhere in the Standards are issues of sufficiency raised without reference to content-oriented approaches to validity.

Although the term "sufficiency" does not appear in the index of the Standards or in any of the relevant standards (e.g., 1.6, 1.7, 3.2, 3.6, 14.8, 14.11), issues related to sufficiency appear in every section that deals with content-related evidence. Issues relating to contamination, on the other hand, appear in every section that deals with evidentiary bases of inferences. Content-related validity provides an appropriate framework for determining the extent of insufficiency.

Our position is that content-related evidence has the potential to expose insufficiency only if the construct is poorly specified. If the construct is well specified, then insufficiency is not possible in the absence of egregious oversight. Therefore, we recommend that to ensure sufficiency, researchers spend additional effort in better explaining the conceptual foundations of their measure. From our experience, many scale development efforts jump straight into writing items, with little attention paid to a careful explication of the construct that those items are supposedly measuring. Engaging in more "up-front" thinking about the target construct will help ensure sufficiency. In addition, it is useful to think of sufficiency in terms of a battery of tests. If one particular test is insufficient in capturing the range of constructs needed to perform a particular job well, then other tests could be used to supplement that single measure.

For test users, it is very important to compare the critical KSAOs derived from a professionally conducted job analysis to the content of the test items that you are considering using. In terms of understanding whether the test you are considering is reasonably sufficient, the quality of the job analysis is crucial. Many test publishers will help you conduct a job analysis and then use those results to link to tests that represent constructs identified in the job analysis.

Contamination

As noted in the introduction, measurement contamination implies that a particular measure is influenced by unwanted sources of variance, different from the construct of interest. Confirmatory factor analytic (CFA) frameworks are helpful in understanding the complex multidimensional nature of contamination by isolating different sources of variance. As will be noted throughout this section, concern for contamination is motivated not only by the psychometric goal of creating a "pure" measure but also by a desire to minimize sources of irrelevant variance that covary with membership in demographic subgroups that are accorded special protection under U.S. employment law. Therefore, all I-O psychologists should be concerned with the contamination of their instruments. Given the complexity of the analyses that can be used to quantify contamination, we devote more space on this topic than reliability and sufficiency.

Contamination implies that a particular measure is influenced by sources of variance other than the construct of interest. Although these sources of irrelevant variance could arise from methods effects, response styles, or irrelevant constructs, within a selection context the largest concern centers around contamination of sources of irrelevant variance that are due to membership in legally protected classes. U.S. employment law prohibits making employment decisions on the basis of group membership in terms of race, color, religion, gender, nationality (Civil Rights Act of 1964), age (Age Discrimination in Employment Act of 1967), and disability (American with Disabilities Act of 1990), whether or not this is the employer's intent. In this sense, the use of test scores that vary on the basis of race or another protected characteristic can create *adverse impact* in the legal sense, increasing an employer's chances of involvement in litigation (Williamson, Campion, Malos, Roehling, & Campion, 1997). Aside from legal concerns, ignoring measurement differences among subpopulations can negatively impact decisions based on organizational research (Drasgow, 1984, 1987) and diversification efforts (Offerman & Gowing, 1993), and can cause negative applicant reactions to the assessment (Gilliland & Steiner, 1999). Thus, it is imperative for researchers in organizations to examine whether the adequacy of an assessment method is similar across groups that may be legally, practically, or theoretically important. In addition to legal and practical concerns, the consideration of potential differences across subpopulations has scientific value. For example,

hypotheses about cultural differences can be tested by examining the ways in which people from different cultures respond differently to certain items.

How to Use Validity Evidence to Help Choose a Test

The first point to remember is that the validity of the test you are using depends on the particular purpose for which it is being used. Test publishers who claim that their test is valid without specifying the context should not be treated seriously. A knowledge test that has been shown to predict success for actuarial scientists will likely not be valid for predicting whether a comedian would generate consistent applause and attendance. A reputable test publisher should be able to provide validation evidence from previous studies to help another test user decide whether a particular test is likely to be valid for a particular usage. Although in earlier times, I-O psychologists were concerned about situational specificity, which stated that validities might vary significantly across situations (with an ambiguous understanding of what situational factors mattered), with the popularization of meta-analyses and validity generalization, these concerns have been lessened. Strong meta-analytic research has shown that cognitive ability tests have validity for nearly all occupations, though the validity is higher for more complex jobs (e.g., Schmidt & Hunter, 1998). In addition, evidence shows that personality traits such as conscientiousness have validity across a wide variety of occupations (e.g., Barrick & Mount, 1991).

It is not enough, however, just to rely on a general statement of validity generalization such as "Our test of conscientiousness is valid because meta-analyses have shown such tests to be valid across a wide range of occupations." First, just because a test is asserted to measure a particular construct does not mean that it actually does. *Validity by assertion* is not a technique recognized by I-O psychologists and respected in courts of law! A reputable test publisher will have correlated its particular test with other tests that measure similar constructs, demonstrating convergent validity. Second, it is important to assess whether the range of occupations for which the test has been used is similar to the ones for which you will be using the test. Finally, it is important to assess the similarity of the situations for which the test is being used. Has it only been validated for personal development and self-insight or has it been validated for high-stakes decision making?

In terms of adverse impact and measurement invariance, reputable test developers should make mean differences for sex and race available for review as well as differential validity statistics. These statistics allow an organization to determine whether a particular test might impact the diversity of hiring decisions. In some cases, organizations may still choose a test with adverse impact because it may have the highest validity compared to other alternatives.

OTHER CONSIDERATIONS FOR CHOOSING A TEST

Although reliability and validity should be the foundation for decisions about whether to choose a particular test or not, we realize that consumers of tests care about many other factors. In this section, we briefly review some more practical issues, such as test security, efficiency, access to norms, and delivery.

Test Security

Test consumers want to make sure that the scores that assessments yield are representative of the KSAOs of the person who is taking the test possesses. Without test security, it may be difficult to know if the score for a person truly represents that individual's construct score or not. Candidates might have other candidates take the test, or may have access to the test beforehand, or may be able to access content from the exam via ancillary sources. Test security may be a

primary consideration for some test users and may be of minor importance to others. For tests that are used as the primary basis for making an important decision, test security may be a prime concern. For tests that are relatively low stakes, such as initial screening tests or tests used for developmental purposes, test security may not be a concern at all.

For those who need high test security, there are several options. One of the best solutions may be to use computerized adaptive testing (CAT), which relies on item response theory (IRT) to match individual items that are most appropriate to an individual. Well-designed CATs are high in security because the test that each test taker receives differs from that of other test takers. Creation of a well-designed CAT, however, is a serious endeavor that requires a large number of items that are pre-calibrated. To develop a CAT may be beyond the capabilities of most organizations, though it may be possible for smaller organizations to use the same CAT as others or use a test from a test publisher that has the resources and client base to develop an effective CAT. If a company is unwilling to invest in a CAT, one simple solution that increases security, though not as much as a CAT, would be to randomize the order of items throughout a test. This can make it more difficult for respondents to remember a string of answers, though the challenge is still not insurmountable. Besides modifying the order of items, general advice to keep materials as secure as possible seems warranted.

Efficiency

Another major consideration for choosing a test is efficiency. How much time does the test take to complete? Unfortunately, there tends to be a tradeoff in terms of efficiency and reliability. Increasing the number of items in an assessment (assuming they are good items) increases the measurement precision of the particular test, although it increases testing time. Clearly, in some situations testing time is a premium. For example, a company may wish to bring an applicant in and have him/her complete some psychological tests, while providing the individual with a recruiting tour of the corporate facilities. In addition, for some jobs, applicants may not be willing to complete a test if it takes too long. Target stores have a computerized kiosk where people can complete the assessment before or after shopping. Clearly, if the assessment took two hours, many good applicants would give up and carry on with their other activities.

CAT is a great solution to the tradeoff between efficiency and measurement precision because good CATs eliminate ineffective items. If you are a mathematical genius, it is a waste of time to ask you basic algebra items. And consequently, if you are less adept at mathematics, asking you to solve two simultaneous unknown equations is futile. Without CAT, test users need to determine how long an assessment takes for most applicants and also determine if there are shorter and longer forms of a test so there can be some flexibility. In our field, there seems to be a trend to make sure all scales are as short as possible. The danger of administering three-item personality tests is that the reliability tends to be so low as to preclude making decisions about individuals based on those test scores.

Norms

Another consideration for choosing a particular test may be the availability of relevant norms. It might be extremely useful to compare individual scores to norms within a particular industry or country or across the general population. In fact, some tests are more useful because the organization responsible for the tests has collected norms across a variety of organizations, demographic groups, cultures, and industries. These norms can be extremely useful in interpreting individual scores, especially if you have a small number of people who will be completing your test. Of course, with norms it is important to determine whether they are relevant to your population. Knowing that your eighth grader is at the second percentile of all individuals who have completed the GRE Psychology Subject test is not a good indication of her/his potential for success in a psychology doctoral program.

CONCLUSIONS

The outlandish claims made by test promoters often make it difficult to determine whether a particular test will work as intended. Fortunately, the science of test development and evaluation has a long history and can be used to see through some of the claims used to advertise tests. We hope that this chapter's review of some of the fundamentals of reliability and validity provides a useful background for test users to make informed decisions.

REFERENCES

Age Discrimination in Employment Act of 1967, 29 U.S.C. § 621 (1967).

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Americans with Disabilities Act of 1990, 42 U.S.C. § 12101 (1990).

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

Civil Rights Act of 1964, 42 U.S.C. § 253 (1964)

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, *80*, 565–579.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin*, *95*, 134–135.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*, 19–29.

Gilliland, S. W., & Steiner, D. D. (1999). Applicant reactions. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 69–82). London: Sage Publications.

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology*,*83*, 960–968.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183–1192.

Offerman, L. R., & Gowing, M. K. (1993). Personnel selection in the future: The impact of changing demographics and the nature of work. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 385–417). San Francisco: Jossey-Bass.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85*, 612–624.

Roznowski, M., & Hanisch, K. A. (1990). Building systematic heterogeneity into work attitudes and behavior measures. *Journal of Vocational Behavior*, *36*, 361–375.

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.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Society for Industrial-Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th Ed.). Bowling Green, OH: Society for Industrial-Organizational Psychology.

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist, 50*, 912–927.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored Internet testing in employment settings. *Personnel Psychology*, *59*(1), 189–225.

Williamson, L. G., Campion, J. E., Malos, S. B., Roehling, M. V., & Campion, M. A. (1997). Employment interview on trial: Linking interview structure with litigation outcomes. *Journal of Applied Psychology*, *82*, 900–912.

Woehr, D. J., Sheehan, M. K., & Bennett Jr, W. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology*, *90*, 592–600.