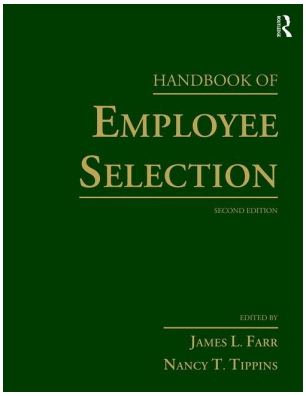

Handbook of Employee Selection

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

Test Administration and the Use of Test Scores

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-8
Jeff W. Johnson, Frederick L. Oswald
Published online on: 22 Mar 2017

8

Test Administration and The Use of Test Scores

JEFF W. JOHNSON AND FREDERICK L. OSWALD

USE OF TEST SCORES

Before you were born, you may have already taken a test (a prenatal test). You likely have been tested repeatedly since then, with the results of many of those tests having meaningful consequences in your life. This chapter focuses on the use of tests and test scores relevant to employment selection settings. Personnel selection is only part of a system of practices that, together, contribute toward meeting a variety of organizational goals (e.g., improved job performance, more effective teamwork, improved learning outcomes, higher motivation, reduced turnover). In other words, an organizational problem with any complexity to it is usually not purely a "selection problem," meaning that it is most effectively evaluated and addressed through an integrated approach to practice that involves a broad set of professional, strategic, and technical skills (Huselid, Jackson, & Schuler, 1997). Just as selection systems do not exist in isolation, neither do selection test scores. Decisions about the type of test scores to collect might be influenced by a number of considerations, such as the type of training that will or will not be provided to job applicants once they are selected; the knowledge, skills, abilities, and other characteristics (KSAOs) of applicants who are the focus of the organization's recruiting strategies; or the time and budget available for test administration. Because of the broad context of selection systems and testing, selection researchers and practitioners can meaningfully improve their skills and their work by remaining connected with the literature in training, motivation, leadership, teamwork, technology, and other relevant substantive areas outside of their usual niche.

DECISIONS TO MAKE BEFORE COLLECTING TEST SCORES

Although a chapter on the use of test scores implies the scores have already been collected, most of the decisions about how test scores will be used must be made at the outset of the testing program. This section presents several issues that must be addressed prior to data collection when determining how test scores in selection will be used.

First, for what purpose will the test scores be used? We focus on selection issues in this chapter, but testing is useful for several organizational purposes. General and common uses of test scores in organizations are (a) to select job applicants for employment, (b) to identify developmental or training needs, (c) to award licensure or certification of professional knowledge and

training, and (d) to promote current employees or determine who will be put on a fast track for later promotion.

Second, what are the major goals of the selection program? Examples of goals are (a) increasing the level of task-specific or contextual job performance among employees; (b) maximizing the number of employees who meet a minimum level of skill proficiency; (c) improving employee commitment and retention; (d) increasing the diversity of skill sets in the organization; and (e) minimizing counterproductive behaviors such as theft, absenteeism, unsafe work behavior, or drug abuse. Taxonomies of job performance have helped sharpen the models, definitions, and metrics that underlie the organizational objectives of selection, but they also have opened our eyes to the fact that multiple objectives are usually of interest, making the operationalization of selection goals complicated. The likely inherent tradeoffs between satisfying certain selection goals mean that some sort of decision must be made about the relative importance of each goal to the organization. If major organizational goals are overlooked in designing a personnel selection system, then decisions made during the selection process could contradict or compete with decisions made with respect to other practices and policies of the organization.

Third, what characteristics do you want to measure in selection? Measures of characteristics that are important for the current job, for future job requirements, or for person-organization fit are often the foundation for a selection program. When the test or test battery measures characteristics that are closely related to job performance, such as technical knowledge and skill, the predictive accuracy of selection tests are usually most favorable, as is face validity (acceptability) of the tests in the eyes of job applicants. Job knowledge tests and work samples might be more costly, or require more company-specific tailoring, than measures of more indirect determinants of job performance, such as ability and personality. Furthermore, it may not be possible or appropriate to test for specific job knowledge and skill in some situations, such as in entry-level positions or positions in which applicants are expected to receive training to remedy any knowledge and skill deficiencies.

Fourth, what is the volume of testing necessary in the selection context? Tests can be used in a wide variety of ways, such as (a) selecting a single individual from a small number of applicants into a specific position (such as CEO); (b) selecting large volumes of applicants into entry-level positions, such as in the fast-food service industry; (c) small businesses having to select from limited pools of applicants who happen to be applying to multiple jobs simultaneously (Scullen & Meyer, 2012); or (d) classifying a large pool of individuals into a wide range of jobs, as is done in the military. Testing in small versus large organizations can influence the nature of a selection process in fundamental ways. For example, large organizations tend to have the resources that would allow them to customize test content and scoring, and their own local data might be extensive enough to allow for stable psychometric and validity analyses, as well as the use of statistical methods for weighting and combining scores in relatively complex ways. By contrast, smaller organizations may be limited to buying off-the-shelf tests that are scored and interpreted by outside vendors using a relevant but broader set of norms, and support for the use of those tests may require relying more heavily on multiple sources of external information (e.g., job analysis, transporting validity from other situations, and meta-analyses; see McPhail, 2007).

Finally, the mode of administration can influence how test scores are used. Will the test be administered via paper and pencil, computer, role play, work sample, or interactive voice response (IVR)? Will the test be proctored or unproctored, one-on-one or group administration? The shift to Internet or computer-based testing opens the door to a much wider range of test formats (e.g., video, interactive, adaptive) and new constructs that can be measured more reliably (e.g., interpersonal skills, ability to speak a foreign language). Tests using these innovative formats must hold up to the same high psychometric standards as those for their more traditional counterparts, with comparability being a concern when multiple formats of a measure are used in a selection setting (e.g,. web vs. paper-and-pencil tests; tests translated into multiple languages).

The decisions we have briefly covered influence later decisions in the testing process. The remainder of this chapter discusses (a) collection of test scores, (b) computation of test scores, and (c) selection decisions on the basis of test scores.

COLLECTION OF TEST SCORES

Several decisions pertaining to the collection of test scores can influence data quality as well as applicant satisfaction with and legal defensibility of the process. In this section, we discuss issues associated with test security, mode of administration, testing time, and retesting.

Maintaining Test Security

In high-stakes testing settings, organizations and test vendors have a keen interest in test security for two primary reasons. First, test owners want to protect their proprietary rights to the content, format, and unique aspects of the administration and scoring of the test items. Second, test users want to maintain the test's fairness and validity by preventing the spread of test-specific information that would allow for cheating (e.g., individuals posting information about a test or the testing procedure on the Internet or passing such information on to their friends; applicants taking pictures of test questions or memorizing them so they unfairly benefit upon a retest). Organizations also have an ethical responsibility to communicate and maintain the privacy and security of individuals' test scores in high-stakes testing situations, and the data are likely to be of higher quality as a result (e.g., informing applicants about the confidentiality of item responses and test scores may reduce test taker anxiety).

Different circumstances make security breaches more likely. For example, paper-and-pencil administration requires hard copies of tests that are easier to steal (computer-administered tests are more difficult to appropriate if proper safeguards are in place). A larger number of examinees means greater likelihood that unscrupulous examinees will be among those tested, greater demand for obtaining test information among applicants, and larger testing sessions that are more difficult to proctor. Using older and/or commercial tests provides more opportunity for the test to be compromised, especially given a limited number of test forms or a small item pool. Finally, unproctored Internet tests taken off-site are now commonplace, which means relying more extensively on the good faith of the examinee that test content will not be taken and shared with others.

Alternate Forms

A common and effective strategy for enhancing test security is creating alternate forms of the test, thus increasing test security while maintaining comparable scores across forms. Given that a test reflects a representative sample of content from the construct domain of interest, it should be possible to develop alternate measures of the same construct that exhibit similar psychometric properties in terms of reliability and validity. High correlations between scores across test forms provides evidence that scores from a particular test are reflective of an applicant's standing on an underlying construct rather than reflective of an applicant's understanding of content unique to a particular test form. The following subsections review (a) creating alternate forms, (b) equating alternate forms, and (c) developing dynamically administered tests.

Creating Alternate Forms Creating alternate forms for some constructs may be a relatively simple task. For example, when creating alternate forms that test for the ability to multiply two 2-digit numbers, substituting different numbers for the original numbers will suffice. When constructs are defined with greater conceptual breadth, however, it is very important that the constructs to be tested are well defined and theoretically driven. Carroll's (1993) hierarchical taxonomy of human abilities would serve as a good reference in the cognitive domain, and the Big Five has proven useful in the personality domain for generating test content (e.g., the International Personality Item Pool, or IPIP, at www.ipip.ori.org; Goldberg et al., 2006). Alternate test forms should be developed so that psychometric characteristics across forms have similarly

high reliability and patterns of criterion-related validity, which can be accomplished by sampling items representatively from a well-defined construct domain. A good approach to creating alternate forms of an ability test is similar to the approach for developing measures in general (DeVellis, 2016). Given that a pool of items will be winnowed down based on the quality of the item content coupled with psychometric characteristics, a rule of thumb is to write about three times as many items as will reliably measure the construct on one form. After developing items, reviewing content, making appropriate revisions, and administering the test in a very small sample of test takers, test developers then collect data from a larger sample for all items in a pilot test (say, $N = 300$) and assign items with similar content and psychometric properties (e.g., proportion correct, corrected item-total r) to alternate forms. Experience tells us that about one-third of the items will drop out because of inadequate psychometric characteristics. Next, ensure that the test forms have similar internal consistency reliabilities and that the correlation between forms is high—at least $r = .90$, after correcting for unreliability using alphas from each form (which should be relatively high for unidimensional constructs). For speeded tests (e.g., some ability tests) and for tests with extremely heterogeneous content (e.g., SJTs), alpha reliability and item-total correlations are generally not appropriate reliability measures (Ployhart & MacKenzie, 2011). Other approaches are to be used in these cases, such as alternate forms and test-retest reliability (Catano, Brochu, & Lamerson, 2012). Finally, when possible, determine whether criterion-related validities are similar across forms when predicting outcomes. Items can be moved across alternate forms to improve the comparability of the forms in terms of reliability, validity, adverse impact, and other factors.

Creating alternate forms for job knowledge tests can follow the same procedure, but it is usually more difficult because of the specificity of knowledge items (e.g., there may not be an alternative item when an examinee needs to know the location of the emergency switch at a nuclear power plant). Also, test developers often lack familiarity with the content area, particularly when job knowledge is highly technical. A good strategy is to have subject matter experts (SMEs; e.g., trainers, supervisors, incumbents) write test items and to instruct them to write an "item buddy" for each item they write. The item buddy would be similar to the original item in terms of content and difficulty, but different enough that knowing the answer to one does not easily give away the answer to the other.

Developing alternate forms for situational judgment tests (SJTs) is a challenge because SJT content can be very wide-ranging in terms of content and constructs assessed. Lievens and Sackett (2007) explored three methods for creating alternate forms for SJTs: (1) assigning items randomly to forms, (2) creating forms with similar situations, and (3) creating forms with similar situations and similar item responses. The latter two methods did show higher test-retest correlations, indicating that random assignment of SJT items may not be a sound approach to developing alternate forms. Oswald, Friede, Schmitt, Kim, and Ramsay (2005) developed multiple parallel forms of an SJT, where each form sampled content across 12 broad dimensions of college student performance (e.g., continuous learning, leadership, ethics). The authors winnowed 10,000 computer-generated forms down to 144 tests with scores having similar means and standard deviations (SDs), high estimated alpha reliability, high estimated validity, and low item overlap. Thus, all test forms were as similar as could be accomplished feasibly in terms of the desired practical qualities of the SJT.

In the personality domain, it is possible to create alternate forms using the same domain-sampling procedure as for ability tests, because there are many potential items to measure personality constructs. Alternate forms may not be necessary, however, because the most desirable answer in a personality test is usually not difficult to determine (Viswesvaran & Ones, 1999). Therefore, there is little to gain in terms of preventing cheating by creating alternate forms of a personality test. A common practice in this case is simply to randomize the presentation order of personality items when creating alternate forms.

In general, the importance of having alternate forms corresponds to the extent to which a single correct answer can be determined for the test questions. For example, biodata tests ask the candidate about experiences and attitudes that are often linked to performance through empirical keying. If a candidate is responding honestly to a biodata or personality test, there is no single "correct" answer because the candidate is just providing a self-description. Thus, alternate

forms are less important for personality and biodata tests; more important for SJTs, interviews, and work simulations; and most important for knowledge or ability tests.

A recent trend is the development of many forms (e.g., 10 instead of 2) in an attempt to minimize cheating and keep a testing program operating if one of the test forms is compromised. Oswald et al. (2005) extended a method proposed by Gibson and Weiner (1998) for creating many test forms on the basis of the statistics from a pool of items that may not have appeared on the same test form or have been given to the same sample. This method of generating parallel forms potentially minimizes the exposure of any single test item; in fact, item exposure, item testing time, item-level validity, and other item characteristics can be built into the mathematical constraints that drive the procedure for generating appropriate alternate test forms (see the linear programming models in van der Linden, 2005). A more common strategy for creating parallel forms is to create three or four unique alternate forms, then create additional forms by changing the item order and/or by taking some items from each of the original forms. Note, however, that mixing items from two or more unique forms to create a new form means that some parts of the unique forms are compromised if the new form were to be stolen.

Equating Alternate Forms When alternate forms are used, it is necessary to equate test scores across forms so that a given score on one form is as psychometrically equivalent to the same score on the other form as possible. Although equating can introduce some additional error into the selection process as opposed to using the same measure across all applicants, careful attention to the process of test development and the establishment of equivalent forms reduces such error. When samples are randomly equivalent, and common anchor items across test forms have similar parameters across samples, several different equating methods based on item response theory (IRT) yield similarly good results (Kim, Choi, Lee, & Um, 2008). In cases in which sample sizes are smaller (less than $N = 500$ per item) or IRT assumptions are untenable, it is necessary to rely on other equating methods. Two other kinds of equating methods are linear equating and equipercentile equating. In linear equating, individual scores across two tests are considered to be equated if they correspond to the same number of standard deviation units from the mean (i.e., the same z scores). Because linear equating is entirely analytical and does not require data at each point in the test score range, it offers the advantages of allowing a mapping of scores from one version to the other throughout the entire range of scores and requires smaller sample sizes than IRT methods. A disadvantage of linear equating is that it requires the assumption that any differences in the shapes of the raw-score distributions for each form are trivial.

In equipercentile equating, scores on two tests are considered to be equated if they correspond to the same percentile rank (for details, see Livingston, 2004, pp. 17–23). A problem with equipercentile equating is that it requires very large sample sizes to precisely equate the entire range of scores on each test. In this method, large errors of estimation are likely in score ranges where data are scant or erratic, so there must be many observations for each possible score on each form (Petersen, Kolen, & Hoover, 1989). Methods are available that smooth the empirical distribution of the data, allowing for more reasonable equipercentile equating in ranges of the scale with less data, assuming the smoothed distribution is the correct one underlying the data (Dorans, Pommerich, & Holland, 2007). If the only score that needs to be equated is a cut score and only a pass-fail decision is communicated to applicants, then we recommend equating the cut score on the basis of equipercentile equating because that will lead to the same pass rate within each form. Whether simpler methods are as useful as IRT-based approaches is an empirical question, but given that very consistent empirical and practical outcomes between IRT and classical test theory have often been identified (Fan, 1998; MacDonald & Paunonen, 2002), we suspect that different methods may often yield similar results.

Dynamically Administered Tests Another way of enhancing test security without having to develop and equate multiple forms of the test is by creating a large item pool and selecting items from that pool in real time as the candidate is completing the test, such as through "linear on the fly" (LOFT) or adaptive testing. Using a LOFT procedure, items are selected dynamically from a large item pool such that different, but equivalent, tests are randomly administered, thus

providing a customized assessment for each applicant. As a result, it is difficult to compromise test security by copying the items and disseminating them to others.

Computer adaptive tests (CATs) also enhance test security because they provide different items to test takers depending on their previous responses, essentially creating a unique form for each applicant. For example, an applicant who answers an ability test item incorrectly will be given an easier item next, whereas an individual who answers that item correctly will be given a more difficult item next. In an adaptive personality test, applicants are presented with item pairs, with each item representing a different level of the target trait. An applicant who chooses the statement at a lower trait level would be given a subsequent item pair that is lower on the trait continuum in an attempt to refine trait-level estimation for that applicant. Even for unidimensional measures, developing ability CATs requires very large sample sizes (at least $N = 500–1,000$) and very large item pools. If this investment can be made, however, the advantage of CAT is that fewer items need to be administered to each individual, reducing test fatigue and testing time while maintaining high reliability across levels of the construct to be measured. Although adaptive testing based on large item pools improves overall test security, it is certainly possible for a savvy test taker trying to see many items to purposely answer certain items incorrectly to ensure that additional items are presented. Other potential disadvantages of CATs are that test takers are not allowed to review their previous answers to correct them, and the test scores resulting from a CAT may be more sensitive to initial responses than to subsequent ones (Chang & Ying, 2008). Investment in CAT development in the private sector is still relatively new, so the conditions under which the cost-benefit of CAT is superior to that of traditional test formats largely remains to be seen.

Mode of Administration

Test administration mode generally refers to (a) paper-and-pencil versus computer administration, and (b) proctored versus unproctored administration. In general, we recommend using computer administration and computer scoring to enhance test security where possible. Although we have noted how computer-based testing does not resolve all security issues, the elimination of paper forms that are more easily stolen is a clear advantage. To maintain the security advantage of computer-administered tests, however, strict physical and information technology (IT) security measures must be established and enforced on a continuous basis to protect against unauthorized access to the testing software. Reputable vendors that specialize in online testing will have extensive security procedures for protecting their intellectual property. We do not recommend that organizations conduct large-scale testing using their own computer systems, unless extensive and up-to-date security measures are in place, and the entire testing system has been thoroughly tested by experienced testing and IT professionals for this purpose.

The International Test Commission (ITC; 2006) has published a set of guidelines for best practice in delivering computer-based testing, especially through the Internet. The guidelines address four primary issues: (1) ensuring that the hardware and software technology at both the server and client side are appropriate for the use of the test; (2) ensuring the quality of test materials and the testing process; (3) controlling the way tests are delivered and who is completing the tests; and (4) ensuring test security, data protection, privacy, and confidentiality. The guidelines are designed to advise test developers, publishers, and users.

One aspect of computer-based testing specifically addressed by the ITC guidelines is unproctored Internet testing (UIT). UIT is increasingly common in selection practice, and it presents unique problems for maintaining test security (see Tippins, 2009). Several things can be done to help minimize test exposure and motivation to cheat in this situation (e.g., Bartram, 2009; Burke, 2009; ITC, 2006; Tippins et al., 2006). First, the system should allow applicants to take the test only once; returning applicants are not allowed to retake the test without the knowledge and approval of the hiring organization. Second, an applicant tracking system should be used that collects relevant identification information and links it to all selection data collected on the applicant. Third, applicants should be encouraged to be honest in the information they provide to ensure a good fit to the organization; they also should be warned about the consequences of

cheating under UIT and how their identity and their answers may be verified. Fourth, each test administration should be unique, through randomizing the item order or selecting items adaptively from large item banks.

Finally, UIT might be used for initial testing or to screen out candidates who are highly unlikely to be suitable for the job (Nye, Do, Drasgow, & Fine, 2008). To verify those scores obtained in the unproctored environment, proctored adaptive tests (Makransky & Glas, 2011) or proctored subtests of the larger unproctored test (Segall, 2001) can still keep testing time short. To date there is little evidence of cheating or inconsistency between unproctored test scores when there is a proctored follow-up test (Kantrowitz & Dainis, 2014). That said, ethical issues should be considered when a follow-up test is used for score confirmation (Bartram, 2009; ITC, 2006). For instance, failure to confirm an unproctored test score is not definitive proof of cheating; it may merely suggest some necessary additional follow-up questioning by hiring managers.

Ultimately, organizations using UIT must accept that the test items are in the public domain and will be exposed to anyone who really wants access, with potential implications for compromising the reliability of test scores and thus the integrity of a selection system. Burke (2009) presented some precautions that can be adopted to determine the extent to which test security has been breached. These precautions include searching the Internet to find sites that inappropriately offer access to test content. When these sites are found, most of them can be taken down by informing the site operator or Internet provider of activity that goes against their stated policies. Another approach to determining the extent to which test content may be compromised is to apply data forensic algorithms that search for evidence of aberrant scores or scores that are highly unlikely under honest test-taking conditions (e.g., fast response times associated with high correct answer rates; matches in correct and incorrect answer profiles, suggesting test taker collusion; long response latencies associated with few correct answers could suggest item harvesting). Ironically, the concern about cheating and test security associated with UIT has led to the development of technologies and procedures that could make UIT more secure than traditional proctored assessment (Bartram, 2009).

Testing Time

The amount of time available for testing is often a practical constraint that influences decisions about the types of tests to be administered, the mode of administration, the number of tests included in a test battery, the number of items in a test, and the number of stages in the selection system. For example, if financial or other considerations place a strict time limit of one hour on test administration, that limits the number of different constructs that can be assessed, the types of constructs that can be assessed (e.g., reading comprehension takes longer than perceptual speed and accuracy), and the testing method (e.g., SJTs generally take longer than biodata inventories). Our general advice when presented with such constraints is to go for depth over breadth. Because testing time (and the patience of test takers) is finite, organizations should maintain their focus in effectively measuring a relatively small handful of key constructs relevant for selection purposes, fully recognizing that not all constructs of potential relevance can be measured. Job analysis should be the essential guide for determining which selection-relevant constructs or characteristics are most important and feasible to measure. Another factor to consider when determining what to include in a test battery is potential adverse impact. When selection tests measure constructs with large subgroup differences, it can be very difficult to neutralize this in a test battery by (a) replacing it with another measure measuring the same construct with equivalent reliability and validity but lower subgroup mean differences (because such measures are very hard to locate or develop) and/or (b) including measures of other constructs that have lower or no subgroup differences (because they do not have as strong of a mathematical effect as one may think, and because this adds breadth and therefore more testing time). Nonetheless, all attempts should be made to balance the goals of validity, reliability, and fairness of the selection battery.

One way to reduce the size of a test battery is to remove tests that are highly correlated with other tests in the battery and do not provide incremental validity. Such a reduced test battery can often maintain its worth in terms of reliability and validity for its intended purposes (Donnellan,

Oswald, Baird, & Lucas, 2006; Stanton, Sinar, Balzer, & Smith, 2002). As mentioned, many organizations cut down on in-house testing time by using unproctored web administration of predictors (Tippins et al., 2006), inviting those meeting a minimum score to the proctored testing session, where more tests are administered. If the unproctored test scores can be linked to the proctored test scores via the applicant name or identification number, then the number of constructs assessed can be increased without increasing proctored testing time.

Power tests are defined as those tests for which speed is not relevant to the measurement of the construct, such as for many cognitive ability and achievement tests. Thus, test takers should be given adequate time to complete the entire test. Because unlimited time is not available for test administration, however, a rule of thumb for allocating a minimum amount of testing time to power tests is the time it takes for 90% of the examinees to complete 90% of the items. Unlike power tests, *speeded tests* require test takers to perform quickly on simpler tasks within the time allotted. Speeded tests must be long enough and the time limit short enough that virtually no one is able to finish. The test score is then scored as the number of correct responses minus a correction for guessing (Cronbach, 1990).

Issues surrounding equity in testing could meaningfully impact decisions on the amount of testing time allocated. Consider job applicants who are nonnative speakers of the language in which the test is written. If it is safe to assume that language skills are not relevant with respect to the construct being assessed by the test, then one should consider providing additional testing time to nonnative speakers so that the influence of unfamiliarity with the language on test scores is greatly diminished or removed. Alternatively, such tests could be redesigned so that written language or other factors irrelevant to the construct are minimized, such as administering a video-based form of a test instead of a traditional written form (Chan & Schmitt, 1997; Weekley & Jones, 1997).

Consideration of individuals covered by the Americans with Disabilities Act (ADA; 1990) is an organizational and legal imperative. If the test is not speeded, applicants with disabilities preventing them from completing the test in the normally allotted time should be given extra time to complete the test. Other reasonable accommodations would include providing larger-font test materials and providing assistance with the answer sheet. See Campbell and Reilly (2000) for an excellent discussion of ADA accommodations in testing.

Alternate Test Formats

The need to create alternate test formats may arise when seeking to test different subgroups of individuals in a comparable manner. Not only might disabled job applicants require a reasonable accommodation, but testing applicants from different countries might also require that the test be translated accurately into several languages. Although in some cases it seems reasonable to assume that differences in test format are merely cosmetic and have no bearing on construct measurement (e.g., a static personality inventory administered on computer vs. administered on paper), research suggests that the psychometric characteristics of tests may differ across formats (Meade, Michels, & Lautenschlager, 2007). Therefore, in most cases it is necessary to test statistically whether format differences lead to score differences that are irrelevant to the constructs being measured.

Numerous studies have found empirical support for the measurement invariance/equivalence of psychometric properties across formats (e.g., for web vs. paper-and-pencil format equivalence, see De Beuckelaer & Lievens, 2009; Naus, Philipp, & Samsi, 2009; Ployhart, Weekley, & Holtz, 2003). Ideally, scores from different test formats would exhibit strict measurement invariance (e.g., similar patterns and magnitudes of factor loadings, similar error variance estimates; Vandenberg & Lance, 2000), but much more often, the data support either scalar invariance (equal loadings and intercepts) or metric invariance (equal loadings only). Furthermore, when some items do not psychometrically function in the same way across subgroups, then partial invariance is said to exist. Items contributing to partial invariance may lead one either to delete or revise them, or to allow them to have unique estimates across subgroups. Recent simulation and empirical work suggests that partial invariance may only have a slight impact on selection

outcomes in many practical situations (Millsap & Kwok, 2004; Stark, Chernyshenko, & Drasgow, 2004; for a thorough review of tests of measurement invariance, see Schmitt & Kuljanin, 2008).

Large sample sizes are desirable in conducting appropriate tests for measurement invariance (e.g., $N = 400$ for supporting metric invariance; Meade & Bauer, 2007); otherwise, it may be necessary to rely on a strong rational basis for the equivalence of test formats (e.g., providing evidence that the formats of the different tests do not influence the construct of interest and should be unrelated to differences in the samples tested). Note that even when measures are found to be psychometrically nonequivalent across formats, this may not stop an organization from proceeding with both formats of a test (e.g., when moving toward an Internet-only format, but using paper forms while in transition), though it should be done with the understanding that nonequivalence prevents the direct comparability of the measures.

Retesting

According to professional guidelines, allowing candidates to retest at a later date is a best practice (Society for Industrial and Organizational Psychology, 2003; U.S. Department of Labor, 1999). Because any single assessment can be influenced by various types of systematic measurement error unrelated to the construct being measured (e.g., illness, extreme anxiety, not meeting the testing prerequisites), it is reasonable to offer the opportunity to retest when it is appropriate and feasible. Retesting is typically not a relevant concern for small-scale testing programs. If a small company is testing to fill a specific position, then the opportunity to retest cannot be offered once the position has been filled. Candidates should be given the chance to retest in a timely manner if job opportunities are continuously available, especially if it involves internal candidates.

On the other hand, how are the psychometric qualities of the test affected by allowing candidates to take it two, three, or four times? Is the purpose of the test defeated if a candidate can retake the test as many times as it takes to finally pass? Score increases from retesting could be partially due to regression to the mean, because low test scorers are more likely to retest, and lower retest scores would tend to increase by chance alone. Score increases could also be due to practice effects on the test-specific content (e.g., memorizing the correct answers) or in picking up on effective test strategies (e.g., learning that picking the longest multiple choice answer is beneficial when in doubt). These undesirable effects need to be considered, if not prevented by not allowing for a retest. On the other hand, retest scores can also reflect desirable effects. Some test takers could be less anxious about the test when they retake it, which could lead to score increases that better reflect their true level of knowledge even if their underlying knowledge upon retest remains the same. Test takers may also consider the types of questions they missed when they first tested and concentrate subsequent study in those areas so that the retest reflects true increases in the construct being measured.

Some recent studies have examined the effect of retesting on scores for different types of tests. In a study of admission exams for medical students, Lievens, Buyse, and Sackett (2005) found standardized mean gains in test scores (after correcting for test-retest reliability) of 0.46 for a cognitive ability test, 0.30 for a knowledge test, and 0.40 for a SJT. Retesting was conducted using alternate forms, suggesting that increases were due to increases in the respective constructs being measured and not due to simply memorizing item-specific content. Raymond, Neustel, and Anderson (2007) found standardized mean gains of 0.79 and 0.48 for two different medical certification exams. In both cases, these gains were nearly the same whether the identical test or a parallel test was used. This latter finding was contrary to a recent meta-analysis of retesting effects on cognitive ability tests that found an adjusted overall effect size of 0.46 for identical forms and 0.24 for alternate forms (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007).

Across all studies (combining samples using identical and alternate forms), the Hausknecht et al. (2007) meta-analysis found an effect size of 0.26 (based on adjusted SD units) between

Time 1 and Time 2 and an effect size of 0.20 between Time 2 and Time 3. The portion of these effects that was due to regression to the mean was estimated to be less than 10%. Test coaching, defined as instruction aimed at improving test scores (through learning either test-related skills or test-taking strategies), generally had a large effect. The effect size for individuals who had some form of test coaching was 0.70, as opposed to 0.24 for individuals who did not have any coaching. Another moderator of gains upon retesting was the type of cognitive ability assessed, with tests of quantitative and analytical ability showing larger mean gains (0.30 and 0.32, respectively) than tests of verbal ability (0.19).

Do scores on a retest tend to reflect the individual's standing on a construct better than scores on the initial test? Lievens et al. (2005) examined this question by using a within-person analysis to compare the validity coefficients for original scores and retest scores for those examinees who did not pass the medical school admissions exam, elected to retest with an alternate form, and subsequently were admitted to medical school. When predicting GPA with the knowledge test, Lievens et al. hypothesized and found significantly higher validity coefficients for retest scores (r corrected for range restriction = .37) than for initial scores (corrected r = .23, N = 556). Also as hypothesized, there were no statistically significant differences in validity coefficients for the cognitive ability test or for the SJT.

Note that the time intervals varied in studies examining retesting. There is little research to inform the decision about how long a candidate should have to wait before being allowed to retest, although the length of the time intervals used in practice generally do not appear to influence test score gains when alternate forms are used. Hausknecht et al. (2007) noted that score gains tended to decrease as the time interval increased, but only for identical forms. Raymond et al. (2007) found no effect of time delay on score increases regardless of whether identical or alternate forms were used. Using identical forms, Burke (1997) found different retest gains across components of a cognitive and psychomotor ability selection battery, but the retesting time interval (ranging from 1 to 5 years) did not moderate these gains by any significant amount, suggesting that the retest effect tends to be stronger than any time effect. The appropriate time interval for retesting in employment settings depends on factors such as the availability of alternate forms and the type of test, but also administrative concerns such as the cost of testing and the need to fill positions. A minimum of 6 months between administrations has been a common rule of thumb for large-scale ability or knowledge testing programs, and 30–60 days is more typical for certain types of skills tests (e.g., typing, software proficiency).

Although retesting effects are usually of interest primarily for cognitive ability or knowledge tests, the effect of retesting on personality test scores has also been examined. Landers, Sackett, and Tuzinski (2011) observed that the use of extreme responses (all 1s and 5s) increased for internal applicants who chose to retake a personality test after initial failure. This type of faking increased over time, suggesting that applicants were being coached on how to respond to maximize scores. An interactive warning indicating that an extreme response pattern was not consistent with paying careful attention to each item reduced the incidence of extreme response faking. This study suggests that retaking a personality test increases the risk of faking to increase scores, but steps can be taken to mitigate this faking.

Finally, in addition to these retesting findings just summarized, there is also a need for research on the effect of retesting on adverse impact. Specifically, are expected score gains upon retesting equivalent for different gender or racial/ethnic subgroups? Are candidates within different subgroups equally likely to retest when given the opportunity? Questions such as these must be answered before it is possible to speculate on the impact of retesting on adverse impact ratios found in selection practice.

COMPUTATION OF TEST SCORES

After test data have been collected, scores must be computed and transformed into a single score or multiple scores that will be used as the basis for making the selection decision. In this section, we discuss creating predictor composites and reporting test scores.

Creating Predictor Composites

It is common for organizations to use more than one predictor for decision making and to combine the scores on each predictor into a single composite score. For example, standardized scores on a reading comprehension test, a conscientiousness test, and an SJT may be added together to create a single score that describes a job applicant better than each test considered in isolation. This is a compensatory approach, in which high scores on one predictor can compensate for low scores on another predictor. This is contrasted with a noncompensatory or multiple-hurdles approach, in which selected applicants must meet a minimum passing score on each predictor. Issues associated with a multiple-hurdles approach are discussed later in this chapter. Two key decisions must be made when compiling a predictor battery and computing a composite score: (1) what predictors should be included in the battery? and (2) how should each predictor be weighted in computing the composite score?

Choosing Predictors

A common problem faced by personnel selection researchers and practitioners is choosing a set of predictors from a larger set of potential predictors for the purpose of creating a predictor battery. In large organizations, an experimental predictor battery may have been assembled for a validation study, and the choice of which predictors to include is based on the psychometric characteristics of the predictors as measured in that study, and possibly validity information where available. In smaller organizations, the choice may be based on examining published norms or meta-analysis results for a wide range of job-relevant predictors. Usually, there are additional practical constraints, such as test availability, cost of the tests, and required testing time.

When assembling a predictor battery, there is often a tradeoff between maximizing criterion-related validity and minimizing adverse impact against protected groups (e.g., racial/ethnic minorities or females). Creating a composite of several valid predictors is a common strategy for reducing the degree to which a selection procedure produces group differences (Campbell, 1996; Sackett & Ellingson, 1997). The problem is that some of the most valid predictors of performance are cognitive in nature, but those predictors tend to have the largest potential for adverse impact. Therefore, adding a cognitive predictor that increases the validity of the composite will often have the simultaneous effect of increasing adverse impact (Sackett & Ellingson, 1997).

Compounding the problem is the fact that adding a predictor with little adverse impact to a predictor with large adverse impact typically does not reduce the adverse impact of the composite to the extent that would generally be expected (Potosky, Bobko, & Roth, 2005; Sackett & Ellingson, 1997). Sackett and Ellingson (1997) gave an example of two uncorrelated predictors. One predictor had a standardized mean subgroup difference (*d*) of 1.00 and the other had a *d* of 0.00. Most researchers would expect that the two predictors would offset each other, so the *d* of an equally weighted composite of the two predictors would be 0.50. In fact, the *d* of this composite would be 0.71 (the square root of 0.50), and one would have to add two more uncorrelated predictors (three predictors uncorrelated with each other and with a cognitive ability predictor) to achieve a *d* value of 0.50. Potosky et al. (2005) further demonstrated the difficulty in reducing adverse impact with a predictor composite by pointing out that the potential for adverse impact in many predictors has been underestimated because *d* has been computed in range-restricted samples of job incumbents rather than in the full range of job applicant samples. On the other hand, a meta-analysis of cognitive ability and race/ethnic differences found that, although overall Black-White *d* is considered to be around 1.0, the *d* values appear to range roughly between 0.60 for high-complexity jobs and .85 in low-complexity jobs, even after accounting for range restriction within jobs (Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Roth, Switzer, Van Iddekinge, & Oh, 2011).

The mathematical presentation of Sackett and Ellingson (1997) and the meta-analysis of Potosky et al. (2005) both demonstrated that reducing adverse impact by adding predictors

to a composite is not as easy as it seems at first glance. The takeaway message is that reducing adverse impact is not a simple matter of adding a noncognitive predictor or two to a predictor composite that includes a measure of cognitive ability, nor is it to create a test battery that overweights measured job characteristics relative to their actual importance. Researchers and practitioners trying to balance validity against potential adverse impact when creating predictor composites should explore a wider variety of alternative predictors and weighting schemes rather than mechanically relying on offsetting the adverse impact of one predictor with another.

Weighting Predictors

When multiple predictors are used, a decision must be made on how much weight to apply to each predictor when computing a composite score. Two types of weights are considered here: (a) statistical weights and (b) rational weights. Statistical weights are data driven, whereas the researcher specifies rational weights, perhaps with input from SMEs. The most common statistical weights are derived using multiple regression because, in a given sample, regression weights maximize the prediction of the criterion (in the sense of minimizing the sum of squared errors). However, regression weights have numerous limitations that often make alternative weighting schemes more desirable. First, criterion scores are not always available, such as when a content-oriented validation strategy is used. Second, regression weights focus entirely on prediction, so they cannot take into account adverse impact or other practical considerations. Third, regression weights can be difficult to interpret and explain to stakeholders, especially when predictors are correlated. Finally, the question of primary interest is how well regression weights predict in other independent samples (e.g., samples of future job applicants), not in the specific sample in which the weights were derived. Sampling error variance and/or correlated predictors make regression weights unstable and thus prone to inaccuracy in other samples compared with unit weights (i.e., a simple sum of standardized predictors), especially when sample sizes are relatively small (less than about 180; Schmidt, 1971) and predictor intercorrelations are high (i.e., multicollinearity; Green, 1977).

De Corte, Lievens, and Sackett (2007) presented a procedure for weighting predictors in such a way that the tradeoff between selection quality (e.g., validity, average criterion score of those selected) and adverse impact is Pareto-optimized. Pareto optimization means that mean subgroup differences are minimized for a given level of validity (or similarly, that validity is maximized for a given level of mean subgroup differences). The procedure applies optimization methods from the field of operations research and involves nonlinear programming. The authors offer a computer program for application of the procedure. Unfortunately, when the criterion is task performance or other criteria requiring cognitive ability, then the Pareto-optimal composites usually require down-weighting a cognitive ability predictor considerably to reduce mean subgroup differences on the predictor composite by a practically significant amount. This compromises validity significantly in most cases, and thus, the low weights might not be a reasonable reflection of the actual importance of cognitive ability on the job.

Given the limitations of empirical weighting schemes, rational weights are frequently applied. Examples include job analysis ratings of importance (Goldstein, Zedeck, & Schneider, 1993) and expert judgments of predictor importance (Janz, Hellervik, & Gilmore, 1986). Johnson and Carter (2010) found that weighting each predictor by the number of performance dimensions to which it is relevant yielded higher validities than did applying unit weights. Given the multidimensional nature of job performance, the approach of placing greater weight on predictors that are likely to influence a wider range of outcomes is an attractive approach from a conceptual standpoint as well as from a predictive standpoint. There are also other considerations that may influence the weighting scheme, for better or for worse, such as equally weighting cognitive and noncognitive portions of the predictor battery or weighting to please stakeholders (e.g., the CEO thinks the interview should be given more weight than the cognitive test). To the extent that adverse impact is not triggered as a function of weighting, the resulting composite scores

from alternative weights become less concerning from a legal standpoint, so long as the weights are applied similarly across all members of the applicant pool.

An important point to keep in mind when considering alternative weighting schemes is that whatever weights are directly applied to a set of predictors—called *nominal weights*—may not have the desired effect on the final composite scores. Oswald, Putka, and Ock (2015) provide several examples demonstrating how predictor variables are not actually weighted as intended unless they are completely uncorrelated (also see Brannick & Darling, 1991). This is because applying a nominal weight to one predictor also applies an implicit weight to all other correlated predictors being used. As a result, the *effective weight* applied to a given variable is not the nominal weight, but instead reflects a combination of the nominal weight and the implicit weights resulting from that variable's correlation with each of the other variables in the composite (see Guion, 2011, p. 275 ff.). Because many components of a selection battery are likely to be positively correlated to some extent, the composite scores created by weighting predictors will not actually reflect the intended weighting scheme. Statistical methods exist for translating nominal weights into effective weights, and although there is no single correct method for doing so (Brannick & Darling, 1991; Oswald et al., 2015), the attempt at translation allows one to understand better whether each predictor contributes to the composite in the manner that was originally intended.

Before spending time arriving at a predictor weighting scheme and worrying about the extent to which our explicit weights correspond to our effective weights, we should first ask to what extent does differential weighting of predictors influence the overall composite score. Both Koopman (1988) and Ree, Carretta, and Earles (1998) demonstrated that very different sets of weights can lead to highly correlated composite scores—often above .95. Bobko, Roth, and Buster (2007) reviewed the literature on the usefulness of unit weights and concluded that unit weights are highly appropriate under many circumstances, including when adopting a content-oriented validation strategy. Unit weights are the easiest to calculate, the easiest to explain, and the most generalizable to different situations. When applying weights, unless sample sizes are large, the number of predictors is small, and predictor intercorrelations are low, it is probably best to use unit weights to keep the weighting of predictors simple (Bobko et al., 2007; Cohen, 1990).

Going beyond choosing and weighting predictors, three review articles take a broader focus on the frequent tradeoff between validity and reducing adverse impact (Kravitz, 2008; Ployhart & Holtz, 2008; Sackett, Schmitt, Ellingson, & Kabin, 2001). These authors examined a wide variety of organizational strategies, indicating which strategies appear to have promise (e.g., administering tests in a video format to reduce an unnecessary language burden), which strategies have not been as fruitful (e.g., modifying tests based on differential item functioning statistics), and which longer-term strategies lack empirical evidence but may be promising (e.g., engaging the organization in broad community-based efforts to increase visibility and attract more qualified minority applicants).

Reporting Test Scores

After job applicants have been tested, it is customary to communicate to them how well they performed. There are no standards for how much information must be provided, nor the format in which to provide it, so score reporting runs the gamut from a simple pass/fail notification to a detailed report of the number correct on each test in the battery and how the information was combined to create an overall score. We are unaware of any research on applicant reactions to how test scores are reported, apart from reactions to how scoring led to making the selection decision (e.g., top-down selection versus test-score banding; Truxillo & Bauer, 1999, 2000). The type of score information provided should depend on the purpose of the assessment and the nature of the applicant. For example, if an assessment center is used to evaluate internal candidates for promotion, it is probably beneficial to the organization and the candidates to provide extensive developmental feedback on how the candidate did on each exercise. This feedback could then lead to targeted training interventions and, ultimately, performance improvement among employees (see Chapter 19, this volume, for details on providing assessment feedback to applicants). On the other hand, if a standardized test is used to screen out a large number of

external applicants who do not meet a minimum cut score, then no more than a pass/fail decision likely needs to be communicated. If a pool of potential candidates is identified that exceeds a cut score, and the highest-scoring individuals in that pool will be given the first opportunity for job openings, then some information about where the individual stands among other candidates (e.g., a ranking or percentile score) might be appropriate to communicate some idea of how likely a job offer is in the near future.

Raw test scores are often transformed to make them more interpretable to applicants. This is especially important when the selection instrument contains tests that do not have right and wrong answers (e.g., personality tests). To avoid confusion and negative impressions, we recommend not reporting test scores with negative values, which may occur when computing raw scores from items that have negatively scored response options (e.g., many biodata items) or when computing z-scores, where scores below the mean are negative. A common transformation for score reporting is to report T-scores, which standardize scores to have a mean of 50 and standard deviation of 10. This makes scores look like percentiles, because most scores range from 10 to 90, and the scores provide information about how the applicant did on each test but do not explicitly state the number correct. Another linear transformation is to convert the total score to a 100-point scale, and when a cut score is used, the cut score can be set at a value such as 70. This reporting method is easy to understand, because scores are similar to grades in school, and provides applicants with a reasonable idea of where they stand in terms of passing and failing and compared with the maximum score. A downside of both T-scores and this latter conversion is that both may incorrectly imply that the score represents the percentage of items answered correctly. A third alternative is to place scores into categories for the purposes of communication to management or other constituencies (e.g., "excellent," "good," "borderline," "poor"). These coarser categories are sometimes used for selection, although coarsening scores will reduce the validity of a selection measure, as discussed in the next section.

MAKING SELECTION DECISIONS

Once final test scores or composite scores are computed, they must be translated into a final selection decision. There are many ways to arrive at the decision to select, reject, or move to the next phase of the selection process. In this section, we discuss top-down selection, setting cut scores, banding, multiple hurdles, selection to fit a profile, and context-based selection.

Methods of Selection

Given personnel selection data that demonstrate linear prediction of a meaningful criterion, top-down selection using a linear composite of the standardized scores is statistically the best method for maximizing the utility of criterion-related validity coefficients as they apply to data outside of the sample. This selection method assumes there is no useful curvilinear predictive relationship to be considered in the selection process, yet there is recent literature suggesting some amount of curvilinear prediction exists in the personality arena (Carter, Dalal, Boyce, O'Connell, Kung, & Delgado, 2014; Converse & Oswald, 2014; Le et al., 2011), but perhaps not in the ability arena (Coward & Sackett, 1990; Cullen, Hardison, & Sackett, 2004). The top-down approach also does not account for meaningful nonrandom attrition, such as when the top-ranked talent is more likely to turn down the offer and take a job elsewhere (Murphy, 1986).

Any alternative to top-down selection usually means compromising the validity and utility of the test at least to some extent, and sometimes to a great extent (Schmidt, 1991; Schmidt, Mack, & Hunter, 1984). As we have seen, however, maximizing validity often raises the potential for adverse impact effects against protected racial/ethnic subgroups whenever selection tests have a cognitive ability component. Thus, many larger organizations seek out alternatives to strict top-down selection, trading off validity to some extent in hopes of a resulting increase in diversity.

Four major selection alternatives are prominent in the selection literature and practice. The first alternative is setting a cut score, above which applicants are selected and below which applicants are rejected. Everyone who passes the test is then considered qualified, but the number of job openings is often smaller than the number of qualified applicants. In this case, job offers may be made in a top-down fashion, but then the cut score almost becomes irrelevant. Alternatively, other considerations may come into play once the cut score is passed, such as job experience or other skill sets. In these cases, selection is operating more like a multiple-hurdle selection system.

In fact, setting a cut score for a test as part of a multiple-hurdle selection system is the second major alternative to top-down selection. Those scoring above the cut score move to the next stage of the selection process, which may be another test, or something as simple as an unstructured interview or reference check. Given a large enough sample size to ensure stable and generalizable results, it is possible to establish a multiple-hurdle selection system such that selection cutoffs and the order of the predictors reduce adverse impact ratios, while retaining the highest possible levels of mean predicted performance (De Corte, Lievens, & Sackett, 2006; Sackett & Roth, 1996). Multiple hurdles offer the advantage of reducing the overall cost of the selection procedure, because not all applicants complete each component. This allows the more expensive or time-intensive tests (e.g., simulations, assessment centers) to be administered to smaller numbers of applicants at the end of the process. However, the multiple-hurdle approach critically depends on the assumption that low scores at an early stage of selection should not be compensated for by high scores at a later stage, because they cannot be for those applicants who do not pass a hurdle. A disadvantage of multiple hurdles is that the reliability of the entire selection system is lower compared with the formation of predictor composites, because the reliability of the entire system is the product of the reliabilities of each hurdle in the system established by each measure (Haladyna & Hess, 1999).

The third alternative is the use of test-score banding procedures. Banding is a broad term that encompasses any selection procedure that groups test scores together and considers them to be equivalent. For example, standard error of the difference (SED) banding considers scores to be equivalent unless they are significantly different from each other (Cascio, Outtz, Zedeck, & Goldstein, 1991). The problem with this type of banding is that less reliable measures lead to wider bands and the conclusion that scores are functionally equivalent, meaning that it is a process that generally goes against the principles of good measurement (Schmidt, 1991). Using bands can also make scores less valid, similar to the practice of dichotomizing predictor scores (Cohen, 1983). Several empirical papers explore the tradeoff between maximizing validity or mean predicted performance and minimizing adverse impact as a function of test-score banding (e.g., Campion et al., 2001; Sackett & Roth, 1991; Schmitt & Oswald, 2004). As one would expect, the tradeoff tends to be larger when the bands are larger, when the selection ratio is smaller, and when the standardized mean difference between groups is larger. These rules are not set in stone, however, because results also depend on the statistical banding method used and how the size of the band aligns with the cutoff point for selection in a particular data set. Despite its good intentions, there is surprisingly little evidence that banding has much of a practical effect in reducing adverse impact (Barrett & Lueke, 2004). One exception would be top-down selection of protected group members within bands (Sackett & Roth, 1991), but this is not a viable strategy because the Civil Rights Act of 1991 explicitly prohibits selection on the basis of protected class status without a consent decree, and random selection within bands may actually increase adverse impact (Barrett & Lueke, 2004).

The fourth alternative to top-down selection is to place candidate scores into categories representing probability of success (e.g., green, yellow, red) and reporting those categories to the hiring manager. The purpose of the test scores is to inform the judgment of the selection decision maker, along with other relevant data such as resumes and interviews. This approach is similar to banding, without the statistical algorithms used to determine where the bands are set. The advantage of this approach is that it provides the decision maker with broader latitude for selecting the best candidate without being overly influenced by differences in test scores that may be viewed as very small from a practical standpoint. This may be a disadvantage as well, however, because a more informal approach to selection provides more opportunities for idiosyncratic biases to influence decisions, removing many of the advantages of standardized testing.

Although there may be many situations in which the use of these alternatives to top-down selection would make sense for the organization, we do not recommend adopting them solely for the purpose of reducing adverse impact. When cognitive ability measures are incorporated into a selection battery, large tradeoffs between adverse impact and validity are often impossible to avoid (Sackett & Ellingson, 1997). Although nothing in today's U.S. legal system bars an organization from sacrificing validity to reduce adverse impact, doing so fails to take full advantage of what selection research on validity has to offer in terms of improving the quality of talent that is hired at an aggregate level (Pyburn, Ployhart, & Kravitz, 2008). Furthermore, we would agree that the courts could perceive a procedure that deliberately decreases the validity of the predictors (cognitive and/or noncognitive) as a deliberate decrease in the job relevance of the selection system. Addressing adverse impact concerns in a proactive manner should go well beyond simple predictor selection, weighting, or banding approaches, such as via the active recruitment of minorities, fostering a climate for diversity, and engagement in the diverse communities that the organization serves.

Setting Cut Scores

When it is necessary in a selection system to set one or more cut scores, there are numerous methods from which to choose (see Kehoe & Olson, 2005, and Mueller, Norris, & Oppler, 2007, for extensive reviews). In general, these methods can be distinguished by their use of either judgmental methods or empirical methods. The most common judgmental method is the Angoff (1971) method, in which expert judges (SMEs) estimate the probability that a minimally qualified candidate will answer each test item correctly. These estimates are summed across items to calculate the expected value of the mean test score for minimally qualified candidates. A legitimate criticism of the Angoff method is that judges generally find it difficult to estimate these probabilities, often tending to overestimate them, which leads to higher cut scores than those determined by other methods. The cut score is often adjusted, such as by lowering it one or two standard errors of measurement of the test. Despite this general problem in estimation, Angoff-like methods have the particular advantage of being well received by the courts (Kehoe & Olson, 2005).

Empirical methods for establishing cut scores are generally based on the relationship between test performance and criterion performance, so a criterion-related validation study is required to use these methods. In the regression technique, the minimum criterion score associated with successful job performance is determined, and linear regression is used to find the test score (or test composite score) corresponding to that predicted criterion score. Forward regression regresses criterion scores on test scores (as is done in selection), and reverse regression regresses test scores on criterion scores (to predict the test or composite score cutoff). These methods usually produce different cut scores, so both methods could be used to produce a range of cut scores, and expert judgment could be used to set the appropriate cut score within that range (Mueller et al., 2007).

To illustrate the practical effects of selection, *expectancy charts* usefully depict the relationship between test performance and criterion performance and, optionally, can be used to help set a cut score. Expectancy charts graphically display either the expected criterion performance scores within given ranges of predictor scores or the percentage of those selected who are expected to meet the standard for success on the job, given a set of alternative cut scores. The advantages of using expectancy charts to set cut scores are (a) they are easy for decision makers to understand, (b) the cut score is based on expected criterion performance, and (c) the courts have shown support for these types of methods (Kehoe & Olson, 2005).

Methods for setting cut scores deserve a great deal more attention because cut scores have increasingly been subject to legal challenge. Whenever test users decide to implement cut scores, they should put as much effort into setting them as they should invest in establishing the reliability and validity of the test itself. Test users should carefully consider the need for a cut score, because a top-down selection strategy is often a more desirable alternative. Providing a legal and professional defense for a top-down strategy may be easier than defending how and where a cut

score was established, because there is often a great deal of room for interpretation in the legal and professional literature on cut scores (e.g., what constitutes minimum qualifications). Test users must be careful to have a transparent, justifiable, and consistent rationale based on sound professional judgment for what is done at each step of the cut-score-setting process. Because cut scores are likely to remain in use in many selection and related contexts (e.g., licensure and certification), future research should continue to investigate the major substantive and methodological factors involved in the justification and setting of cut scores.

Selection to Fit a Profile

The selection methods we have reviewed thus far are based on the straightforward linear relationship between predictors and criteria, but some have advocated selection on a more sophisticated basis, such as how well an individual fits a given profile (e.g., McCulloch & Turban, 2007). There are many types of fit (e.g., person-job, person-organization, person-group, person-supervisor; Kristof-Brown, Zimmerman, & Johnson, 2005), but person-organization (P-O) fit seems to be commonly advocated for selection. P-O fit is typically conceptualized as congruence between individual and organizational values or culture and is strongly related to organizational attitudes, such as job satisfaction, organizational trust, and commitment (Kristof-Brown et al., 2005).

Conceptually, P-O fit is thought to predict important organizational criteria such as job performance and turnover in ways that traditional selection measures do not. Empirical support for this is found in a meta-analysis of P-O fit by Arthur, Bell, Villado, and Doverspike (2006), who found corrected mean validities of .15 for predicting job performance and .24 for predicting turnover. Indications were that work attitudes partially mediated the relationships between P-O fit and these criteria, so selection on P-O fit may be more on the basis of job satisfaction than on job performance (see Schmitt, Oswald, Friede, Imus, & Merritt, 2008). This meta-analysis recommended not using P-O fit to make selection decisions in the absence of a local validation study, and to use fit measures as tools post-selection for developmental purposes, such as exploring fit—and changes in fit—when working with employees who may develop performance issues or who are withdrawing and may be considering leaving the organization because of some type of misfit.

A major issue with selection for any type of fit is calculating a fit score. Common techniques are difference scores and correlations between the person's profile and the profile of the organization, where smaller differences between corresponding profile scores and larger correlations across the profile scores are both thought to imply greater fit. Difference scores and their variants (e.g., Euclidean distance) suffer from several methodological problems, including not knowing how much each component of the difference scores contributes to validity, and the compound attenuating effects of measurement error variance on the reliability of the difference scores in the profiles (Edwards, 1994). Arthur et al. (2006) found stronger criterion-related validities when fit was calculated via correlations than via difference scores. Correlations between person and organization profiles are also problematic, however, in that they reflect similarity in profile shape but not the absolute differences between person and organization scores. Also, the relationship between fit correlations and outcomes might be driven by specific variables within the profile. Scores on certain variables rather than the pattern of scores may predict performance. Edwards (1994) demonstrated that polynomial regression is the appropriate analysis method when evaluating the relationship between fit and a criterion, thus overcoming several methodological problems inherent in difference scores. Unfortunately, polynomial regression is a data analysis method and not a method for assigning scores to individuals.

On the basis of current research, we recommend that P-O fit be used for selection only when the goal is to minimize turnover and only when a local validation study is possible. A procedure similar to that of McCulloch and Turban (2007) holds promise and should be legally defensible. These authors had call center managers describe the characteristics of a call center by way of a Q-sort that had them place 54 work descriptors into a normal distribution that was defined along a 9-point scale. Managers came to a consensus solution that defined the call center profile. Call center representatives then sorted the same descriptors in terms of how much

they valued each characteristic. The P-O fit score was the correlation between the individual profile and the call center profile. This score correlated .36 with employee retention and was uncorrelated with job performance, which again is aligned with the general idea of P-O fit being correlated more with attitudinal constructs than with performance.

Context-Based Selection

Related to selection to fit a profile is the idea of taking context into account when making selection decisions. Johns (2006) defined *context* as any aspect of the situation that could affect the occurrence of behavior in an organization or the relationship between variables (e.g., test scores and the criterion). Context could be a cause of organizational behavior (i.e., a main effect), could interact with other variables to influence behavior (i.e., a moderator), or could influence predictor or criterion scores in other ways (e.g., organizational culture may influence how a personality inventory is interpreted in an indirect or multilevel manner). By measuring relevant context variables, the prediction of performance can sometimes be improved by considering the unique aspects of the organization and/or job. Attending to context could, therefore, influence the use or weighting of test scores in making selection decisions.

As an illustration of the influence of context on variable relationships, Tett, Jackson, Rothstein, and Reddon (1994, 1999) showed that personality scales may be positively correlated with a performance dimension in some situations, yet have negative correlations in other situations. For example, a person high in agreeableness may do well in an organization that has a team-based, cooperative culture, but that same person may have difficulty in an organization with a culture that is highly competitive. This suggests that the first organization should select applicants who score high on agreeableness, but the second organization would be better off selecting applicants with a measure of competitiveness or achievement orientation.

Research on context-based selection is in its infancy because of the large sample sizes necessary and measurement issues associated with identifying moderator variables in personnel selection research. Johns (2006) presents a thorough review of issues associated with studying context. For example, context often has a cross-level effect (e.g., organizational strategy influences the evaluation of individual behavior), whereas selection research typically focuses on measuring variables at the individual level (e.g., test scores, performance ratings). When measuring variables at a higher level than the individual, there must be enough data at the higher level to adequately evaluate its impact on variables at the lower level. This is very difficult if one is evaluating the impact of an organizational-level variable, because data must be collected from multiple organizations, and there must be enough variability across organizations to properly evaluate the effect. The typical local validation study cannot meet this requirement, although a consortium study or meta-analysis may allow for tests of organization-level moderators. Team- or role-level context variables, however, would likely show reliable variance within an organization.

One way to test the effects of organization-level moderators without conducting a multilevel analysis is to evaluate the impact of raters' perceptions of organization-level variables (e.g., organizational strategy, business priorities) on their perceptions of individual job performance. Of course, if raters within an organization agree on the organization's standing on these variables, the limitation of not having enough variance to test for moderation still exists. As an example of this type of analysis, Johnson (2016) used data collected across multiple organizations as part of the CEB Leadership Study (LoVerde & Schmidt, 2016) to demonstrate the moderating effect of several organizational context variables. For example, when managers had stronger perceptions that the organization's future growth would come through innovation, the relationship between a personality measure of network leadership potential and manager ratings of network leadership performance was stronger than when innovation was not as much of a priority (i.e., there was a moderator effect).

There is a risk in studying context effects in isolation, because there are always many different contexts operating simultaneously and possibly interacting. It is difficult or impractical to consider multiple contextual variables simultaneously. Similarly, it is impossible to identify, much less measure, all contextual variables that could be relevant. Nevertheless, it is still better to consider

a small number of contextual variables than to ignore context altogether, as is often done. Considering context in selection research has the potential to (a) improve prediction beyond what is typically seen in criterion-related validation studies; (b) identify candidates who are in alignment with the organization's culture, strategy, and priorities; and (c) match candidates to specific roles they are most ready and equipped to perform.

CONCLUSION

Personnel selection makes a critical contribution to the system of organizational policies and practices to which it is related. Key selection questions are worth asking and addressing repeatedly as organizational researchers and practitioners, because both the questions and the answers adapt to fit the organizational setting and the current state of the art. In this chapter, we highlighted questions that are essential to address (a) at the outset of the testing program, (b) with regard to collection of test scores, (c) when computing test scores, and (d) when making selection decisions. If these questions are not addressed mindfully, they will likely be addressed by default. Test users should be aware of the organizational implications of each decision to ensure that the testing program is consistent with other organizational goals, and they should be aware of the current legal context and implications of each decision made to avoid potential litigation problems. When used properly by experienced professionals in the context of other organizational practices, selection test scores have proven to have a very positive influence on individuals and organizations alike.

REFERENCES

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Arthur, W., Bell, S. T., Villado, A. J., & Doverspike, D. (2006). The use of person-organization fit in employment decision making: An assessment of its criterion-related validity. *Journal of Applied Psychology, 91*, 786–801.

Barrett, G. V., & Lueke, S. B. (2004). Legal and practical implications of banding for personnel selection. In H. Aguinis (Ed.), *Test-score banding in human resource selection: Technical, legal, and societal issues* (pp. 71–111). Westport, CT: Praeger.

Bartram, D. (2009). The International Test Commission guidelines on computer-based and Internet-delivered training. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 11–13.

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.

Brannick, M. T., & Darling, R. W. (1991). Specifying importance weights consistent with a covariance structure. *Organizational Behavior and Human Decision Processes, 50*, 395–410.

Burke, E. F. (1997). A short note on the persistence of retest effects on aptitude scores. *Journal of Occupational and Organizational Psychology, 70*, 295–301.

Burke, E. F. (2009). Preserving the integrity of online testing. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 35–38.

Campbell, J. P. (1996). Group differences and personnel decisions: Validity, fairness, and affirmative action. *Journal of Vocational Behavior, 49*, 122–158.

Campbell, W. J., & Reilly, M. E. (2000). Accommodations for persons with disabilities. In J. F. Kehoe (Ed.), *Managing selection in changing organizations* (pp. 319–370). San Francisco, CA: Jossey-Bass.

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.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

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., 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.

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment*, *20*, 333–346.

Chan, D., & Schmitt, N. (1997). Video versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143–159.

Chang, H. H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items on adaptive testing. *Psychometrika, 73*, 441–450.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249–254.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304–1312.

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. (1990). *Essentials of psychological testing* (5th ed.). New York, NY: HarperCollins.

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89*, 220–230.

De Beuckelaer, A., & Lievens, F. (2009). Measurement equivalence of paper-and-pencil and Internet organisational surveys: A large scale examination in 16 countries. *Applied Psychology: An International Review, 58*, 336–361.

De Corte, W., Lievens, F., & Sackett, P. R. (2006). Predicting adverse impact and mean criterion performance in multistage selection. *Journal of Applied Psychology, 91*, 523–537.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92*, 1380–1393.

DeVellis, R. F. (2016). *Scale development: Theory and applications* (4th ed.). Newbury Park, CA: Sage.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18*, 192–203.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Potential solutions to practical equating issues*. New York, NY: Springer.

Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes, 58*, 51–100.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357–381.

Gibson, W. M., & Weiner, J. A. (1998). Generating random parallel test forms using CTT in a computer-based environment. *Journal of Educational Measurement, 35*, 297–310.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84–96.

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

Green, B. F. (1977). Parameter sensitivity in multivariate methods. *Multivariate Behavioral Research, 12*, 264–288.

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions*. New York, NY: Routledge.

Haladyna, T., & Hess, R. (1999). An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. *Educational Assessment, 6*, 129–153.

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*, 373–385.

Huselid, M. A., Jackson, S. E., & Schuler, R. S. (1997). Technical and strategic human resource management effectiveness as determinants of firm performance. *Academy of Management Journal, 40*, 171–188.

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

Janz, T., Hellervik, L., & Gilmore, D. (1986). *Behavior description interviewing: New, accurate, cost effective*. Boston, MA: Allyn and Bacon.

Johns, G. (2006). The essential impact of context on organizational behavior. *Academy of Management Review*, *31*, 386–408.

Johnson, J. W. (April 2016). Enhancing our understanding of the network leadership construct. In M. A. LoVerde (Chair), *Overview and selected finding from a multi-organizational, multi-level leadership study*. Symposium

conducted at the 31st Annual Conference of the Society for Industrial and Organizational Psychology, Anaheim, CA.

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

Kantrowitz, T. M., & Dainis, A. M. (2014). How secure are unproctored pre-employment tests? Analysis of inconsistent test scores. *Journal of Business and Psychology, 29*, 605–616.

Kehoe, J. F., & Olson, A. (2005). Cut scores and employment discrimination litigation. In F. J. Landy (Ed.), *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives* (pp. 410–449). San Francisco, CA: Jossey-Bass.

Kim, D-I., Choi, S. W., Lee, G., & Um, K. R. (2008). A comparison of the common-item and random-groups equating designs using empirical data. *International Journal of Selection and Assessment, 16*, 83–92.

Koopman, R. F. (1988). On the sensitivity of a composite to its weights. *Psychometrika, 53*, 547–552.

Kravitz, D. A. (2008). The validity-diversity dilemma: Beyond selection—The role of affirmative action. *Personnel Psychology, 61*, 173–193.

Kristof-Brown, A. L., Zimmerman, R. D., & Johnson, E. C. (2005). Consequences of individuals' fit at work: A meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. *Personnel Psychology, 58*, 281–342.

Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology*, *96*, 202–210.

Le, H., Oh, I-S., Robbins, S. B., Ilies, R., Holland, E., & Westrick, P. (2011). Too much of a good thing: Curvilinear relationships between personality traits and job performance. *Journal of Applied Psychology, 96*, 113–133.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58*, 981–1007.

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology, 92*, 1043–1055.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service.

LoVerde, M. A., & Schmidt, C. (April 2016). Overview of the CEB Leadership Study (CLS). In M. A. LoVerde (Chair), *Overview and selected finding from a multi-organizational, multi-level leadership study*. Symposium conducted at the 31st Annual Conference of the Society for Industrial and Organizational Psychology, Anaheim, CA.

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person characteristics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*, 921–943.

Makransky, G., & Glas, C. A. W. (2011). Unproctored internet testing verification: Using adaptive confirmation testing. *Organizational Research Methods, 14*, 608–630.

McCulloch, M. C., & Turban, D. B. (2007). Using person-organization fit to select employees for high-turnover jobs. *International Journal of Selection and Assessment, 15*, 63–71.

McPhail, S. M. (2007). *Alternative validation strategies: Developing new and leveraging existing validity evidence*. San Francisco, CA: Jossey-Bass.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 611–635.

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods*, *10*, 322–345.

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*, 93–115.

Mueller, L., Norris, D., & Oppler, S. (2007). Implementation based on alternate validation procedures: Ranking, cut 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. (1986). When your top choice turns you down: Effect of rejected offers on the utility of selection tests. *Psychological Bulletin*, *99*, 133–138.

Naus, M. J., Philipp, L. M., & Samsi, M. (2009). From paper to pixels: A comparison of paper and computer formats in psychological assessment. *Computers in Human Behavior*, *25*, 1–7.

Nye, C. D., Do, B.-R., Drasgow, F., & Fine, S. (2008). Two-step testing in employment selection: Is score inflation a problem? *International Journal of Selection and Testing, 16*, 112–120.

Oswald, F. L., Friede, A. J., Schmitt, N., Kim, B. H., & Ramsay, L. J. (2005). Extending a practical method for developing alternate test forms using independent sets of items. *Organizational Research Methods, 8*, 149–164.

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 myths and urban legends* (pp. 187–205). New York, NY: Taylor & Francis.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–263). New York, NY: American Council on Education and Macmillan.

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

Ployhart, R. E., & MacKenzie, J. I., Jr. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedek (Ed.), *APA handbook of industrial and organizational psychology* (Vol. 2, pp. 237–252). Washington, DC: American Psychological Association.

Ployhart, R. E., Weekley, J. A., & Holtz. B. C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology*, *56*, 733–752.

Potosky, D., Bobko, P., & Roth, P. L. (2005). Forming composites of cognitive ability and alternative measures to predict job performance and reduce adverse impact: Corrected estimates and realistic expectations. *International Journal of Selection and Assessment, 13*, 304–315.

Pyburn, K. M., Jr., Ployhart, R. E., & Kravitz, D. (2008). The validity-diversity dilemma: Overview and legal context. *Personnel Psychology, 61*, 143–151.

Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology, 60*, 367–396.

Ree, M., Carretta, T., & Earles, J. (1998). In top-down decisions, weighting variables does not matter: A consequence of Wilks' theorem. *Organizational Research Methods, 1*, 407–420.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297–330.

Roth, P. L, Switzer, F. S., III, Van Iddekinge, C. H., & Oh, I-S. (2011). Toward better meta-analytic matrices: How input values can affect research conclusions in human resource management simulations. *Personnel Psychology, 64*, 899–935.

Sackett, P. R., & Ellingson, J. E. (1997). The effect of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*, 707–721.

Sackett, P. R., & Roth, L. (1991). A Monte Carlo examination of banding and rank order methods of test score use in personnel selection. *Human Performance, 4*, 279–295.

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. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56*, 302–318.

Schmidt, F. L. (1971). The relative efficiency of relative and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement, 31*, 699–714.

Schmidt, F. L. (1991). Why all banding procedures in personnel selection are logically flawed. *Human Performance, 4*, 265–277.

Schmidt, F. L., Mack, M. J., & Hunter, J. E. (1984). Selection utility in the occupation of U.S. park ranger for three modes of test use. *Journal of Applied Psychology, 69*, 490–497.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210–222.

Schmitt, N., & Oswald, F. L. (2004). Statistical weights of ability and diversity in selection decisions based on various methods of test-score use. In H. Aguinis (Ed.), *Test-score banding in human resource selection: Technical, legal, and societal issues* (pp. 113–131). Westport, CT: Praeger.

Schmitt, N., Oswald, F. L., Friede, A., Imus, A., & Merritt, S. (2008). Perceived fit with an academic environment: Attitudinal and behavioral outcomes. *Journal of Vocational Psychology, 72*, 317–335.

Scullen, S. E., & Meyer, B. (2012). More applicants or more applications per applicant? A big question when pools are small. *Journal of Management, 14*, 1675–1699.

Segall, D. O. (April 2001). *Detecting test compromise in high-stakes computerized adaptive testing: A verification testing approach*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

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

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55*, 167–194.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item functioning and differential test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497–508.

Tett, R. P., Jackson, D. N., Rothstein, M., & Reddon, J. R. (1994). Meta-analysis of personality-job performance relations: A reply to Ones, Mount, Barrick, and Hunter (1994). *Personnel Psychology*, *47*, 157–172.

Tett, R. P., Jackson, D. N., Rothstein, M., & Reddon, J. R. (1999). Meta-analysis of bi-directional relations in personality-job performance research. *Human Performance, 12*, 1–29.

Tippins, N. T. (2009). Where is the unproctored Internet testing train headed now? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 69–76.

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*, 189–225.

Truxillo, D. M., & Bauer, T. N. (1999). Applicant reactions to test score banding in entry-level and promotional contexts. *Journal of Applied Psychology, 84*, 322–339.

Truxillo, D. M., & Bauer, T. N. (2000). The roles of gender and affirmative action in reactions to test score use methods. *Journal of Applied Social Psychology, 30*, 1812–1828.

U.S. Department of Labor. (1999). *Testing and assessment: An employer's guide to good practices*. Washington, DC: U.S. Department of Labor, Employment and Training Administration.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer.

Viswesvaran, C., & Ones, D. S., (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197–210.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*, 25–49.