1-2023

Personnel selection: A review of ways to maximize validity, diversity, and the applicant experience

Chad H. Van Iddekinge

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg

Paul R. Sackett

Citation

Van Iddekinge, Chad H.; LIEVENS, Filip; and Sackett, Paul R.. Personnel selection: A review of ways to maximize validity, diversity, and the applicant experience. (2023). *Personnel Psychology*.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/7186

Personnel Selection: A Review of Ways to Maximize Validity, Diversity, and the Applicant Experience

Chad Van Iddekinge

Filip Lievens

Paul R. Sackett

In press *Personnel Psychology*

Abstract

Personnel Psychology has a long tradition of publishing important research on personnel selection. In this article, we review some of the key questions and findings from studies published in the journal and in the selection literature more broadly. In doing so, we focus on the various decisions organizations face regarding selection procedure development (e.g., use multiple selection procedures, contextualize procedure content), administration (e.g., provide pre-test explanations, reveal target KSAOs), and scoring (e.g., weight predictors and criteria, use artificial intelligence). Further, we focus on how these decisions affect the validity of inferences drawn from the procedures, how use of the procedures may affect organizational diversity, and how applicants experience the procedures. We also consider factors such as cost and time. Based on our review, we highlight practical implications and key directions for future research.

Personnel Selection: A Review of Ways to Maximize Validity, Diversity, and the Applicant Experience

In industrial and organizational psychology, few domains, if any, have generated such consistent interest among academicians and practitioners as has personnel selection (Lievens & Sackett, 2017). Selection involves identifying the knowledge, skills, abilities, and other characteristics (KSAOs) needed for effective job performance, developing valid and reliable measures of these KSAOs, administering the assessments to job applicants, and using the scores to make hiring decisions (Schmitt & Chan, 1998). Selection is rooted in the "predictive hypothesis" (Guion, 1965) that applicants who possess higher levels of the critical KSAOs also will perform better on the job if selected.

Personnel Psychology has long tradition of publishing selection research (and we italicized citations to articles in the journal throughout our review). The very first issue of the journal included articles about using vision tests to predict job performance (*Kerr, 1948*), using vocational interest measures to predict turnover (*Bolanovich, 1948*), and how using standard selection processes can attract better applicants (*Stromberg, 1948*). Since then, articles published in the journal have made substantial contributions to understanding and improving how organizations make selection decisions. Early articles focused on the criterion-related validity of different selection procedures, including selection interviews (e.g., *Wagner, 1949; Mayfield, 1964*), personality tests (e.g., *Guion & Gottier, 1965*), aptitude tests (e.g., *Ghiselli, 1973*), and work sample tests (e.g., *Asher & Sciarrino, 1974*), as well as their economic utility (e.g., *Brogden, 1949*). After the passing of the Civil Rights Act (1964), issues such as demographic group differences and biases began to receive more attention (e.g., *Cann et al., 1981; Reilly & Chao, 1982; Schein, 1978*). In the 1990s, *Personnel Psychology* published articles that reconsidered the criterion-related validity of personality tests (e.g., *Barrick & Mount, 1991; Cortina et al., 1992*), as well as some of the initial work on situational judgment tests (SJTs; e.g., *Weekly & Jones, 1997*). The journal also published summaries of the Army Project A validation project (e.g., *McHenry et al., 1990*). In more recent years, *Personnel Psychology* has published key primary studies and reviews on topics such as selection across cultures (e.g., *Ryan & McFarland, 1999*), web-based selection (e.g., *Ployhart et al., 2003*), applicant reactions to selection procedures (e.g., *Hausknecht et al., 2004*), adverse impact (e.g., *Ployhart & Holtz, 2008*), and, once again, revisiting the use of personality tests for selection (e.g., *Morgeson et al., 2007*). The journal also published an introduction to the Occupational Information Network (O*Net; *Peterson et al., 2001*), which has had a substantial impact on how organizations and researchers identify attributes for selection.

In this article, we review some of the key questions and findings in the selection literature. In doing so, we focus on development, administration, and scoring decisions that can affect validity, diversity, and the applicant experience. Selection research traditionally has focused on identifying selection procedures that predict future performance (i.e., validity) and minimize subgroup differences that can limit diversity and lead to adverse impact. Applicant reactions to selection procedures also have emerged as a consideration, and more recently, how applicants experience the procedures. Indeed, user experience and brand reputation now play an important role in attracting talent (Landers et al., 2021).

Given the vast amount of research in this area, it was important to establish some parameters for our review. First, although we review research from various sources, we devote particular attention to contributions articles published in *Personnel Psychology* have made to selection. Second, *Personnel Psychology* has made important contributions to understanding constructs organizations often assess during the selection process, such as personality traits (e.g., *Roberts et al., 2005*), vocational interests (e.g., *Barrick et al., 2003*), person-organization fit (e.g., *Kristof, 1996*), and background characteristics such as education (e.g., *Ng & Feldman, 2009*), work experience (e.g., *Quiṅones et al., 1995*), and biographical data (e.g., *Mael, 1991*). The journal also has published key articles on the methods that can be used to assess these and other job-relevant constructs, including assessment centers (e.g., *Arthur et al., 2003*), integrity tests (e.g., *Sackett & Wanek, 1996*), selection interviews (e.g., *Arvey & Campion, 1982*), SJTs (e.g., *McDaniel et al., 2007*), and work samples (e.g., *Roth et al., 2005*). Although we sometimes refer to this research, we focus on decisions surrounding the development, administration, and scoring of selection procedures more generally (see Sackett et al. [2017] for a review of selection constructs and Lievens and Sackett [2017] for a review of selection methods). Third, although research has provided important insights regarding selection in academic contexts (e.g., Oswald et al., 2004; Shultz & Zedeck, 2011), we focus primarily on selection in work organizations.

Fourth, job analysis (e.g., Harvey, 1991; *Raymark et al., 1997*) and competency modeling (e.g., *Lievens et al., 2004; Schipmann et al., 2000*) are foundational to identifying attributes to assess during the selection process. However, we do not review the substantial research on these two processes (see *Campion et al. [2011]* and Sanchez and Levine [2012] for reviews). Fifth, although we address the validity of selection procedures, we do not attempt to cover the myriad issues involved in validating the procedures (see *Van Iddekinge and Ployhart [2008]* for a review). For example, we do not detail decisions involving sampling (e.g., applicants vs. incumbents; Guion & Cranny, 1982) or correction of statistical artifacts such as range restriction (e.g., *Schmidt et al., 2008*) and measurement error (e.g., *Murphy & DeShon, 2000*). Fifth, although accurate specification and measurement of criteria is critical for developing and validating selection procedures, we do not attempt to review the large literature on criteria (see DeNisi and Murphy [2017] for a review). For instance, in the performance domain, we do not review topics such as the dimensionality of performance (e.g., *Campbell et al., 1990*), maximal versus typical performance (e.g., *Ployhart et al., 2001*), dynamic performance (e.g., *Ployhart & Hakel, 1998*), or objective versus subjective criteria (e.g., *Bommer et al., 1995*).

The review will unfold as follows. First, we discuss some of the main criteria on which selection procedures—and comparisons of the relative merit of different selection procedures—are evaluated. Second, we review research relevant to development, administration, and scoring decisions that can affect these criteria. Finally, although we note research gaps throughout the review, in the last section, we focus on several key directions for future research.

Factors on which Selection Procedures are Evaluated

We begin by addressing the main factors on which selection procedures are evaluated, including validity, subgroup differences, applicant reactions, and cost and time. We also address the effects of selection procedures on unit- and firm-level outcomes.

Criterion-Related and Content Validity

Modern conceptualizations of validation use the term to refer to the process of gathering evidence to support the inferences one wishes to draw from a score (Sackett et al., 2012). Thus, validity is not a property of an assessment per se, but of these inferences. Some inferences are descriptive (“score reflects applicant’s current level of job knowledge”), whereas others are predictive (“scores can be used to forecast the level of performance if the person is hired”).

In the selection context, this predictive inference is most central, and different validation strategies can be used to establish this inference (Sackett et al., 2012). Two broad categories of strategies reflect the classic distinction drawn by Wernimont and Campbell (1968) between "signs" and "samples". A "sign" strategy views scores on a predictor as a sign or signal of future performance and relies on evidence that individuals with higher predictor scores subsequently perform better on a criterion on interest (labeled "criterion-related validity"). The strategy does not specify how or why the predictor and criterion are related; the empirical relationship is the basis for offering support for the predictive inference. In contrast, a "sample" strategy views scores on a predictor as a direct or indirect sample of the criterion domain of interest. It is applicable when a predictor directly reflects the criterion (e.g., job tryout) or a reasonable simulation of the criterion domain (e.g., work sample test). The evidence for the predictive inference when relying on a sampling strategy is judgmental (labeled "content validity"). That is, evidence is assembled to support the inference that the behaviors reflected in the predictor reasonably sample the criterion domain.

Operationally, use of the "sign" strategy makes use of criterion-related validity, which is assessed by correlating scores on the predictor variable of interest—obtained from a sample of applicants or current employees—with measures of outcomes of interest (e.g., job performance, turnover, advancement). Developments in the use of this strategy include (a) the key role of sampling error, and hence the need for larger samples (e.g., Schmidt et al., 1976); (b) attenuation due to measurement error in the criterion (e.g., Viswesvaran et al., 1996) and differences in predictor variability between the research sample and applicant pool (e.g., Sackett & Yang, 2000); and (c) a move from relying on local validity evidence to cumulative evidence from other settings (e.g., Schmidt & Hunter, 1981). Meta-analytic estimates of the mean and variance of criterion-related validity of predictors have been obtained (e.g., Sackett et al., 2021; Schmidt & Hunter, 1998), as have meta-analytic estimates of the incremental validity of different predictors (e.g., *Bobko et al., 1999; Cortina et al., 2000).*

"Sampling" commonly relies on a content validation strategy, which assesses the degree to which the predictor samples the criterion domain. There has been debate and confusion about content validation (e.g., *Guion, 1978; Tenopyr, 1978*), with the *Uniform Guidelines on Employee Selection Procedures* (1978) arguing that content validity is not applicable to predictors that measure constructs. This is at odds with contemporary perspectives that argue that constructs are evoked whenever a substantive interpretation is applied to a score (e.g., Schmidt, 2012; Sackett, 2012). Useful models to operationalize content validation have been put forward (e.g., Goldstein et al., 1993), as have approaches to quantify the judgments involved in content validation (e.g., Colquitt et al., 2019; *Lawshe, 1975*). Further, although criterion-related validation requires large samples, content validation is applicable to settings that vary in size and scope *(Robinson, 1981).*

Subgroup Differences, Adverse Impact, and Bias

The Civil Rights Act of 1964 was a pivotal event for selection research and practice in the U.S, with other countries subsequently influenced by developments in the U.S. (Shen et al., 2017). Court decisions interpreting the Civil Rights Act led to a legal framework that (a) defines adverse impact as when a procedure selects members of minority groups (regarding race, color, religion, sex, and national origin) at lower rates than members of majority groups, (b) requires documenting validity to justify predictor use in the presence of adverse impact, and (c) permits rebutting validity evidence with the demonstration of available alternate predictors with equal validity and less adverse impact (Uniform Guidelines, 1978).

It has become clear that adverse impact for a given predictor or predictor composite is determined by two factors: (a) the standardized predictor mean difference (*d*) between the majority and minority groups of interest and (b) the selection ratio. *Sackett and Ellingson (1997)* provided useful tables showing the adverse impact resulting from various combinations of *d* and selection ratio. Group mean differences have been examined for a range of predictors and for various subgroup comparisons, most frequently male-female, White-Black, and White-Hispanic. Summaries of these differences can be found in Hough et al. (2001), Dahlke and Sackett (2017), and Roth et al. (2017). Dahlke and Sackett provided insight into the drivers of racio-ethnic subgroup differences. Across 38 predictors, the correlation between White-Black *d* and the cognitive load of the predictor (i.e., its correlation with general cognitive ability) was .84.

Considerable research has examined ways to reduce adverse impact associated with cognitive-oriented selection procedures. Sackett et al. (2001) and *Ployhart and Holtz (2008)* reviewed the efficacy of potential strategies, including changing test format, changing the testing environment, dropping problematic items, and supplementing with non-cognitive predictors. Among these options, supplementing with additional predictors has the potential to both reduce subgroup differences and increase validity.

A key question is whether subgroup differences on a predictor reflect true differences or bias. There is general recognition that *d* might reflect true differences between the groups, and that a mechanism other than *d* is needed to identify bias. Comparing the regression lines relating predictor and criterion scores for majority and minority subgroups emerged as the preferred approach to identifying predictive bias (Cleary, 1968). Evidence of predictive bias has been compiled for some predictors, but not for others. In the U.S. legal regulatory system, investigating predictive bias is required only if adverse impact is found and, thus, there has been little examination of low-*d* predictors (e.g., personality tests). In contrast, there has been extensive investigation of cognitive ability measures (e.g., *Bartlett et al., 1978*; Hartigan & Wigdor, 1989; Sackett et al., 2021). The general finding is that test scores do not show predictive bias against racio-ethnic subgroups (Berry, 2015).

Finally, three recent insights into predictive bias merit attention. First, the appropriate analysis is at the level of a selection system, not at the level of the individual predictor in cases where a composite of multiple predictors is used. Failure to do so creates an omitted variable problem, a violation of the assumptions of the underlying regression model (Sackett et al., 2003). Second, traditional approaches focus on the statistical significance of differences between group-specific regression lines (Berry, 2015), whereas modern approaches focus more on the magnitude of difference between the regression lines (e.g., Dahlke, & Sackett, 2021; Nye & Sackett, 2017). Third, a useful distinction between the terms “fairness” and “bias” has emerged (SIOP, 2018). Bias is a technical term and can be assessed as discussed above, whereas fairness is a value statement. Some may assert than any predictor exhibiting subgroup differences is unfair, whereas others may assert that group differences would only signal unfairness if accompanied by predictive bias against the minority group(s).

Applicant Reactions

Recently, the importance of “the candidate experience” has increased. The candidate experience focuses on how applicants experience an organization’s staffing process, from first recruitment to assessment and selection to onboarding (Miles & McCamey, 2018). In the selection literature, considerable research has examined how applicants perceive the assessment stage of the staffing process. Initial work was based on organizational justice theory and focused on how applicants perceive the fairness of selection procedures (e.g., Gilliland, 1993, 1994). Subsequent research identified a range of applicant reactions. For example, McCarthy et al. (2013) distinguished between dispositional-based reactions such as test-taking anxiety and efficacy versus situational-based reactions such as job-relatedness and opportunity to perform.

Some of the first and most influential studies in this area were published in *Personnel*

Psychology. For example, *Arvey et al. (1990)* developed one of the first measures of applicant reactions, and *Bauer et al. (2001)* developed and validated the Selection Procedural Justice Scale, which has been widely used in subsequent applicant reactions research. In addition, *Smither et al. (1993)* and *Macan et al. (1994)* were among the first studies to compare how applicants perceive different selection procedures. More recently, *Hausknecht et al. (2004)* published a seminal meta-analysis that revealed that applicants tend to react more favorably to procedures such as interviews, work samples, and resume screens than to personality inventories, integrity tests, and biodata. Further, applicant reactions were positively related to perceived and actual performance on the selection procedures. In addition, applicants who reacted favorably to selection procedures had higher intentions to accept job offers and to recommend the employer to others.

A persistent question is how much attention should applicant reactions receive when choosing and developing selection procedures. Studies that have tried to address this question have reported mixed results. *McCarthy et al. (2009)* conducted two studies in which they correlated candidate reactions to a promotional exam to performance on the exam. Further, candidates did not yet know the outcome of the process, so their reactions were not affected by the outcome. Reactions were more strongly related to self-rated exam performance than to actual exam performance. *Harold et al. (2016)* found that fairness perceptions predicted job offer acceptance beyond factors such as perceived person-organization fit and recruiter behaviors. In contrast, Konradt et al. (2017) found that reactions did not predict job acceptance decisions. McCarthy et al. (2013) discovered that although reactions were related to performance on selection procedures, they did not affect (e.g., weaken) the criterion-related validity of the selection procedures. Finally, although some studies suggest that reactions may help explain demographic differences in test scores (e.g., Arvey et al., 1990), these effects generally have not panned out (Hausknecht et al., 2004; McCarthy et al., 2017). Overall, although applicant reactions can affect how they perceive and perform on selection procedures, reactions do not tend to have consistent effects on outcomes such as predictive validity or subgroup differences.

Cost and Time

Consideration of cost is an inevitable part of the selection system design process. There is obvious appeal to inexpensive selection systems (e.g., single- vs. multi-component systems, self-reports vs. simulations or work samples). A case for a costly selection system typically relies on other valued attributes, such as greater validity or a better candidate experience.

One attempt to formally integrate costs into the evaluation of selection systems involves the use of utility analysis (*Brogden, 1949*). Conceptually, the utility of a selection system can be viewed in terms of a simple equation: Quantity (number selected using the system x average tenure) x Quality (the difference in the mean dollar-valued performance of those selected using the system vs. an alternative system) - Cost (development costs plus per candidate assessment costs). Various models incorporating these ideas have been presented, including different approaches of estimating the dollar value of performance (e.g., Cascio, 1982; *Schmidt et al., 1982*) and additional inputs to consider, such as variable costs and taxes (e.g., *Boudreau, 1983*). However, concerns about conveying these ideas credibly to organizational decision makers have received considerable attention (*Cronshaw, 1997; Latham & Whyte, 1994)*.

Another approach to selection system design views costs as a constraint. De Corte et al. (2011) offered an approach to deriving selection systems that are optimal with regards to a specified set of outcomes (e.g., validity, adverse impact reduction). Their approach permits the specification of a cost constraint (e.g., no more than X dollars per candidate), and thus limits the set of potential selection systems to those meeting the constraint. For example, a cost constraint may preclude a system that includes an in-person interview with all candidates but permits a system that screens out 80% of candidates based on less expensive predictors, with the remaining 20% interviewed.

Time is another consideration in designing selection systems. All else equal, longer assessments are more reliable, and including more job-relevant attributes in an assessment battery increases the validity of the resulting system (Dahlke & Sackett, 2021). One concern is that lengthy assessments will lead candidates to withdraw from the process, particularly stronger applicants who have more options. However, recent research has challenged this perspective. Hardy et al. (2017) examined the rate of non-completion of assessments across 69 different selection systems with a mean assessment length of 67 minutes. On average, 21% of individuals who began an assessment failed to complete it. Across the systems, test length was not related to withdrawal, nor to accepting an initial invitation to take the assessment. There was less withdrawal for jobs with higher salary and for organizations with a more positive image. Hardy et al. (2021) extended this by examining candidate quality. Candidates who failed to complete an assessment had lower scores on the completed components than candidates who completed the assessment. This finding counters the claim that stronger candidates are less likely to persist in the face of lengthier assessments. Thus, evidence to date does not support sacrificing other desirable attributes (e.g., reliability, validity) as means of reducing non-completion rates.

Effects on Unit- and Firm-level Outcomes

In recent years, researchers have devoted much more attention to the effects of HR practices—including selection—on unit- and firm-level performance and retention. Selection is thought to influence collective outcomes through its influence on human capital resources, which are a unit-level resource created from the emergence of employees' KSAOs (Ployhart & Moliterno, 2011). Selection is the primary means by which organizations acquire human capital resources, particularly more generic types such as general mental ability. These resources, in turn, are thought to provide a basis for the development of more specific human capital resources, such as job-specific knowledge and skills (Ployhart et al., 2011).

Most research has included selection (or recruiting and selection) as one of several high-performance work practices (HPWPs, Huselid, 1995) and typically has not reported the unique effects of selection practices. However, a few studies focused on selection or reported separate results. *Terpstra and Rozell (1993)* conducted one of the first studies using this approach. They surveyed the HR leaders about their firm's staffing practices, such as the use of cognitive ability tests, biographical data, and structured interviews, as well as the extent to which firms validated their selection procedures. Results revealed a positive, yet modest, relationship between the number staffing practices used and measures of financial performance. Similarly, Hatch and Dyer (2004) found that manufacturing firms that used a selection test to assess technical skills had fewer product defects. In a meta-analysis of HPWPs, *Combs et al. (2006)* reported a correlation (corrected for measurement error in both variables) of .14 (k = 15, N = 3,689) between "selectivity" and firm performance (but they did not define selectivity nor describe how the primary studies measured it). More recently, Kim and Ployhart (2014) found that selective staffing (operationalized as the selection ratio) predicted labor productivity, which, in turn, predicted profits. Kim and Ployhart (2018) showed that a composite of 11 selection practices (e.g., use of cognitive ability tests, interviews, and assessment centers) was positively related to productivity (e.g., sales per employee), negatively related to collective turnover, but unrelated to profitability once other factors were considered. Further, among firms with lower turnover or that operate in dynamic industries, the effects of selection on productivity were positive, whereas selection demonstrated negative effects among firms with higher turnover or that operate in more stable environments.

Research likely underestimates the value of selection given the coarse measures most studies have used, such as “yes/no” judgments concerning whether an organization uses certain types of selection tests (e.g., Hatch & Dyer, 2004) or the proportion of employees who were given a selection test prior to hiring (e.g., Huselid, 1995). Other studies (e.g., *Gerhart et al., 2000*) have highlighted limitations of measuring selection and other HPWPs using subjective evaluations of individual raters (e.g., HR managers) due to factors such as variability in the practices across the organization (e.g., selection tests are used for some jobs or units but not others). In addition, most studies have correlated selection with distal outcomes (e.g., financial performance) that are influenced by myriad other factors. The few studies that have used more rigorous measures of effective selection and more proximal outcomes have tended to find stronger effects. For instance, Van Iddekinge et al. (2009) found that use of a validated selection test was associated with higher unit-level customer service performance and retention, which, in turn, were related to better financial performance. Finally, research suggests that HPWPs (including selection) often are similarly related to past, current, and future outcomes (e.g., Wright et al., 2005). This raises questions about whether HPWPs cause the outcomes such as firm performance, whether the outcomes cause HPWPs (e.g., firms have more resources to fund such practices), or whether the two variables are reciprocally related.

Development, Administration, and Scoring Decisions that Can Affect Validity, Subgroup Differences, and the Applicant Experience

We now review decisions involving how to develop, administer, and score selection procedures. We used an iterative process to identify decisions relevant to each of these three steps in the selection process. Specifically, we identified an initial list of decisions and continually revised the list as we reviewed the literature. The final list reflects decisions selection researchers and practitioners commonly would need to make and that have received research attention. Table 1 provides implications related to each decision.

Development Decisions

Use multiple selection procedures. Although much selection research has focused on the validity of a single predictor, it is common for selection systems to include multiple predictors. One reason is an expectation that a system that captures more of the important attributes related to job performance will better predict performance than a narrower system. Another reason is that a broader selection system is likely to be more legally defensible. In the U.S. legal system, if a proposed system results in adverse impact against one or more protected groups, there is an expectation that the organization conduct a search for alternatives with comparable validity but less adverse impact. Should a job analysis suggest the relevance of a broader set of attributes—particularly those for which there is evidence of smaller subgroup differences—a multi-predictor system emerges as a desirable alternative (Sackett et al., 2001).

One body of literature frames the issue in terms of an initial selection system with a single predictor with a large subgroup difference (e.g., cognitive ability), and explores the effects on validity and subgroup differences of adding additional predictors. *Sackett and Ellingson (1997)* offered formulas for estimating the effect of adding additional predictors on the resulting composite subgroup difference. Other research has explored the issue with differing combinations of predictors (e.g., Pulakos & Schmitt, 1996; Ryan et al., 1998).

Dahlke and Sackett (2021) examined a meta-analytic matrix—initially assembled by *Roth* *et al. (2011)* and refined by Song et al. (2017)—containing validity, White-Black subgroup differences, and correlations among five predictors: cognitive ability, conscientiousness, biodata, structured interviews, and integrity tests. With regression weighting, mean validity increased consistently as the number of predictors increased from 1 to 5. This finding supports the notion that capturing more criterion-relevant attributes contributes to higher validity if predictors are appropriately weighted. However, mean subgroup differences also increased as the number of predictors increased. Adding predictors can either increase or decrease mean differences. If one starts with a predictor with a large difference, adding small-difference predictors will result in a smaller composite difference. In contrast, if one starts with a small-difference predictor, adding a larger-difference predictor will result in a larger composite difference.

The intercorrelation among predictors also helps determine the effects of additional predictors on mean differences (*Sackett & Ellingson, 1998)*. For example, adding a small-difference predictor to an existing large-difference predictor will reduce the difference on the composite to a greater degree as the correlation between the two predictors increases. Moreover, Dahlke and Sackett (2021) showed that the mean level of predictive bias tends to decrease as the number of predictors increases. Predictive bias reflects the correspondence between subgroup differences on the predictor composite and on the criterion. For example, with an overall performance criterion, adding more predictors that provide incremental validity leads to closer correspondence between predictor and criterion, and hence less bias.

Once a multi-predictor system is identified, *LeBreton et al. (2007)* demonstrated how relative weights analysis can be used to shed light on the relative role of each predictor. Finally, using multiple predictors can also increase time and costs. Thus, decisions about how many predictors to use should consider tradeoffs between these features and validity, group differences, and predictive bias.

Add structure. The strategy of adding structure became popular in the context of employment interviews to increase reliability, validity, and fairness (e.g., *Campion et al., 1988; Campion et al., 1997; Pulakos & Schmitt, 1995*). Campion et al. (1997) defined structure as "any enhancement of the interview that is intended to increase psychometric properties by increasing standardization or otherwise assisting the interviewer in determining what questions to ask or how to evaluate responses" (p. 656). They identified 15 components of structure categorized by the content of the interview (e.g., base questions on job analysis, ask each applicant the same questions) and the evaluation of interviewees' responses (e.g., use anchored rating scales, use multiple interviewers). According to *Levashina et al. (2014*), 12 meta-analyses found that structure increased the criterion-related validity of interviews. In addition, structure reduces idiosyncratic interviewer effects (e.g., demographic similarity, *McCarthy et al., 2010*), produces higher reliability (e.g., Huffcutt et al., 2013), and smaller subgroup differences (e.g., Huffcutt & Roth, 1998), which is partly explained by its lower cognitive load (Berry et al., 2007). That said, structure has beneficial effects to a level where validity asymptotes (e.g., Huffcutt & Arthur, 1994). Similarly, although more consistency leads to higher procedural fairness (*Hausknecht et al., 2004*), too much consistency lowers interactional justice (e.g., too much structure is seen as "cold"; Conway & Peneno, 1999).

Although the notion of structure is more than 20 years old, it is still very relevant. Technology (and the COVID-19 pandemic) has put synchronous video interviews (e.g., via ZOOM) and asynchronous video interviews (also known as "on demand" interviews) at center stage. In asynchronous video interviews, applicants receive a standard set of questions on their screen (e.g., through an avatar) they need to answer within a predefined time. Their answers (captured via webcam and microphone) are later evaluated by recruiters or AI; thus, asynchronous video interviews are noninteractive and typically include the question standardization element of structured interviews (Lukacik et al., 2020). Results about asynchronous interviews mirror many of the trends above for highly structured interviews. Although participants appreciate the consistency, flexibility, and preparation time of these interviews (Basch et al., 2021; Langer et al., 2017), a study in 46 countries found that applicants rated asynchronous video interviews as less effective and satisfying than synchronous virtual interviews (Griswold et al., 2021). Other studies (e.g., Acikgoz et al., 2020; Langer et al., 2017; Newman et al., 2020) revealed that participants perceived asynchronous interviews lower on interpersonal treatment, privacy, controllability, social presence, and media richness. In turn, this often lowered an organization's attractiveness. Interestingly, asynchronous interviews also impaired the impression management found in traditional interviews and led to shorter answers (e.g., Langer et al., 2020).

The positive effects of structure also have been found in ACs. For instance, *Reilly et al. (1990*) reported better construct-related validity evidence when behavioral checklists were used. In addition, drawing on trait activation theory (Tett & Burnett, 2003), Lievens et al. (2015) discovered that assessors who were familiarized with standardized role-player prompts noted more behavioral observations and provided ratings with higher validity and accuracy. Cybervetting (i.e., inspecting people's social media information) is another domain where structure may be beneficial, although initial evidence has been mixed (*Roulin & Levashina, 2019*; Zhang et al., 2020). Finally, structure also is important in the integration of different pieces of information into an overall evaluation. Indeed, research consistently has found that algorithmic integration of information produces better predictive validity than judgmental integration (e.g., Kuncel et al., 2013).

Contextualize procedure content. Contextualization refers to the extent to which test stimuli are embedded in a detailed and realistic context (Lievens & Sackett, 2017). Generally, contextualization aims to enhance validity (via increasing the overlap with the criterion) and applicant perceptions (via job relevance; *Hausknecht et al., 2004*). Contextualization has been especially tried out with personality tests. The history of personality tests is characterized by pendulum swings, as attested by the more pessimistic view in the 60s (e.g., *Guion & Gottier, 1965*), the more positive picture in the 90s based on meta-analyses on their criterion-related validity (e.g., *Barrick & Mount, 1991; Tett et al., 1991*), and in more recent years the skepticism of many journal editors (*Morgeson et al., 2007*). Currently, there is relative consensus that efforts should be undertaken to improve personality measurement in selection. Contextualization fits into these efforts because the general frame-of-reference evoked by noncontextualized personality items (e.g., "I pay attention to details") might lead to ambiguity and measurement error, thereby lowering validity. Contextualizing aims to alleviate these problems by asking applicants to adopt a frame-of-reference (e.g., "I pay attention to details *at work*"). A meta-analysis confirmed that adding such at-work tags increased mean validities from .11 to .24 across the Big Five personality traits (*Shaffer & Postlethwaite, 2012*). Contextualized personality scores also can provide incremental validity over generic ones (e.g., Bing et al., 2004). Research suggest that the superior validity of contextualized personality test is not only due to higher reliability but also to better predictor-criterion matching (Lievens et al., 2008).

Although contextualization can enhance validity, there is limited evidence in terms of its effects on subgroup differences and applicant reactions. For example, making cognitive ability items more contextualized via business-related graphs (Hattrup et al., 1992) or placing them in a social relations context (DeShon et al., 1998) did not lower ethnic subgroup differences. In one of the few studies to examine applicant reactions, Holtz et al. (2005) found nonsignificant differences between students' perceptions of contextualized versus generic personality measures.

Whereas the studies above attempted to increase the contextualization of selection procedures, another research stream addressed the opposite: Is context necessary in selection procedures like SJTs? Krumm et al. (2015) demonstrated that a substantial proportion of SJT items (between 43 to 71%) could be answered correctly when the job-related situations were omitted in the item stems. Follow-up studies showed that this finding remained stable across different SJTs, constructs, samples, presentation formats, response instructions, and item characteristics (e.g., Schäpers, Lievens et al., 2020). Although excluding the situation descriptions did not affect construct-related validity or decrease applicant perceptions, criterion-related validity effects seemed to depend on the breadth of the criteria (Schäpers, Mussel et al., 2020): SJT validity was not affected for predicting broad criteria (i.e., overall job performance) but it was for specific criteria (e.g., interpersonal adaptability). Another study designed an SJT with items from various occupations and found that the SJT predicted behavior in contexts that were not even featured in the items (Motowidlo et al., 2016). Overall, this research suggests many SJT items tap into general domain knowledge (i.e., implicit trait policies; Lievens & Motowidlo, 2016) and detailed contexts are less important than originally thought. This sheds a different angle on recent efforts to make the stimulus material in selection as realistic as possible.

Minimize cognitive load. The cognitive load of a predictor reflects how strongly scores covary with cognitive ability. As discussed, cognitive load is the primary driver of racio-ethnic subgroup differences (Dahlke & Sackett, 2017). A crucial distinction, however, is between construct-relevant and construct-irrelevant cognitive load. For example, job knowledge is near the cognitive end of a cognitive-non cognitive continuum, and well-designed knowledge measures can be expected to have a substantial construct-relevant cognitive load. In contrast, a SJT that targets proactivity (e.g., Bledow & Frese, 2009) would be closer to the non-cognitive end of the continuum and, thus, likely would have a substantially smaller cognitive load.

Lievens and Sackett (2017) developed a modular framework of method factors to consider when designing predictors that can influence outcomes such as cognitive load. These factors include stimulus format (e.g., written vs. video), response format (e.g., multiple choice vs. constructed response), contextualization (e.g., context-neutral vs. context specific), etc. Although much remains to be learned about choices between method factors, work to date offers various useful suggestions. For example, in the domain of SJTs, audiovisual stimulus presentation tends to produce higher validity (*Christian et al., 2010*) and lower cognitive saturation (Lievens & Sackett, 2006) than written stimulus presentation. And for a job knowledge test, a written constructed response format produced smaller subgroup differences and comparable criterion-related validity as a multiple-choice format (Edwards & Arthur, 2007).

Chan and Schmitt (1997) offered an illustration of the role of inadvertent cognitive load. They transcribed the content of a video-based interpersonal skills SJT and found a much larger White-Black mean difference in the written condition (d = .95) than in the video condition (d = .21). A written reading comprehension test correlated with the written SJT, but not the video SJT, shedding light on the mechanism behind the findings. Similarly, Lievens et al. (2019) administered a video SJT with three different response formats: traditional multiple-choice, an open-ended written format, and an open-ended spoken (i.e., webcam-based) format. Non-native speakers (e.g., immigrants) were compared with natives. Subgroup differences paralleled Chan and Schmitt: largest in the multiple-choice condition (d = .91) and smaller in constructed response conditions (*d* = .41 and .09 in the open-ended written and spoken conditions, respectively). Importantly, criterion-related validity was comparable across conditions.

Thus, the issue of cognitive load needs to be carefully evaluated in each specific setting. For example, inadvertent load may have a large effect when the range of reading ability in the applicant pool is large (e.g., entry-level jobs), but may have little to no effect when reading ability is generally high (e.g., applicant pools of college graduates). The effects of inadvertent cognitive load also may depend on the nature of the criterion. For example, cognitive load might harm validity when the criterion is non-cognitive (e.g., citizenship behavior) but increase validity if the criterion itself has a high cognitive load (e.g., performance on technical tasks).

Review procedure content. Research concerning the fairness and legal defensibility of selection procedures has tended to focus on whether overall scores on the procedures produce adverse impact or differential prediction. We review two approaches for ensuring the *items* that make up the procedures are free from bias. In a sensitivity review, panels of subject matter experts and/or test developers examine test content to identify items that might prevent applicants from responding in ways that allow for valid inferences regarding the target constructs (Zieky, 2006). Examples of problematic items include those that contain offensive language, emotionally provocative content, portrayal of stereotypes, or contextual information unfamiliar to applicants from certain backgrounds (Golubovich et al., 2014).

Although sensitivity reviews are widely used for evaluating standardized tests such as the ACT (Education Testing Service, 2009) and GMAT (Rudner, 2012), we know very little regarding whether or how organizations conduct such reviews or their effectiveness. Golubovich et al. (2014) surveyed professionals who served as sensitivity reviewers and found that most reviewers did not receive formal training. In addition, item reviewers reported that they encounter relatively few problematic items, and when they do, items tend to the more subtly problematic, such as including content or vocabulary that might be differentially familiar to applicants. Golubovich et al. also asked the reviewers to evaluate a set of problematic items and discovered relatively modest interrater consistency regarding perceptions of item sensitivity. A few studies have examined individual differences that could influence the accuracy with which reviewers identify problematic items. One somewhat consistent factor appears to be sex, whereby female reviewers are more likely to perceive items as problematic compared to male reviewers (Grand et al., 2013; Mael et al., 1996). However, Grand et al. found that items female reviewers flagged as problematic did not function differently based on respondent sex.

In contrast to sensitivity reviews, differential item functioning (DIF) refers to a set of statistical approaches that can identify problematic items. DIF exists when individuals with equal standings on the total score or latent trait (e.g., verbal ability) have different item responses based on group membership (e.g., male vs. female applicants; Drasgow, 1984). DIF has been used to examine differences between applicants from different racial-ethnic groups (e.g., Stark et al., 2004) and cultures (e.g., Meade, 2010), as well as between groups with or without incentive to fake (e.g., Stark et al., 2001). For instance, Chan et al. (1999) examined sex- and race-based DIF on the ASVAB and found higher rates of DIF among items that assessed more semantic content than among items that focused on general skills and principles. Stark et al. (2001) found evidence of DIF in a personality test, such that some items reflected different constructs in applicant versus non-applicant samples.

Further, DIF often varies in direction, with some items favoring one group and other items favoring another. As such, differential functioning at the measure level (i.e., differential test functioning, or DTF) may be negligible because the item-level differences cancel out (e.g., Stark et al., 2004). Stark et al. offered a useful formulation of overall differences between groups in observed scores as the sum of DTF and true differences between groups. Using this approach, they concluded that racio-ethnic differences on a licensing exam and the ACT reflected true differences rather than bias. Other research has found that DTF can be mitigated by removing one or more problematic items (e.g., Chan et al., 1999). Because large samples typically are needed to examine DIF, observed effects can be statistically significant but often are so small they may not be practically significant. Thus, recent studies have provided formulas and programs for estimating the magnitude of DIF effects (e.g., Meade, 2010; Nye & Drasgow, 2011). Finally, it is important to try to understand the underlying reason(s) for any DIF that may emerge. For example, Whitney and Schmitt (1997) found only limited evidence that Black-White DIF on a biodata measure was due to cultural value differences between the two groups.

Limit opportunities to fake. Researchers have long been interested in faking on selection procedures, which reflects “an intentional response behavior aimed at exerting a positive influence on the hiring decision” (Griffeth et al., 2011, p. 345). Most research has studied self-reports of noncognitive constructs, which is the focus of our review (see Melchers et al. [2020] for a review of interview faking research). Early research investigated whether people could fake their responses and found that when instructed or given an incentive to do so, people tend to score higher on personality tests (e.g., *Dunnette et al., 1962*), biodata inventories (e.g., *Schrader & Osborn, 1977*), and vocational interest measures (e.g., *Abrahams et al., 1971*). Integrity tests also are susceptible to faking (e.g., Alliger & Dwight, 2000).

Second, research has examined actual faking during the selection process. Between-person studies comparing scores of applicants who completed personality tests for selection to scores of incumbents who completed these tests for research found consistently higher scores among applicants (e.g., Hough, 1998). Within-person studies compare scores obtained in different circumstances. Ellingson et al. (2007) found that mean score increases on a personality test generally were small but were largest among those who first completed the test for development (i.e., no incentive to fake) and then for selection (i.e., where there was an incentive to fake). However, the personality test (California Personality Inventory, Gough & Bradley, 1996) uses a dichotomous response format, which may have limited the magnitude of potential differences (see Hogan et al. [2007] for a similar issue). Griffeth et al. (2007) found larger differences between personality test scores obtained during the selection process and a follow-up test in which applicants were instructed to respond honestly.

A third issue relates to the consequences of faking for selection decisions. Faking can affect construct-related validity by producing an "ideal employee" factor in applicant settings (e.g., Schmit & Ryan, 1993). As to whether faking affects criterion-related validity, studies in which incumbents were instructed (or determined) to respond honestly or fake tended to find higher correlations with job performance (e.g., McCartney et al., 1962) or other relevant variables (e.g., *Pannone, 1984*) in the honest group. Faking also can affect rank-order changes, which, in turn, can affect which applicants are selected (e.g., Mueller-Hanson et al., 2003). This is particularly concerning given fakers may comprise most of the candidates hired under low selection ratios (e.g., Rosse et al., 1998).

Finally, research on ways to reduce or control for faking can be distinguished in terms of reactive versus preventive approaches (Fan et al., 2012). Reactive approaches attempt to address faking post hoc. For instance, using social desirability scales to adjust scores for faking tends to have little or no effect on outcomes such as criterion-related validity (e.g., *Christiansen et al., 1994*; Ones et al., 1996). In studies examining item response patterns, people instructed to fake often adopt an extreme response set (i.e., a disproportionate favor for scale endpoints), have quicker response times, and exhibit less eye fixation (e.g., van Hooft & Born, 2012). Including bogus items or those that allow applicants to exaggerate their knowledge (i.e., overclaiming) can detect faking and even predict subsequent deviant behavior (e.g., Dunlop et al., 2020).

Preventive approaches attempt to prevent faking from occurring in the first place. Forced-choice (FC) measures and applicant warnings are commonly studied. FC-formatted measures require respondents to select among statements that assess different characteristics but are similarly attractive. Although research suggests reduced faking with FC measures, the specific format appears to matter. For example, faking is lower when FC measures match statements on social desirability or job relevance, ask applicants to pick the statement most like them, and use normative scoring (Cao & Drasgow, 2019). Further, normative scoring of FC measures is required to make between-person comparison required for selection. There also is evidence that FC measures capture somewhat different constructs than single-stimulus measures as suggested by stronger correlations with cognitive ability (e.g., Vasilopoulos et al., 2006).

Another preventive approach involves requiring applicants to provide information to support their answers. Several studies have found that such elaboration tends to reduce scores on selection procedures (e.g., *Levishina et al., 2012*; Lievens & Peeters, 2008; *Schmitt & Kunce, 2002*). For example, Levishina et al. asked job applicants to elaborate their responses to certain items on a biodata measure and found that doing so yielded lower scores compared to items for which applicants were not asked to elaborate. However, item verifiability had a larger effect, such that scores were lower on items that could be verified and higher on items that could not be easily verified. Further, these two factors interacted, such that required elaboration was somewhat more likely to reduce scores of non-verifiable items compared to verifiable ones.

Finally, warning applicants not to fake generally produces lower scores (e.g., Dwight & Donovan, 2003). A "test warning-retest" approach whereby applicants who respond affirmatively to items designed to detect faking are warned and given an opportunity to change their answers or retake the assessment shows promising initial results (e.g., Fan et al., 2012). However, this approach requires organizations to develop valid measures of faking and determine cut scores for flagging fakers. Other potential concerns are that warnings can suppress the responses of honest applicants (Kuncel & Borneman, 2007), influence construct-related validity, such as by increasing variance related to cognitive ability (e.g., Vasilopoulos et al., 2005), and produce negative emotions and fairness perceptions (e.g., Li et al., 2022).

In sum, research suggests that applicants can and do fake, which, in turn, can affect which applicants are selected. Further, faking can negatively influence the construct- and criterion-related validity of inferences based on selection procedures. Steps such as using FC measures, warnings, and elaborations can reduce faking, but they also can change the meaning of test scores. Overall, faking continues to be a vexing issue in employee selection.

Consider gamifying selection procedures. Attempts to optimize validity, diversity, and the applicant experience are well-illustrated by the growing popularity of gamifying selection procedures and game-based assessment (GBA). Gamification involves adding game like elements (e.g., action language, conflict/challenge, game fiction, immersion, rules/goals, Bedwell et al., 2012) to existing selection procedures such as personality tests (e.g., Landers & Collmus, 2022) or SJTs (e.g., Georgiou et al., 2019). Gamification is not the same as GBA. Whereas gamification denotes an activity of an assessment developer to (re)design existing selection procedures, GBA represents a novel method for assessing known constructs. Another key distinction is made between "theory driven" GBA—in which a GBA is designed with the intent of measuring a specific construct (e.g., Landers et al., 2022) —and "data driven" GBA—in which an empirical or other AI approach is used to build a predictive model of outcomes or existing construct measures directly (e.g., Wu et al., 2022).

Gamifying selection procedures and GBA have the potential to increase fidelity, immersion, motivation, flow, and reduce anxiety, thus improving the applicant experience. Key questions are whether they also reduce faking, decrease subgroup differences, and increase criterion-related validity, while maintaining the same construct-related validity. Landers et al. (2021) examined many of these issues in a GBA of cognitive ability. First, they observed applicant experience benefits mostly in terms of improved intrinsic motivation (1.00 *SD* higher than traditional assessments) for the GBA. For other applicant perceptions, improvements were more modest, varying from 0.13 to 0.30 (see also results for gamifying SJTs, Georgiou & Lievens, 2022; Georgiou & Nikolaou, 2020). Second, they found that the true-score correlation between the GBA and a traditional cognitive ability test battery was .97 (see also Quiroga et al., 2019). However, the validity of the operational composite (.16) for predicting grade point average was lower than the one (.22) of traditional tests. In addition, the GBA had no incremental validity over the traditional assessments, whereas these assessments did predict over the GBA. Although the validity results are based on a small sample (*N* = 49), this initial evidence suggests that GBA's increased fidelity and immersion do not lead to higher validity and might even introduce measurement error (see Arthur et al., 2017; Bhatia & Ryan, 2018). Other research (Wu et al., 2022) also found that a GBA (in this case designed to assess conscientiousness) risks that unintended constructs (i.e., cognitive ability) are measured.

A third key conclusion related to GBAs' potential to reduce subgroup differences, Landers et al. (2022) found that race differences were similar across the GBA and traditional assessments. However, females performed worse on most of the GBAs. Importantly, similar to traditional assessments, there was no evidence for differential prediction.

It is important to end with two caveats. First, given GBA is an alternative method for assessing known constructs, it is difficult to derive general conclusions about their effects on validity and subgroup differences *regardless of the construct* measured. As another testament of this, Landers and Collmus (2022) found that a personality inventory that included narrative elements ("storification") was resistant to faking for assessing conscientiousness but not for openness. Second, gamification and GBAs are technical products whose *design quality* can vary greatly. As such, conclusions regarding the efficacy of these methods depend on the design options chosen (Landers & Marin, 2021).

Administration Decisions

Consider pre-test explanations, practice tests, and coaching. This section focuses on steps organizations can take to help applicants prepare for selection procedures. Perhaps the most basic step is to provide applicants information regarding what the procedures will entail. Research on the effects of pre-test information has been mixed. For example, Truxillo et al. (2002) found that providing information about how procedures are related to the job and the feedback applicants will receive was positively to applicant perceptions of these factors but did not relate to other outcomes, such as pursuit intentions or selection outcomes. Lievens et al. (2003) reported that providing reliability and validity information did not increase fairness perceptions. Burns et al. (2008) found that whether applicants were provided an information packet about the selection procedures was unrelated to performance on the procedures and applicant reactions. However, receiving the packet was associated with more positive reactions among applicants who failed the test battery.

A more substantial step is to provide applicants opportunities to practice the selection procedures. Practice tests are like retesting but may involve different dynamics because practice tests are not “for keeps.” In addition, providing practice tests is less expensive for the organization than is providing opportunities to retest. In one of the few studies on this topic, Campion et al. (2019) found that applicants who scored higher on a practice test were more likely to apply for the job. Also, applicants who took the practice test performed better on the actual test than applicants who did not take it, though self-selection for taking the practice test also could contribute to this effect. Further, score gains between the practice and actual tests were larger for Black and Hispanic applicants than for White applicants.

The most involved step is to provide selection procedure orientation or coaching programs. Orientation or test preparation programs are short in duration and introduce applicants to the types of selection procedures they will take, as well as provide test-taking strategies (Sackett et al., 2001). *Ryan et al. (1998)* found that whether applicants attended a test preparation program was unrelated to how they performed on a cognitive ability test. Further, although female and Black applicants were more likely to attend the program, attendance did not reduce racio-ethnic differences on the test. In contrast, Chung-Herrera et al. (2009) found that test preparation program attendance was positively correlated with scores on a job knowledge test, though there were no Black-White differences in program attendance. Like Ryan et al., test preparation did not appear to improve test performance more for one subgroup than for the other.

Coaching programs are more extensive and typically involve practice and feedback (Sackett et al., 2001). In an experimental study, *Kurecka et al. (1982)* found that coaching that provided behavioral examples of good performance and practice with feedback yielded higher scores on a leaderless group discussion than did no coaching. In a meta-analysis of coaching effects on cognitive ability tests, Hausknecht et al. (2007) reported larger mean score improvements for coached than for uncoached groups of (*d* = .70 vs. .24, respectively). However, these estimates do not account for other factors that could play a role. For example, based on Hasuknecht et al.'s regression results, we estimated a much smaller effect (*d* = .09) for a coaching program of average length, in an operational setting, and with retesting on an alternate form. Mauer and colleagues (1998, 2001) found that interview coaching was positively related to subsequent interview performance, and Lievens et al. (2012) discovered that medical school applicants who paid for formal coaching scored higher on an SJT.

Thus, most research has focused on whether test preparation influences how applicants perform on selection procedures. Conversely, few studies have examined the effects on validity. Maurer et al. (2008) found that a situational interview predicted job performance only among applicants who participated in a voluntary coaching program. Internal consistency and interrater reliabilities also were higher in the coached group. Stemig et al. (2015) discovered that the availability of coaching did not degrade the predictive validity of an SJT. Additional research is needed to determine whether the higher scores test preparation programs tend to produce reflect changes in performance-relevant variance (e.g., increases in the target KSAOs), irrelevant variance (e.g., test wiseness), or extraneous variance (e.g., test unfamiliarity or anxiety).

Consider revealing target KSAOs to applicants. The issue whether to make the constructs assessed in selection procedures transparent to applicants is another key administration decision. Generally, transparency improves fairness perceptions (e.g., opportunity to perform, *Ingold et al., 2016*), applicant performance (e.g., Jacksch & Klehe, 2016), and construct measurement (e.g., by reducing error variance, Kleinmann et al., 1996). Nevertheless, transparency might improve applicant performance only for constructs on which subgroups are not negatively stereotyped. Jacksch and Klehe (2016) found that transparency improved both male and female performance on a gender-neutral dimension (planning), but it decreased female scores on a gender stereotyped dimension (leadership). In addition, Ingold et al. (2016) found that transparency lowered the criterion-related validity of AC scores because it made the situation stronger, thereby suggesting to applicants what they should do, rather than allow them to choose what to do. This explanation of transparency removing construct-relevant variance is consistent with evidence on the validity of applicants' *spontaneous* inferences about constructs assessed (i.e., their ability to identify criteria; Jansen et al., 2013).

Rockstuhl and Lievens (2021) delved deeper into the effects on criterion-related validity by cuing the performance criteria associated with the constructs assessed (i.e., prompts related to effective behavior). Candidates who received performance cues relied more on their knowledge to solve intercultural scenarios and improved predictions of in-role performance (an outcome often predicted by cognitive constructs). Conversely, candidates who did not receive cues regarding the criteria relied more on their behavioral tendencies (perspective taking and openness) and this increased the prediction of intercultural performance (an outcome often predicted by personality). This study thus shows that revealing the performance criteria does not lower validity per se. Instead, conceptually different constructs (knowledge vs. personality) are activated and different outcomes are predicted depending on how prompts are specified.

Recently, the transparency question has grown in importance in light of AI and game-based assessment because it is often unclear for stakeholders (e.g., applicants, HR users, general public) what these assessments measure and how the algorithms work (Langer & Landers, 2021). In this research, a distinction has been made between process information (i.e., making transparent what is measured and what will happen) and process justification (i.e., telling why the procedure is used). Langer et al. (2021) found that process justification demonstrated the largest effect on applicant perceptions (see also Newman et al., 2020).

Consider retesting. Many organizations allow unsuccessful applicants to retake selection procedures. Doing so can expand applicant pools, which allows organizations to be more discerning about whom they select and, in turn, increase the utility of selection procedures. *Lievens et al. (2005)* introduced a framework that outlines the retest effects that can occur, how to test the different effects, and their practical implications. Subsequent research has examined various aspects of retesting and has revealed several trends (Van Iddekinge & Arnold, 2017).

First, applicant scores on selection procedures tend to increase upon retesting. Score increases tend to be small for cognitive-oriented tests (e.g., Hausknecht et al., 2007), small to moderate for performance-based assessments such as interviews and ACs (e.g., Schleicher et al., 2010), and moderate to large for job knowledge tests (e.g., *Raymond et al., 2007)* and personality tests (e.g., Landers et al., 2011). Consistent with these findings, Schleicher et al.'s (2010) research on several selection procedures revealed larger score gains for more novel, less g-loaded, performance-based, and more fakeable procedures.

Research also has identified person and situational factors that can affect how much retesters improve. Among person factors, score improvements tend to be larger among high-ability, White, female, and younger test-takers (e.g., Rapport et al., 1997; Schleicher et al. 2010). Regarding situational factors, score improvements are typically larger when tests are completed for selection than for research or development purposes, particularly personality tests (e.g., Ellingson et al., 2007). In addition, whether test-takers pass the initial test (but are not selected) or fail the initial test also seems to matter. For instance, *Hausknecht (2010)* found that scores of applicants who initially passed a battery of personality and cognitive tests, but who were not selected, were not significantly different from their retest scores. In contrast, retest scores of applicants who initially failed the test battery were significantly higher on six of eight personality test scales. Similarly, Holladay et al. (2013) reported that applicants who were told they were not selected due to their personality test score improved more upon retesting than applicants who were not provided this feedback. Finally, regression to the mean also plays a role. Van Iddekinge and Arnold (2017) reported that such effects accounted for between 9% and 55% of score improvements in the studies they reviewed.

Another consistent finding is that between-person variance tends to be larger in retest scores than in initial test scores (e.g., Raymond et al., 2007; Schleicher et al., 2010; Van Iddekinge et al., 2011). This suggests the existence of individual differences in how test-takers interpret, react to, and attempt to address their initial test performance. The tendency for retest scores to be more varied is particularly interesting given that retest scores often are higher, which generally decreases score variance (e.g., a ceiling effect).

Perhaps relatedly, retest scores also tend to be more reliable than initial test scores, including estimates of internal consistency reliability (e.g., Hogan et al., 2007) and interrater reliability (e.g., Schleicher et al., 2010). In addition, estimates of criterion-related validity tend to be larger for retest scores. For instance, *Lievens et al. (2005)* compared initial and retest scores as predictors of grades among medical school students who retested after initial failure. Validity coefficients for a composite of three predictors were .21 and .38 for initial and retest scores, respectively. However, one-time applicants' scores on the test composite were a significantly better predictor of GPA than were repeat applicants' scores. Van Iddekinge et al. (2011) found that retest scores on a job knowledge test were significantly better predictors of job performance (.43) than were initial scores (.31).

Although retest scores tend to be higher, more varied, more reliable, and better predictors of performance than initial test scores, possible downsides to retesting also exist. For example, retest effects on personality tests may reflect faking, and we do not know how this may affect the criterion-validity of repeat scores on such tests. Further, test scores of one-time applicants may be more valid than retest scores (e.g., Lievens et al., 2005). If so, retesting could lower validity overall. Finally, as noted, applicants from some minority groups (e.g., Blacks and Hispanics, older applicants) might not benefit as much from retesting as majority group members.

Scoring Decisions

Weight predictors and criteria. Once one has settled on a set of predictors as components of a selection system, one faces the decision of how to combine the predictor scores. A key decision is whether to use the predictors in a compensatory or non-compensatory manner. Often this is a values-driven decision: one organization planning to use measures of, say, cognitive ability and conscientiousness may opt for compensatory use, such that a higher level of one attribute can make up for a lower level of the other. In contrast, another organization may be unwilling to make this tradeoff and instead set separate cutoff scores on each predictor. In some circumstances, a minimum amount of the attribute(s) a predictor assesses may be a job requirement, rather than a values choice, as in the case of a job requiring the regular lifting of objects up to 50 pounds.

When a compensatory strategy is used, one faces the decision of the appropriate weights to assign each predictor when combining them into a composite. We outline four approaches. The first is regression weighting whereby weights are chosen that maximize the validity of the composite in the validation sample. The second is unit weighting in which all predictors equally weighted. This has pragmatic appeal in terms of being easier to apply and explain. In addition, with smaller sample sizes, unit weighting can produce better results upon cross-validation than regression weighting (Bobko et al., 2007). The third approach involves weighting schemes such as Pareto-optimization that attempt to optimize multiple criteria rather than validity alone (e.g., DeCorte et al., 2011; Rupp et al., 2020). For example, these approaches can be used to identify weights that give the highest level of diversity at any given level of validity, and vice versa. This permits an organization to see, for example, the level of improvement in diversity attainable at a specific loss in validity and to use this information in choosing among sets of weights. The final approach relies on job analytic information, rather the relationship between predictors and criteria, as the basis for predictor weighting. Here analyst ratings of the relative importance of various attributes underlying the set of predictors determine the relative weighting of predictors (e.g., Goldstein et al., 1993).

The selection literature is replete with studies documenting the incremental validity of one predictor over another, with the findings commonly used to advocate selecting on a composite that adds a new predictor to an existing one. Sackett et al. (2017) showed that these findings are specific to the use of regression weights, and that validity often decreases, rather than increases, when a new predictor is added using unit weights rather than regression weights. Thus, weighting decisions should be made with care.

Finally, although the literature has paid considerable attention to predictor weights, criterion weights also merit consideration. For example, Rotundo and Sackett (2002) showed that supervisors often weight task, citizenship, and counterproductive components differently when making overall performance judgments. These weighting choices can subsequently affect predictor weights (*Murphy & Shiarella, 1997*). Task performance has been shown to be more cognitively loaded than citizenship and counterproductive behavior (Gonzalez-Mulé et al., 2014). Thus, if task performance is assigned more weight in an overall performance measure, regression weighting would, in turn, assign greater weight to a cognitive-oriented predictor and less weight to non-cognitive predictors. Thus, criterion weighting is a values choice on the part of the organization: which aspects of the criterion domain are of greatest concern? In sum, the major takeaways from this work are (a) one should choose a weighting approach consistent with valued objectives (e.g., maximize validity vs. balance validity and diversity) and (b) consider the relative importance of the criterion to be predicted, which, in turn, can affect how much weight each selection procedure should be given.

Consider artificial intelligence. Although scoring and mechanical aggregation have a long tradition in selection (Kuncel et al., 2013), AI has put all of this in a broader perspective. AI is an "umbrella term for a wide array of models, methods, and prescriptions used to simulate human intelligence, often when it comes to collecting, processing, and acting on data" (Köchling & Wehner, 2020, p. 798). Some AI domains are computer vision (image recognition), machine learning[1], and language processing (Kaplan & Haenlein, 2019; Paschen et al., 2020).

As AI can work with a variety of inputs such as written text, audio, and video, it provides opportunities to model selection data that were previously thought to be too large, noisy, and unstructured. For example, Naim et al. (2015) analyzed facial expressions, language, and prosodic information (e.g., pitch) of interviewees and were able to predict interview performance rated by humans with *r*s of .70 or higher. Chen et al. (2017) found similarly high accuracy for AI models taught to predict people's personality and hiring recommendations based on

[1] In machine learning, algorithms model the relationship between inputs and outputs on the basis of training data. When the trained model is applied to new data, the model predictions are compared to known outputs and convergence speaks to the model's predictive accuracy (Tay et al., 2022). In supervised machine learning selection applications, human ratings (e.g., self-reports, interviewer ratings) typically serve as "ground truth", whereas there is no ground truth in unsupervised machine learning.

asynchronous interviews rated by trained experts. Spoken text was the most useful for the predictions, whereas facial and prosodic information was not. Hickman et al. (2022) also trained models to predict trait ratings in interviews. There was stronger evidence for convergent and discriminant validity when the models were trained on interviewer ratings than on self-report ratings. They also examined the generalizability of their models to new interview questions and found adequate test-retest reliabilities only for extraversion and conscientiousness. Yet, convergent and discriminant validity evidence generalized to new interviews. The models (trained based on interviewer ratings) also predicted academic outcomes similarly as self- and other-rated personality.

Other studies used social media data as inputs to infer personality traits. Park et al. (2015) trained models to predict self-rated personality based on Facebook texts. In another sample, these models correlated with self-reports and informant reports, predicted incremental variance over other-reports, were stable over 6-month intervals, and correlated with external criteria (e.g., life satisfaction) to a similar degree as personality scales. Kosinski et al. (2013) used Facebook “likes” to successfully predict variables such as personality, intelligence, and political views. Two meta-analyses concluded that the convergence between AI predictions based on social media and self-reports ranged from .29 (agreeableness) to .40 (extraversion; Azucar et al., 2018) and was a bit higher for AI than for a manual (e.g., via recruiters) approach (Tay et al., 2020).

Finally, Campion et al. (2016) reported one of the first AI studies based on selection data. AI scoring of accomplishment records correlated over .60 with human ratings and demonstrated comparable evidence of reliability and construct-related validity. Subgroup differences between Whites and minorities were low for both AI and human scores. Yet, substantial savings could be made via the AI approach.

Although these studies showcase the possibilities of AI, they are also limited due to their focus on the convergence between AI and human scoring and their lack of job-related criteria such as performance or turnover. In addition, the theoretical rationale for including various inputs in the predictive model is often scant. As an example of a more theory-driven approach, Sajjadiani et al. (2019) built models using application data (e.g., work experience, turnover history) and considered underlying mechanisms so that the model features were explainable to stakeholders (e.g., hiring managers, applicants). For instance, they categorized turnover reasons as reflecting either an "avoidance" (indicated by a history of leaving bad jobs) or an "approach" disposition (indicated by a history of approaching better jobs, see Elliot & Thrash, 2002). Work experience relevance and approaching better jobs predicted job performance and were negatively related to voluntary turnover, whereas leaving bad jobs negatively predicted performance and was positively linked to involuntary turnover. This study demonstrates the benefit of combining data-driven methods with theory-based derivation of possible pivotal features.

Consider score banding. Banding involves grouping scores from selection procedures within given ranges and treating scores within a band as equivalent. Many uses of banding are based on a desire to avoid overinterpreting small differences (which may represent nothing but measurement error) or to simplify score reporting and use. Common examples include A-F grading in academic settings and the use of stanine scores that convert scores to a 0 to 9 metric. Banding received increased attention after it was proposed to increase diversity (e.g., Casio et al., 1991). This form of banding relied on the standard error of measurement as the basis for defining bandwidth, such that scores not significantly different from the highest score are placed in the same band and viewed as interchangeable. Banding can increase diversity if additional minorities are included in the top band(s) of candidates from which selection occurs.

Sackett and Wilk (1994) identified three key features on which banding approaches differ. The first is whether there is minority preference within a band. If all applicants in a band are viewed as equivalent, an organizational goal of diversity might be made for giving minority preference within the band. The second is the basis for selection within a band, with options including selection on some additional predictor (e.g., experience, seniority) and random selection. The third is the use of fixed bands (where lower bands come into play only when the top band is exhausted) versus sliding bands (where the band is recalculated once the highest scorer in the band is selected). Research suggests that diversity benefits are largest with minority preference and sliding bands (Cascio et al., 1991; *Murphy et al., 1995*; Sackett & Roth, 1991).

There has been much debate regarding logical, statistical, and philosophical issues surrounding banding (e.g., *Campion et al., 2001*; Schmidt, 1991; Zedeck et al., 1991). The legality of banding also is somewhat uncertain. Henle (2004) reviewed the legal status of banding and concluded there does not appear to be a clear consensus. Similarly, Barrett and Lueke (2004) found over 30 cases that mentioned banding. In some cases, banding was proposed as potential approach to addressing disparities, whereas in other cases the fairness of banding was challenged by nonminority or minority plaintiffs. However, use of banding with minority preference as a sole basis for selection within a band has not survived legal scrutiny (Henle, 2004). Use of minority status as a “plus factor” along with other information has received support, but absent the dominant use of minority status as the basis for selection within a band, banding has little effect. Selection decisions within a band can be based on factors that do not show group differences, such as seniority, experience, or non-cognitive predictors (*Campion et al., 2001*). Yet, Laczo and Sackett (2004) showed that if one has an additional predictor with a small subgroup difference, it is more useful to include it as part of the selection system than to save it for use as tool for selection within a band. In addition, variables such as seniority and experience tend to have weak criterion-related validity (e.g., Ng & Feldman, 2010; Van Iddekinge et al., 2019), so using such variables to select within a band could be suboptimal.

In recent years, there has been very little new research on banding, nor has there been new legal or professional guidance regarding their use. Further, the most recent revision to the *The Principles* (2018) says very little about banding beyond that it generally will yield lower expected criterion outcomes and utility, but it may increase administrative ease, as well as diversity depending on how within-band selections are made (p. 32). A few studies have examined formulaic issues with creating bands (e.g., Bobko et al., 2005; Gasperson et al., 2013). Interestingly, there has been little or no field research that directly examines the extent to which banding affects the criterion-related validity of selection procedures.

Key Directions for Future Research

Table 2 provides a list of future research directions in each area addressed in our review. In addition, we conclude by describing several areas we believe are in particular need of additional research attention.

Bias and Opacity in AI

The sections above documented the rapid growth of AI use in selection. Although there was often an impressive convergence between AI based algorithms and human raters, there are still challenges to tackle. Critically, more field-based research on the validity and adverse impact of AI based scores in actual selection settings is needed. Subgroup differences in the context of AI algorithms deserve particular attention (Oswald et al., 2020). Algorithmic bias can occur from (a) skewed input data (i.e., some groups are overrepresented), (b) human biases in the input data (hence, algorithms replicate the bias), (c) constraints in the algorithm (i.e., technical bias), and

(d) population changes after constructing the algorithm (i.e., emergent bias, Köchling & Wehner, 2020; Sanchez-Monedero et al., 2020). Unfortunately, little is known about how AI vendors deal with these potential biases. Raghavan et al. (2020) concluded that most vendors remove model features that are correlated with protected attributes. Although the 4/5ths rule is typically used as benchmark for such "debiasing", differential prediction should also be a core concern. Along these lines, Tay et al. (2022) developed an approach to test for differential functioning of the trained ML model between subgroups. More broadly, Landers and Behrend (2022) proposed to conduct psychological audits, which are impartially conducted evaluations of AI systems based on components such as input data, model design, interpretations by stakeholders and cultural context.

Another challenge deals with the opacity (i.e., lack of transparency) associated with AI solutions. On one hand, AI opacity might result in selection procedures that are more difficult to "game" by applicants. On the other hand, opacity might result in a lack of trust and controllability on behalf of applicants as well as hiring managers (e.g., Li et al., 2021). Langer and König (2021) listed three reasons for the opacity of AI solutions (see also Burrell, 2016: Mahmud et al., 2022): (a) the system is too complex to explain, (b) technical stakeholder illiteracy, and (c) intentional decisions of developers to keep the system opaque. They also sketched the implications of opacity for five stakeholder groups (users, affected people, deployers, developers, and regulators; see also Langer & Landers, 2021) and provided several strategies (technical solutions, education and training, and regulation and guidelines) for reducing the "black box" associated with AI. We encourage more field studies that compare different decision-making modalities (e.g., AI, human, a combination, see Langer et al. 2020) and examine whether it makes sense to adjust design characteristics to "humanize" AI (Kaplan et al., 2021; Landers & Marin, 2019; Roesler et al., 2021). Whereas such anthropomorphisms might increase trust, people might also expect AI to be AI (i.e., to be consistent instead of human).

Selection Across Cultures

Another issue in need of additional research concerns cross-cultural issues in selection. Differences across countries and cultures in how selection systems are designed and evaluated come into play in various ways (Ryan & Tippins, 2009). First, the legal environment for selection varies widely across countries, with differing groups offered protection (Shen et al, 2017). At the extreme, practices prohibited in some countries may be required in others. For example, setting quotas for racio-ethnic groups is prohibited in the U.S. but is required in South Africa (Myors et al., 2008). Thus, multinational firms need to attend to the legal requirements of each setting in which they operate.

Second, cultural values can affect the use of different selection procedures. *Ryan et al. (1999)* and Ryan et al. (2017) found considerable variability in selection practices across countries. For example, Ryan et al. (1999) found that although structured interviews are widely advocated in the U.S., resistance to them is substantial in some other countries. However, both studies reported a lack of success in predicting cross-cultural differences in selection using cultural values dimensions (e.g., performance orientation, uncertainty avoidance, cultural tightness). Thus, this remains an area for future research.

Third, whether research findings, such as evidence of criterion-related validity of different predictors, generalize across countries and cultures remains relatively unexplored. Examples of such evidence for subsets of countries can be found in meta-analyses of the relationship between cognitive ability test and job performance, with U.S.-specific evidence (e.g., Schmidt & Hunter, 1998) complemented by a meta-analysis in the European community

(*Salgado et al., 2003)* and another in Great Britain (Bertua et al., 2005). In this instance, similar findings were observed across settings.

Fourth, firms attempting to use the same selection procedures across countries need to attend to issues of measurement equivalence. For instance, great care needs to be taken in translating test items from one language to another. Even in cases where language is common, test content can be specific to one country (e.g., using math problems involving U.S. currency). There are examples of developing measures for global use, rather than retrofitting a measure designed for use in a different setting, such as *Schmit et al.'s (2000)* global personality inventory.

Selection for Gig and Telework Work

The term "gig" comes from musicians who are paid a set fee to perform one show (Kolmar, 2022). Gig work refers to similar short-term arrangements between an employer and a contractor who is hired to serve one function (Ashford et al., 2018; Kolmar, 2002). Examples of gig work include freelance writing, rideshare drivers (e.g., Uber), grocery shopping for others, tutoring, and performing household tasks like cleaning someone's home.

Despite the increased prevalence of gig work, very little is known about selection for such jobs. Ashford et al. (2018) identified several ways in which gig work may differ from traditional work, including less stability and financial security, greater autonomy, less certain career paths, greater transience, and greater physical and relational separation. Given these differences, gig jobs may require additional or different KSAOs, such as resilience, independence, learning agility, and tolerance for ambiguity (Ashford et al., 2018). Furthermore, very little is known about how organizations select people for gig work. Because such jobs tend to involve lower-level and shorter-term work, perhaps organizations use less rigorous selection procedures. For example, some organizations may hire just about any applicant who meets basic qualifications (e.g., at least 18 years of age, acceptable driving record, clean drug test). Another issue is how organizations perceive applicants for traditional jobs who also hold a gig job (i.e., multiple jobholders, Campion & Csillag, 2022). Gig workers are more likely to be racial minorities (particularly Hispanic and Black individuals), as well as from lower income levels (Pew Research Center, 2021). Thus, potential adverse impact or biases associated with demographic and other socioeconomic factors could affect selection decisions involving multiple job holding applicants.

Telework also has increased in recent years, particularly in the wake of the COVID-19 pandemic. As with gig work, remote work may include different tasks or challenges than onsite work. For example, working from home tends to involve more nonwork intrusions, distractions, and multitasking, particularly for female workers (Leroy et al., 2021). Therefore, remote work may require different KSAOs and, thus, different or additional selection procedures. Attributes including independence, attentional focus, and conscientiousness facets such as dutifulness, organization, and initiative, may be particularly important to success in telework jobs.

Another key issue is how organizations assess the performance of gig and remote workers. In traditional work, there are already issues that supervisors often have limited opportunity to observe employees (MacLane et al., 2020). This seems to be exacerbated in gig and remote work. Because many employees plan to remain working remotely (Pew Research Center, 2022), these issues will continue to be relevant.

Conclusion

The study of personnel selection is one of the great success stories of applied psychology, and over the past 75 years, *Personnel Psychology* has been one of the main outlets for this research. Collectively, this work has yielded tremendous insights about how organizations can improve how they select employees and enhance their human capital resources. However, assessing job applicants is challenging, and decisions made based on that information have considerable stakes for organizations, applicants, and society. And although much progress has been made about how to make effective and fair selection decisions, vexing issues persist. Examples include how to simultaneously maximize workforce performance and diversity, how to best assess personality and other noncognitive constructs, and how to design structured selection procedures applicants and decision makers perceive as fair and reasonable to administer. In addition, new questions and challenges continue to emerge, such as the role of gamification and AI in the design and scoring of selection procedures. We hope the present review helped take stock of what we know and still need to know about this important topic.

References

Abrahams, N. M., Neumann, I., & Githens, W. A. (1971). Faking vocational interests: Simulated versus real life motivation. *Personnel Psychology, 24*, 5-12.

Acikgoz, Y., Davison, K. H., Compagnone, M., & Laske, M. (2020). Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment, 28*, 399–416.

Alliger, G. M., & Dwight, S. A. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. *Educational and Psychological Measurement*, 60(1), 59-72. doi:10.1177/00131640021970367

Arthur Jr, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*(1), 125-153.

Arthur Jr, W., Doverspike, D., Kinney, T. B., & O'Connell, M. (2017). The impact of emerging technologies on selection models and research: Mobile devices and gamification as exemplars. In *Handbook of employee selection* (pp. 967-986). Routledge.

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology, 35*(2), 281-322.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*, 695-716. https://doi.org/10.1111/j.1744-6570.1990.tb00679.x

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27*(4), 519–533. https://doi.org/10.1111/j.1744-6570.1974.tb01173.x

Ashford, S. J., Caza, B. B., & Reid, E. M. (2018). From surviving to thriving in the gig economy: A research agenda for individuals in the new world of work. *Research in Organizational Behavior*, *38*, 23-41.

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences, 124*, 150-159.

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: Legal, technical, and societal issues* (pp. 71-111). Quorum Books.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31*, 233–241.

Barrick, M. R., & Mount, M. K. (1991). The Big 5 personality dimensions and job-performance - a meta-analysis. *Personnel Psychology, 44*, 1-26.

Barrick, M. R., Mount, M. K., & Gupta, R. (2003). Meta-analysis of the relationship between the five-factor model of personality and Holland's occupational types. *Personnel Psychology, 56*(1), 45-74.

Basch, J. M., Brenner, F., Melchers, K. G., Krumm, S., Dräger, L., Herzer, H., & Schuwerk, E. (2021). A good thing takes time: The role of preparation time in asynchronous video interviews. *International Journal of Selection and Assessment, 29(3/4)*, 378-392.

Bauer, T.N., Truxillo, D.M., Sanchez, R.J., Craig, J.M., Ferrara, P., & Campion, M.A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology*, *54*(2), 387-419. https://doi.org/10.1111/j.1744-6570.2001.tb00097.x

Bedwell, W. L., Pavlas, D., Heyne, K., Lazzara, E. H., & Salas, E. (2012). Toward a taxonomy linking game attributes to learning: An empirical study. *Simulation & Gaming: An Interdisciplinary Journal, 43*(6), 729–760. doi:10.1177/1046878112439444

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 435-463.

Berry, C. M., Sackett, P. R., & Landers, R. N. (2007). Revisiting interview-cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology, 60*, 837-874.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78*(3), 387-409.

Bhatia, S., & Ryan, A. M. (2018). Hiring for the win: Game-based assessment in employee selection. In J. H. Dulebohn & D. L. Stone (Eds.), *The brave new world of eHRM 2.0* (pp. 81–110). IAP Information Age Publishing.

Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental validity of the frame-of-reference effect in personality scale scores: A replication and extension. *Journal of Applied Psychology, 89*, 150-157.

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology, 62*(2), 229-258.

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*(4), 689–709. https://doi.org/10.1177/1094428106294734

Bobko, P., Roth, P. L., & Nicewander, A. (2005). Banding selection scores in human resource management decisions: Current inaccuracies and the effect of conditional standard errors. *Organizational Research Methods, 8*(3), 259-273.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology*, *52*(3), 561-589.

Bolanovich, D. J. (1948). Reduce factory turnover. *Personnel Psychology, 1*(1), 81-92. https://doi.org/10.1111/j.1744-6570.1948.tb01296.x

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48*(3), 587-605.

Boudreau, J. W. (1983). Economic considerations in estimating the utility of human resource productivity improvement programs. *Personnel Psychology, 36*(3), 551-576.

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology, 2*, 171–183. https://doi.org/10.1111/j.1744-6570.1949.tb01397.x

Burns, G. N., Siers, B. P., & Christiansen, N. D. (2008). Effects of providing pre-test information and preparation materials on applicant reactions to selection procedures. *International Journal of Selection and Assessment, 16***,** 73-77. https://doi.org/10.1111/j.1468-2389.2008.00411.x

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 2053951715622512.

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology, 43*(2), 313-575.

Campion, E. D., & Csillag, B. (2022). Multiple jobholding motivations and experiences: A typology and latent profile analysis. *Journal of Applied Psychology*, *107*(8), 1261-1287.

Campion, M. C., Campion, E. D., & Campion, M. A. (2019). Using practice employment tests to improve recruitment and personnel selection outcomes for organizations and job seekers. *Journal of Applied Psychology, 104*(9), 1089–1102. https://doi.org/10.1037/apl0000401

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101*(7), 958–975.

Campion, M. A., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). Doing competencies well: Best practices in competency modeling. *Personnel Psychology, 64*(1), 225-262.

Campion, M. A., Outtz, J. L., Zedeck, S., Schmidt, F. L., Kehoe, J. F., Murphy, K. R., & Guion, R. M. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology, 54*(1), 149-185.

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50*, 655–702.

Campion, M. A., Pursell, E. D., & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41*(1), 25-42.

Cann, A., Siegfried, W. D., & Pearce, L. (1981). Forced attention to specific applicant qualifications: Impact on physical attractiveness and sex of applicant biases. *Personnel Psychology, 34*, 65-75. https://doi.org/10.1111/j.1744-6570.1981.tb02178.x

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, *104*(11), 1347.

Cascio, W. F. (1982). *Costing human resources: The financial impact of behavior in organizations*. Kent.

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*(4), 233-264.

Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology, 84*, 610-619.

Chan, D., & Schmitt, N. (1997). Video-based 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*(1), 143-159.

Chen, L., Zhao, R., Leong, C. W., Lehman, B., Feng, G., & Hoque, M. E. (2017). Automated video interview judgment on a large-sized corpus collected online. In 2017 *Seventh International Conference on Affective Computing and Intelligent Interaction* (ACII), 504–509.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*(1), 83-117.

Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47*(4), 847-860.

Chung-Herrera, B. G., Ehrhart, K. H., Ehrhart, M. G., Solamon, J., & Kilian, B. (2009). Can test preparation help to reduce the black—white test performance gap? *Journal of Management, 35*(5), 1207-1227. doi:10.1177/0149206308328506

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*(2), 115-124.

Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology, 104*(10), 1243–1265. https://doi.org/10.1037/apl0000406

Combs, J., Liu, Y., Hall, A., & Ketchen, D. (2006). How much do high-performance work practices matter? A meta-analysis of their effects on organizational performance. *Personnel Psychology*, 59(3), 501-528. https://doi.org/10.1111/j.1744-6570.2006.00045.x

Conway, J. M., & Peneno, G. M. (1999). Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology, 13*, 485-506.

Cortina, J. M., Doherty, M. L., Schmitt, N., Kaufman, G., & Smith, R. G. (1992). The "big five" personality factors in the IPI and MMPI: Predictors of police performance. *Personnel Psychology, 45*, 119-140. https://doi.org/10.1111/j.1744-6570.1992.tb00847.x

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology, 53*(2), 325-351.

Cronshaw, S. F. (1997). Lo! The stimulus speaks: The insiders view on Whyte and Latham's "The futility of utility analysis". *Personnel Psychology*, *50*(3), 611-615.

Dahlke, J.A., & Sackett, P. R. (2017). The relationship between cognitive-ability saturation and subgroup differences across predictors of job performance. *Journal of Applied Psychology, 102,* 1403-1420.

Dahlke, J. A., & Sackett, P. R. (2021). On the assessment of predictive bias in selection systems with multiple predictors. *Journal of Applied Psychology.* Advance online publication.

DeCorte, W., Sackett, P. R., & Lievens, F. (2011). Designing Pareto-optimal selection systems: Formalizing the decisions required for selection system development. *Journal of Applied Psychology, 96,* 907-920.

DeNisi, A. S., & Murphy, K. R. (2017). Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology, 102*(3), 421–433. https://doi.org/10.1037/apl0000085

De Shon, R. P., Smith, M. R., Chan, D., & Schmitt, N. (1998). Can racial differences in cognitive test performance be reduced by presenting problems in a social context? *Journal of Applied Psychology, 83*, 438-451.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95*(1), 134–135. https://doi.org/10.1037/0033-2909.95.1.134

Dunlop, P. D., Bourdage, J. S., de Vries, R. E., McNeill, I. M., Jorritsma, K., Orchard, M., Austen, T., Baines, T., & Choe, W.-K. (2020). Liar! Liar! (when stakes are higher): Understanding how the overclaiming technique can be used to measure faking in personnel selection. *Journal of Applied Psychology, 105*(8), 784-799. https://doi.org/10.1037/apl0000463

Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology, 15*, 13-24.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*(1), 1-23.

Educational Testing Service. (2009). *ETS guidelines for fairness review of assessments*. Publisher.

Edwards, B. D., & Arthur, W., Jr. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology, 92*, 794–801.

Ellingson, J. E., Sackett, P. R., & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology, 92*(2), 386–395. https://doi.org/10.1037/0021-9010.92.2.386

Elliot, A. J., & Thrash, T. M. (2002). Approach-avoidance motivation in personality: Approach and avoidance temperaments and goals. *Journal of Personality and Social Psychology, 82*, 804–818.

Equal Employment Opportunity Commission C. S. C, Department of Labor, and Department of Justice. (1978). *Uniform guidelines on employee selection procedures*, Federal Register, 43 (166), 38295–38309.

Fan, J., Gao, D., Carroll, S. A., Lopez, F. J., Tian, T. S., & Meng, H. (2012). Testing the efficacy of a new procedure for reducing faking on personality tests within selection contexts. *Journal of Applied Psychology, 97*(4), 866–880. https://doi.org/10.1037/a0026655

Gasperson, S. M., Bowler, M. C., Wuensch, K. L., & Bowler, J. L. (2013). A statistical correction to 20 years of banding. *International Journal of Selection and Assessment*, *21*(1), 46-56.

Georgiou, K., & Lievens, F. (2022). Gamifying an assessment method: what signals are organizations sending to applicants? *Journal of Managerial Psychology*. Advance online publication.

Georgiou, K., & Nikolaou, I. (2020). Are applicants in favor of traditional or gamified assessment methods? Exploring applicant reactions towards a gamified selection method. *Computers in Human Behavior, 109*, 106356.

Gerhart, B., Wright, P. M., McMahan, G. C., & Snell, S. A. (2000). Measurement error in research on human resources and firm performance: How much error is there and how does it influence effect size estimates? *Personnel Psychology, 53*, 803-834.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26*(4), 461–477. https://doi.org/10.1111/j.1744-6570.1973.tb01150.x

Gilliland, S. W. (1993). The perceived fairness of selection systems: an organizational justice perspective. *Academy of Management Review, 18*, 694–734. https://doi.org/10.5465/amr.1993.9402210155

Gilliland, S. W. (1994). Effects of procedural and distributive justice on reactions to a selection system. *Journal of Applied Psychology, 79*(5), 691–701. https://doi.org/10.1037/0021-9010.79.5.691

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

Golubovich, J., Grand, J. A., Ryan, A. M., & Schmitt, N. (2014). Sensitivity review practices in test development. *International Journal of Selection and Assessment, 22*, 1-11. https://doi.org/10.1111/ijsa.12052

Gonzalez-Mulé, E., Mount, M. K., & Oh, I. S. (2014). A meta-analysis of the relationship between general mental ability and nontask performance. *Journal of Applied Psychology, 99*(6), 1222-1243.

Gough, H. G., & Bradley, P. (1996). *The California Psychological Inventory manual* (3rd ed.). Consulting Psychologists Press.

Grand, J. A., Golubovich, J., Ryan, A. M., & Schmitt, N. (2013). The detection and influence of problematic item content in ability tests: An examination of sensitivity review practices for personnel selection test development. *Organizational Behavior and Human Decision Processes, 121*(2), 158-173.

Griffith, R .L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior, *Personnel Review*, 36(3), 341-355. https://doi.org/10.1108/00483480710731310

Griffith, R. L., Lee, L. M., Peterson, M. H., & Zickar, M. J. (2011). First dates and little white lies: A trait contract classification theory of applicant faking behavior. *Human Performance, 24*(4), 338–357.

Griswold, K. R., Phillips, J. M., Kim, M. S., Mondragon, N., Liff, J., & Stanley, G. M. (2021). Global differences in applicant reactions to virtual interview synchronicity. *The International Journal of Human Resource Management*, 1-28.

Guion, R. M. (1965). *Personnel testing*. McGraw-Hill.

Guion, R. M. (1978). "Content validity" in moderation. *Personnel Psychology*, *31*(2), 205-213.

Guion, R. M., & Cranny, C. J. (1982). A note on concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology, 67*(2), 239–244. https://doi.org/10.1037/0021-9010.67.2.239

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18*, 135–164.

Hardy, J. H., Gibson, C., Carr, A., & Dudley, N. (2021). Quitters would not prosper: Examining the relationship between online assessment performance and assessment attrition behavior. *International Journal of Selection and Assessment*, *29*(1), 55-64.

Hardy, J. H., Gibson, C., Sloan, M., & Carr, A. (2017). Are applicants more likely to quit longer assessments? Examining the effect of assessment length on applicant attrition behavior. *Journal of Applied Psychology, 102*(7), 1148–1158.

Harold, C. M., Holtz, B. C., Griepentrog, B. K., Brewer, L. M., & Marsh, S. M. (2016). Investigating the effects of applicant justice perceptions on job offer acceptance. *Personnel Psychology, 69*(1), 199-227. https://doi.org/10.1111/peps.12101

Hartigan, J., & Wigdor, A. K. (1989). *Fairness in employment testing*. National Academies Press.

Harvey, R. J. (1991). Job analysis. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 71-163). Consulting Psychologists Press.

Hatch, N. W., & Dyer, J. H. (2004). Human capital and learning as a source of sustainable competitive advantage. *Strategic Management Journal, 25*(12), 1155–1178. https://doi.org/10.1002/smj.421

Hattrup, K., Schmitt, N., & Landis, R. S. (1992). Equivalence of constructs measured by job-specific and commercially-available aptitude tests. *Journal of Applied Psychology, 77*, 298-308.

Hausknecht, J. P. (2010). Candidate persistence and personality test practice effects: Implications for staffing system management. *Personnel Psychology*, *63*(2), 299-324. https://doi.org/10.1111/j.1744-6570.2010.01171.x

Hausknecht, J.P., Day, D.V., & Thomas, S.C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639-683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

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. http://dx.doi.org/10.1037/0021-9010.92.2.373

Henle, C. A. (2004). Case review of the legal status of banding. *Human Performance*, 17, 415–432.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology, 107*(8), 1323–1351. https://doi.org/10.1037/apl0000695

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*(5), 1270-1285.

Holtz, B. C., Ployhart, R. E., & Dominguez, A. (2005). Testing the rules of justice. The effects of frame-of-reference and pre-test information on personality test responses and test perceptions. *International Journal of Selection and Assessment, 13*, 75-86.

Holladay, C. L., David, E., & Johnson, S. K. (2013). Retesting personality in employee selection: Implications of the context, sample, and setting. *Psychological Reports, 112*(2), 486-501.

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*(2-3), 209-244.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*(1-2), 152-194.

Huffcutt, A. I., & Arthur Jr, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*, 184-190.

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment, 21*, 264-276.

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in interview evaluations. *Journal* *of Applied Psychology, 83*, 179-189.

Huselid, M. A. (1995). The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of Management Journal, 38*(3), 635–672. https://doi.org/10.5465/256741

Ingold, P. V., Kleinmann, M., König, C. J., & Melchers, K. G. (2016). Transparency of assessment centers: Lower criterion-related validity but greater opportunity to perform? *Personnel Psychology, 69*, 467-497.

Jacksch, V., & Klehe, U. C. (2016). Unintended consequences of transparency during personnel selection: Benefitting some candidates, but harming others? *International Journal of Selection and Assessment, 24*(1), 4-13.

Jansen, A., Melchers, K. G., Lievens, F., Kleinmann, K., Brändli, M., Fraefel, L., & König, C. J. (2013). Situation assessment as an ignored factor in the behavioral consistency paradigm underlying the validity of personnel selection procedures. *Journal of Applied Psychology, 98*, 326-341.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62 (1*), 15-25.

Kerr, W. A. (1948). For precision workers at RCA. *Personnel Psychology, 1*(1), 63-66. https://doi.org/10.1111/j.1744-6570.1948.tb01294.x

Kim, Y., & Ployhart, R. E. (2014). The effects of staffing and training on firm productivity and profit growth before, during, and after the Great Recession. *Journal of Applied Psychology, 99*(3), 361–389. https://doi.org/10.1037/a0035408

Kim, Y., & Ployhart, R. E. (2018). The strategic value of selection practices: antecedents and consequences of firm-level selection practice usage. *Academy of Management Journal, 61*(1), 46-66.

Kleinmann, M., Kuptsch, C., & Köller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centers. *Applied Psychology: An International Review, 45*, 67–84.

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research, 13*, 795-848.

Kolmar, C. (2022, February). *23 essential gig economy statistics [2022]: definitions, facts, and trends on gig work*. Retrieved from https://www.zippia.com/advice/gig-economy-statistics/

Konradt, U., Garbers, Y., Böge, M., Erdogan, B., & Bauer, T. N. (2017). Antecedents and consequences of fairness perceptions in personnel selection: A 3-year longitudinal study. *Group & Organization Management, 42*(1), 113-146.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 5802–5805.

Kristof, A. L. (1996). Person-organization fit: An integrative review of its conceptualizations, measurement, and implications. *Personnel Psychology, 49*(1), 1-49.

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *Journal of Applied Psychology, 100,* 399-416.

Kuncel, N. R., & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment, 15*(2), 220-231.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: a meta-analysis. *Journal of Applied Psychology*, 98(6), 1060–1072. https://doi.org/10.1037/a0034156

Kurecka, P. M., Austin, J. M., Jr., Johnson, W. & Mendoza, J. L. (1982). Full and errant coaching effects on assigned role leaderless group discussion performance. *Personnel Psychology, 35*, 805-812. https://doi.org/10.1111/j.1744-6570.1982.tb02223.x

Laczo, R. M., & Sackett, P. R. (2004). Further Monte Carlo investigation of the effects of banding on performance and minority representation. In H. Aguinis (Ed.), *Test score banding in human resource selection: Legal, technical, and societal issues* (pp. 133-150). Quorum Books.

Landers, R. N., Armstrong, M. B., Collmus, A. B., Mujcic, S., & Blaik, J. (2021). Theory-driven game-based assessment of general cognitive ability: Design theory, measurement, prediction of performance, and test fairness. *Journal of Applied Psychology*. Advance online publication.

Landers, R. N., & Behrend, T. S. (2022). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*. Advance online publication. https://doi.org/10.1037/amp0000972

Landers, R. N., & Collmus, A. B. (2022). Gamifying a personality measure by converting it into a story: Convergence, incremental prediction, faking, and reactions. *International Journal of Selection and Assessment, 30*(1), 145-156.

Landers, R. N., & Marin. S. (2019). Theory and technology in organizational psychology: A review of technology integration paradigms and their effects on the validity of theory. *Annual Review of Organizational Psychology and Organizational Behavior, 8*, 235-258.

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*(1), 202-210.

Langer, M., & König, C. J. (2021). Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management. *Human Resource Management Review*. Advance online publication.

Langer, M., König, C. J., & Busch, V. (2021). Changing the means of managerial work: Effects of automated decision support systems on personnel selection tasks. *Journal of Business and Psychology, 36*(5), 751-769.

Langer, M., König, C. J., & Hemsing, V. (2020). Is anybody listening? The impact of automatically evaluated job interviews on impression management and applicant reactions. *Journal of Managerial Psychology, 35*(4), 271-284.

Langer, M., König, C. J., & Krause, K. (2017). Examining digital interviews for personnel selection: Applicant reactions and interviewer ratings. *International Journal of Selection and Assessment, 25*, 371–382.

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior.* Advance online publication.

Latham, G. P., & Whyte, G. (1994). The futility of utility analysis. *Personnel Psychology*, *47*(1), 31-46.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563-575.

LeBreton, J. M., Hargis, M. B., Griepentrog, B., Oswald, F. L., & Ployhart, R. E. (2007). A multidimensional approach for evaluating variables in organizational research and practice. *Personnel Psychology, 60*(2), 475-498.

Leroy, S., Schmidt, A. M., & Madjar, N. (2021). Working from home during COVID-19: A study of the interruption landscape. *Journal of Applied Psychology, 106*(10), 1448–65.

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology, 67*, 241-293.

Levashina, J., Morgeson, F. P., & Campion, M. A. (2012). Tell me some more: Exploring how verbal ability and item verifiability influence responses to biodata questions in a high-stakes selection context. *Personnel Psychology, 65*(2), 359-383.

Li, H., Fan, J., Zhao, G., Wang, M., Zheng, L., Meng, H., ... & Lievens, F. (2022). The role of emotions as mechanisms of mid-test warning messages during personality testing: A field experiment. *Journal of Applied Psychology*, *107*(1), 40-59.

Li, L., Lassiter, T., Oh, J., & Lee, M. K. (2021). *Algorithmic hiring in practice: Recruiter and HR professional's perspectives on AI use in hiring*. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58*(4), 981-1007.

Lievens, F., Buyse, T., Sackett, P. R., & Connelly, B. S. (2012). Coaching effects on situational judgment tests. *International Journal of Selection and Assessment, 20*, 272-282. https://doi.org/10.1111/j.1468-2389.2012.00599.x

Lievens, F., De Corte, W., & Brysse, K. (2003). Applicant perceptions of selection procedures: The role of selection information, belief in tests, and comparative anxiety. *International Journal of Selection and Assessment, 11*, 67-77. https://doi.org/10.1111/1468-2389.00227

Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology, 93*, 268-279.

Lievens, F., & Peeters, H. (2008). Impact of elaboration on responding to situational judgment test items. International *Journal of Selection and Assessment, 16*(4), 345-355.

Lievens, F., & Sackett, P. R. (2006). The predictive validity of video-based and written situational judgment tests in an operational setting. *Journal of Applied Psychology, 91,* 1181-1188.

Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology, 102,* 43-66.

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. E., & De Soete, B. (2019). Constructed responses formats and their effect on majority-minority differences and validity. *Journal of Applied Psychology, 104,* 715-726.

Lievens, F., Sanchez, J. I., & De Corte, W. (2004). Easing the inferential leap in competency modeling: The effects of task-related information and subject matter expertise. *Personnel Psychology, 57*, 881-904.

Lievens, F., Schollaert, E., & Keen, G. (2015). The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in assessment center exercises. *Journal of Applied Psychology, 100*, 1169-1188.

Lukacik, E.-R., Bourdage, J. S., & Roulin, N. (2020). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*. Advance online publication.

Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology*, *47*(4), 715-738. https://doi.org/10.1111/j.1744-6570.1994.tb01573.x

MacLane, C. N., Cucina, J. M., Busciglio, H. H., & Su, C. (2020). Supervisory opportunity to observe moderates criterion-related validity estimates. *International Journal of Selection and Assessment, 28*(1), 55-67.

Mael, F. A. (1991). A conceptual rationale for the domain and attributes of biodata items. *Personnel Psychology, 44*(4), 763-792.

Mael, F. A., Connerley, M., & Morath, R. A. (1996). None of your business: Parameters of biodata invasiveness. *Personnel Psychology, 49*(3), 613-650.

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change, 175*, 121390.

Maurer, T. J., Solamon, J. M., Andrews, K. D., & Troxtel, D. D. (2001). Interviewee coaching, preparation strategies, and response strategies in relation to performance in situational employment interviews: An extension of Maurer, Solamon, and Troxtel (1998). *Journal of Applied Psychology, 86*(4), 709–717. https://doi.org/10.1037/0021-9010.86.4.709

Maurer, T.J., Solamon, J. M., & Lippstreu, M. (2008). How does coaching interviewees affect the validity of a structured interview? *Journal of Organizational Behavior, 29*, 355-371. https://doi.org/10.1002/job.512

Maurer, T., Solamon, J., & Troxtel, D. (1998). Relationship of coaching with performance in situational employment interviews. *Journal of Applied Psychology, 83*(1), 128–136. https://doi.org/10.1037/0021-9010.83.1.128

McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., & Ahmed, S. M. (2017). Applicant perspectives during selection: A review addressing "So what?,""What's new?," and "Where to next?" *Journal of Management, 43*(6), 1693-1725.

McCarthy, J. M., Hrabluik, C., & Jelley, R. B. (2009). Progression through the ranks: Assessing employee reactions to high-stakes employment testing. *Personnel Psychology, 62*, 793-832. https://doi.org/10.1111/j.1744-6570.2009.01158.x

McCarthy, J. M., & Goffin, R. (2004). Measuring job interview anxiety: Beyond weak knees and sweaty palms. *Personnel Psychology*, 57, 607-637. https://doi.org/10.1111/j.1744-

6570.2004.00002.x

McCarthy, J. M., Van Iddekinge, C. H., & Campion, M. A. (2010). Are highly structured job interviews resistant to demographic similarity effects? *Personnel Psychology, 63*, 325-359.

McCarthy, J. M., Van Iddekinge, C. H., Lievens, F., Kung, M.-C., Sinar, E. F., & Campion, M. A. (2013). Do candidate reactions relate to job performance or affect criterion-related validity? A multistudy investigation of relations among reactions, selection test scores, and job performance. *Journal of Applied Psychology, 98*(5), 701–719. https://doi.org/10.1037/a0034089

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb III, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63-91.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43*, 335-354. https://doi.org/10.1111/j.1744-6570.1990.tb01562.x

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(4), 728–743. https://doi.org/10.1037/a0018966

Melchers, K. G., Roulin, N., & Buehl, A. K. (2020). A review of applicant faking in selection interviews. *International Journal of Selection and Assessment, 28*(2), 123-142.

Miles, S. J., & McCamey, R. (2018). The candidate experience: Is it damaging your employer brand? *Business Horizons, 61*, 755–764.https://doi.org/10.1016/j.bushor.2018.05.007

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

Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance, 29*(4), 331-346.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88,* 348–355. doi:10.1037/0021-9010.88.2.348

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*(4), 873-900.

Murphy, K. R., Osten, K., & Myors, B. (1995). Modeling the effects of banding in personnel selection. *Personnel Psychology*, *48*, 61–84.

Murphy, K. R., & Shiarella, A. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology, 50*, 823– 854. https://doi.org/10.1111/j.1744-6570.1997.tb01484.

Myors, B., Lievens, F., Schollaert, E., Van Hoye, G., Cronshaw, S. F., Mladinic, A., ... & Sackett, P. R. (2008). International perspectives on the legal environment for selection. *Industrial and Organizational Psychology, 1*(2), 206-246.

Naim, I., Tanveer, I., Gildea, D., & Hoque, M. E. (2018). Automated analysis and prediction of job interview performance. IEEE *Transactions on Affective Computing, 9*(2), 191–204.

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes, 160*, 149–167.

Ng, T. W., & Feldman, D. C. (2009). How broadly does education contribute to job performance? *Personnel Psychology, 62*(1), 89-134.

Ng, T. W., & Feldman, D. C. (2010). Organizational tenure and job performance. *Journal of Management, 36*(5), 1220-1250.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96*, 966-980.

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

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*(6), 660–679. https://doi.org/10.1037/0021-9010.81.6.660

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resource management: forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior, 7*(1), 505–533.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89*(2), 187-207.

Pannone, R. D. (1984). Predicting test performance: A content valid approach to screening applicants. *Personnel Psychology, 37*(3), 507-514.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology, 108*(6), 934–952.

Paschen, U., Pitt, C., & Kietzmann, J. (2020). Artificial intelligence: Building blocks and an innovation typology. *Business Horizons 63*(2), 147-155.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y., Campion, M. A., Mayfield, M. S., Morgeson, F. P., Pearlman, K., Gowing, M. K., Lancaster, A. R., Silver, M. B., & Dye, D. M. (2001). Understanding work using the Occupational Information Network (O* NET): Implications for practice and research. *Personnel Psychology, 54*(2), 451-492.

Pew Research Center (December, 2021). *The state of gig work in 2021*. Author.

Pew Research Center (February, 2022). *COVID-19 pandemic continues to reshape work in America*. Author.

Ployhart, R. E., & Hakel, M. D. (1998). The substantive nature of performance variability: Predicting interindividual differences in intraindividual performance. *Personnel Psychology, 51*(4), 859-901.

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*(1), 153-172.

Ployhart, R. E., Lim, B. C., & Chan, K. Y. (2001). Exploring relations between typical and maximum performance ratings and the five factor model of personality. *Personnel Psychology, 54*(4), 809-843.

Ployhart, R. E., & Moliterno, T. P. (2011). Emergence of the human capital resource: A multilevel model. *Academy of Management Review, 36*(1), 127-150.

Ployhart, R. E., Van Iddekinge, C. H., & MacKenzie Jr, W. I. (2011). Acquiring and developing human capital in service contexts: The interconnectedness of human capital resources. *Academy of Management Journal, 54*(2), 353-368.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, 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. https://doi.org/10.1111/j.1744-6570.2003.tb00757.x

Pulakos, E. D., & Schmitt, N. (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology, 48*(2), 289-308.

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9*, 241-258.

Quińones, M. A., Ford, J. K., & Teachout, M. S. (1995). The relationship between work experience and job performance: A conceptual and meta-analytic review. *Personnel Psychology, 48*(4), 887-910.

Quiroga, M. A., Diaz, A., Román, F. J., Privado, J., & Colom, R. (2019). Intelligence and video games: Beyond "brain-games." *Intelligence, 75*, 85–94. https://doi.org/10.1016/j.intell.2019.05.001

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic employment screening: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM* (pp. 469–481).

Rapport, L. J., Brines, D. B., Theisen, M. E., & Axelrod, B. N. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *The Clinical Neuropsychologist, 11*(4), 375-380.

Raymark, P. H., Schmit, M. J., & Guion, R. M. (1997). Identifying potentially useful personality constructs for employee selection. *Personnel Psychology, 50*(3), 723-736.

Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology, 60*(2), 367-396.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1-62. https://doi.org/10.1111/j.1744-6570.1982.tb02184.x

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43*(1), 71–84.

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology, 58*(1), 103-139.

Robinson, D. D. (1981). Content-oriented personnel selection in a small business setting. *Personnel Psychology*, *34*(1), 77-87.

Rockstuhl, T., & Lievens, F. (2021). Prompt-specificity in scenario-based assessments: associations with personality vs. knowledge and effects on predictive validity. *Journal of Applied Psychology, 106*, 122-139.

Roesler, E., Manzey, D., Onnasch, L. (2021). A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics, 6*(58). https://doi.org/10.1126/scirobotics.abj5425.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*(4), 634–644. https://doi.org/10.1037/0021-9010.83.4.634

Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58*(4), 1009-1037.

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.

Roth, P. L., Van Iddekinge, C. H., DeOrtentiis, P. S., Hackney, K. J., Zhang, L., & Buster, M. A. (2017). Hispanic and Asian performance on selection procedures: A narrative and meta-analytic review of 12 common predictors. *Journal of Applied Psychology*, *102*(8), 1178-1202.

Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance for supervisor ratings of overall performance: A policy capturing study. *Journal of Applied Psychology, 87*, 66-80.

Roulin, N., & Levashina, J. (2019). LinkedIn as a new selection method: Psychometric properties and assessment approach. *Personnel Psychology, 72*(2), 187-211

Rudner, L. M. (2012). *Demystifying the GMAT: Guarding against bias*. Graduate Management Admission Council.

Rupp, D. E., Song, Q. C., & Strah, N. (2020). Addressing the so-called validity–diversity trade-off: Exploring the practicalities and legal defensibility of Pareto-optimization for reducing adverse impact within personnel selection. *Industrial and Organizational Psychology*, *13*(2), 246-271.

Ryan, A. M., & McFarland, L. A. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology, 52*(2), 359-392.

Ryan, A. M., Ployhart, R. E., & Friedel, L. A. (1998). Using personality testing to reduce adverse impact: A cautionary note. *Journal of Applied Psychology, 83*, 298-307.

Ryan, A. M., Ployhart, R. E., Greguras, G. J., & Schmit, M. J. (1998). Test preparation programs in selection contexts: self-selection and program effectiveness. *Personnel Psychology, 51*, 599-621. https://doi.org/10.1111/j.1744-6570.1998.tb00253.x

Ryan, A. M., Reeder, M. C., Golubovich, J., Grand, J., Inceoglu, I., Bartram, D., ... & Yao, X. (2017). Culture and testing practices: is the world flat? *Applied Psychology: An International Review, 66*(3), 434-467.

Ryan, A. M., & Tippins, N. T. (2009). *Designing and implementing global selection systems*. Wiley-Blackwell.

Sackett, P. R. (2012). Cognitive tests, constructs, and content validity: A commentary on Schmidt (2012). *International Journal of Selection and Assessment, 20,* 23-27.

Sackett, P. R., Dahlke, J. A., Shewach, O. R., & Kuncel, N. R. (2017). Effects of predictor weighting methods on incremental validity. *Journal of Applied Psychology, 102,* 1421-1434.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*(3), 707-721.

Sackett, P. R., Laczo, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology, 88*(6), 1046–1056. https://doi.org/10.1037/0021-9010.88.6.1046

Sackett, P. R., Lievens, F., Van Iddekinge, C. H., & Kuncel, N. R. (2017). Individual differences and their measurement: A review of 100 years of research. *Journal of Applied Psychology, 102*(3), 254-273.

Sackett, P. R., Putka, D. J., & McCloy, R. A. (2012). The concept of validity and the process of validation. In N. Schmitt (Ed.), *Oxford handbook of assessment and selection* (pp. 91-118). Oxford University Press.

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., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56*(4), 302–318. https://doi.org/10.1037/0003-066X.56.4.302

Sackett, P. R., & Wanek, J. E. (1996). New developments in the use of measures of honesty integrity, conscientiousness, dependability trustworthiness, and reliability for personnel selection. *Personnel Psychology, 49*(4), 787-829.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, *49*(11), 929- 954

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*(1), 112-118.

Sackett, P. R., Zhang, C, & Berry, C. M. (2021). Challenging conclusions about predictive bias against Hispanic test-takers in personnel selection. *Journal of Applied Psychology.* Advance online publication.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2021). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*. Advance online publication. https://doi.org/10.1037/apl0000994

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology, 104*(10), 1207–1225.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., De Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology, 88*(6), 1068–1081. https://doi.org/10.1037/0021-9010.88.6.1068

Sanchez, J. I., & Levine, E. L. (2012). The rise and fall of job analysis and the future of work analysis. *Annual Review of Psychology, 63*(1), 397-425.

Sanchez-Monedero, J., Dencik, L., & Edwards, L. (2020) What does it mean to 'solve' the problem of discrimination in hiring? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM.*

Schäpers, P., Lievens, F., Freudenstein, J.-P., Hüffmeier, J., König, C. J., & Krumm, S. (2020). Removing situation descriptions from situational judgment test items: does the impact differ for video-based versus text-based formats? *Journal of Occupational and Organizational Psychology, 93*, 472-494.

Schäpers, P., Mussel, P., Lievens, F., König, C. J., Freudenstein, J.-P., & Krumm, S. (2020). The role of situations in situational judgment tests: Effects on construct saturation, predictive validity, and applicant perceptions. *Journal of Applied Psychology, 105*, 800-818.

Schein, V. E. (1978). Sex role stereotyping, ability and performance: Prior research and new directions. *Personnel Psychology, 31*, 259-268. https://doi.org/10.1111/j.1744-6570.1978.tb00445.x

Schippmann, J. S., Ash, R. A., Battista, M., Carr, L., Eyde, L. D., Hesketh B., Kehoe, J., Pearlman, K., Prien, E. P., & Sanchez, J. I. (2000). The practice of competency modeling. *Personnel Psychology, 53*, 703-740.

Schleicher, D. J., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2010). If at first you don't succeed, try, try again: Understanding race, age, and gender differences in retesting score improvement. *Journal of Applied Psychology, 95*(4), 603–617. https://doi.org/10.1037/a0018920

Schmidt, F. L. (1991). Why all banding procedures in personnel selection are logically flawed. *Human Performance, 8*(3), 165-177, DOI: 10.1207/s15327043hup0803_3

Schmidt, F. L. (2012). Cognitive tests used in selection can have content validity as well as criterion validity: A broader research review and implications for practice. *International Journal of Selection and Assessment*, *20*(1), 1-13.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, *36*(10), 1128–1137. https://doi.org/10.1037/0003-066X.36.10.1128

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on workforce productivity. *Personnel Psychology, 35*(2), 333-347.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61*(4), 473-485.

Schmidt, F. L., Shaffer, J. A., & Oh, I. S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology, 61*(4), 827-868.

Schmit, M. J., Kihm, J. A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology, 53*(1), 153-193.

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*(6), 966–974. https://doi.org/10.1037/0021-9010.78.6.966

Schmitt N., & Chan, D. (1998). *Personnel selection: A theoretical approach*. Sage.

Schmitt, N., & Kunce, C. (2002). The effects of required elaboration of answers to biodata questions. *Personnel Psychology, 55*(3), 569-587.

Schrader, A. D., & Osborn, H. G. (1977). Biodata faking: Effects of induced subtlety and position specificity. *Personnel Psychology, 30*, 395-404.

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65*, 445-494.

Shen, W., Sackett, P. R., Lievens, F., Schollaert, E., Van Hoye, G., Steiner, D. D., ... & Cook, M. (2017). Updated perspectives on the international legal environment for selection. In *Handbook of employee selection* (pp. 659-677). Routledge.

Shultz, M. M., & Zedeck, S. (2011). Predicting lawyer effectiveness: Broadening the basis for law school admission decisions. *Law & Social Inquiry, 36*(3), 620-661.

Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R.W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46*, 49-76. https://doi.org/10.1111/j.1744-6570.1993.tb00867.x

Society for Industrial and Organizational Psychology (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). Author.

Song, Q. C., Wee, S., & Newman, D. A. (2017). Diversity shrinkage: Cross-validating pareto-optimal weights to enhance diversity via hiring practices. *Journal of Applied Psychology, 102*, 1636-1657.

Stark, S., Chernyshenko, O. S., Chan, K.-Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86*, 943–953. doi:10.1037/0021-9010.86.5.943

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*(3), 497–508. https://doi.org/10.1037/0021-9010.89.3.497

Stemig, M. S., Sackett, P. R., & Lievens, F. (2015). Effects of organizationally endorsed coaching on performance and validity of situational judgment tests. *International Journal of Selection and Assessment, 23*(2), 174-181.

Stromberg, E. L. (1948). Testing programs draw better applicants. *Personnel Psychology, 1*(1), 21-29. https://doi.org/10.1111/j.1744-6570.1948.tb01290.x

Tay, L., Woo, S. E., Hickman, L., Booth, B., & De Mello, S. K. (2022). A conceptual framework for investigating and mitigating Machine Learning measurement bias (MLMB) in psychological assessment. *Advances in Methods and Practices in Psychological Science.* Advance online publication.

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in Machine Learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality, 34*, 826–844.

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology, 37*, 47-54.

Terpstra, D. E., & Rozell, E. J. (1993). The relationship of staffing practices to organizational level measures of performance. *Personnel Psychology, 46*(1), 27–48. https://doi.org/10.1111/j.1744-6570.1993.tb00866.x

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500-517.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology, 87*(6), 1020–1031. https://doi.org/10.1037/0021-9010.87.6.1020

van Hooft, E. A. J., & Born, M. P. (2012). Intentional response distortion on personality tests: Using eye-tracking to understand response processes when faking. *Journal of Applied Psychology, 97*(2), 301–316. https://doi.org/10.1037/a0025711

Van Iddekinge, C. H., & Arnold, J. D. (2017). Retaking employment tests: What we know and what we still need to know. *Annual Review of Organizational Psychology and Organizational Behavior, 4*, 445-471.

Van Iddekinge, C. H., Arnold, J. D., Frieder, R. E., & Roth, P. L. (2019). A meta-analysis of the criterion-related validity of prehire work experience. *Personnel Psychology*, *72*(4), 571-598.

Van Iddekinge, C. H., Ferris, G. R., Perrewé, P. L., Perryman, A. A., Blass, F. R., & Heetderks, T. D. (2009). Effects of selection and training on unit-level performance over time: A latent growth modeling approach. *Journal of Applied Psychology, 94*(4), 829–843. https://doi.org/10.1037/a0014453

Van Iddekinge, C. H., Morgeson, F. P., Schleicher, D. J., & Campion, M. A. (2011). Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. *Journal of Applied Psychology, 96*(5), 941–955. https://doi.org/10.1037/a0023562

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology*, *61*(4), 871-925.

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance, 19*(3), 175-199.

Vasilopoulos, N. L., Cucina, J. M., & McElreath, J. M. (2005). Do warnings of response verification moderate the relationship between personality and cognitive ability? *Journal of Applied Psychology, 90*(2), 306–322. https://doi.org/10.1037/0021-9010.90.2.306

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*(5), 557-572.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*, 25-49. https://doi.org/10.1111/j.1744-6570.1997.tb00899.x

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52*(5), 372-376.

Whitney, D. J., & Schmitt, N. (1997). Relationship between culture and responses to biodata employment items. *Journal of Applied Psychology, 82*(1), 113–129. https://doi.org/10.1037/0021-9010.82.1.113

Wright, P. M., Gardner, T. M., Moynihan, L. M., & Allen, M. R. (2005). The relationship between HR practices and firm performance: Examining causal order. *Personnel Psychology*, *58*(2), 409–446. https://doi.org/10.1111/j.1744-6570.2005.00487.x

Wu, F. Y., Mulfinger, E., Alexander, L. III, Sinclair, A. L., McCloy, R. A., & Oswald, F. L. (2022). Individual differences at play: An investigation into measuring Big Five personality facets with game-based assessments. *International Journal of Selection and Assessment*, *30*(1), 62-81.

Zedeck, S., Outtz, J., Cascio, W. F., & Goldstein, I. L. (1991). Why do "testing experts" have such limited vision? *Human Performance, 4*(4), 297-308.

Zhang, L., Van Iddekinge, C. H., Arnold, J. D., Roth, P. L., Lievens, F., Lanivich, S. E., & Jordan, S. L. (2020). What's on job seekers' social media sites? A content analysis and effects of structure on recruiter judgments and predictive validity. *Journal of Applied Psychology, 105(12), 1530–1546.*

Zieky, M. (2006). Fairness reviews in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359–376). Lawrence Erlbaum Associates.

Table 1

Implications for Selection Research and Practice

Evaluation factors	**Implications**
Criterion-related and content validity	• Establishing validity, drawing from multiple available strategies, is a fundamental issue for all credible selection systems.
Subgroup differences, adverse impact, and bias	• The cognitive load of predictors is the main driver of racio-ethnic subgroup differences. • Predictive bias analyses should be conducted at the selection system level rather than at the individual procedure level.
Applicant reactions	• Applicants tend to react more positively to selection procedures that are more "sample like" (e.g., interviews, work samples) than procedures that are more "sign like" (e.g., cognitive ability tests, integrity tests). • Reactions can affect factors such as selection test performance and organization attraction, but they do not tend to affect criterion-related validity or subgroup differences.
Cost and time	• Use pareto optimization methods to assess potential tradeoffs among criterion-related validity, adverse impact, and costs. • Initial research suggests that the time applicants spend completing selection procedures does not influence factors such as withdrawal from the process.
Effects on unit- and firm-level outcomes	• Initial research suggests that selection practices are more strongly related to unit-level than to firm-level outcomes. • To examine the effects of selection on unit- or firm-level outcomes, use more specific measures of selection practices (e.g., actual test scores) than more general measures such as practice use or selection ratios.
Developmental decisions	
Use multiple selection procedures	• The construct(s) each selection procedure assesses is key to changes in criterion-related validity and subgroup differences resulting from additional procedures. • When predicting overall job performance, adding more job relevant predictors tends to reduce predictive bias.
Add structure	• A moderately high level of structure may balance goals related to criterion-related validity, legal defensibility, and applicant reactions. • Algorithmic integration of selection information produces better criterion-related validity than judgmental integration.

Contextualize procedure content	• Placing personality test items in a work context tends to improve criterion-related validity but does not reduce subgroup differences or enhance applicant reactions. • Contextual information does not appear to exert a strong influence on the criterion-related validity of SJTs.
Minimize cognitive load	• The distinction between construct relevant versus irrelevant cognitive load appears important. • Method factors such as stimulus (textual vs. video presentation) and response formats (open vs. close ended responses) can affect cognitive load.
Review procedure content	• Have SMEs review the content of selection procedures to identify potentially problematic items, such as those that include offensive or unfamiliar language. • Item review panels should be diverse to balance potential differences in perceptions of item sensitivity. • Consider assessing differential item functioning (DIF), particularly the magnitude of the effects at the level at which the scores will be used to make selection decisions.
Limit opportunities to fake	• Although prevalence of faking varies based on factors such as the selection procedure and sample, faking is a concern because it can negatively affect validity and influence selection decisions • The most effective methods to deter and detect faking appear to be certain types of forced-choice measures. Requiring applicants to elaborate the basis for their responses and "warning and retest" approaches also appear promising.
Gamify selection procedures	• Gamification tends to enhance applicant reactions to selection procedures. • Initial research suggests that gamification does not increase criterion-related validity and may decrease it and/or other sources of validity evidence by altering the constructs assessed. • Initial research suggests gamification does not affect racio-ethnic differences, but it can produce sex differences that favor males.
Administration decisions	
Pre-test explanations, practice tests, and coaching	• Pre-test explanations do not tend to affect applicant reactions. • Initial research suggests that providing practice tests can increase application rates and improve scores on the actual tests, particularly for racio-ethnic minorities. • Coaching programs tend to increase applicants' performance on selection procedures (although increases can be small) and do not appear to negative impact criterion-related validity and may even enhance it.
Revealing target KSAOs	• Revealing the target KSAOs tends to improve applicant reactions and performance on selection procedures.

	• Effects of transparency on criterion-related validity are still unclear. Some evidence suggests revealing the KSAOs can affect which constructs are activated, which, in turn, can affect prediction of different criteria.
Retesting	• Applicants tend to improve upon retesting, particularly for selection procedures that are more novel, less g-loaded, performance-based, and more fakeable. • Score improvements tend to be larger for high-ability, White, female, and younger applicants. • Retest scores tend to be more reliable and demonstrate stronger criterion-related validity. However, because the scores of one-time applicants may be more valid, allowing retesting could lower validity overall.
Scoring decisions	
Predictor and criterion weighting	• Choose a weighting approach consistent with valued objectives (e.g., validity maximization vs. administrative ease vs. balancing validity and diversity). • Consider the relative importance of the criterion to be predicted, which, in turn, can affect how much weight each selection procedure should be given.
Artificial intelligence	• AI-scoring of selection procedures tends to converge to a moderate to large degree with self- and other-reports. • Although AI-based scores can predict relevant criteria, data on job-related criteria are lacking. • Preliminary evidence suggests that AI-scoring is associated with similar subgroup differences as human scoring.
Score banding	• Although minority preference within bands is the most effective way to increase diversity, the courts have not upheld its sole use as a basis for selection. • It is possible to use variables that have small subgroup differences to make within-band selection decisions. However, using such variables earlier in the process as traditional predictors tends to be more effective.

Table 2

Directions for Future Selection Research

<u>Factors on which selection procedures are evaluated</u> • Can we continue to improve validity estimation methods (e.g., via increased insight into appropriate range restriction and reliability corrections)? • Can we increase the evidence base for new or understudied predictors (e.g., game-based assessments, technology-enhance assessments and subgroups (e.g., White-Black subgroup differences have been examined for a much broader range of predictors than have White-Hispanic differences)? • What are the tradeoffs between selection procedure length and other evaluation factors such as selection process withdrawal and criterion-related validity? • What are the effects of selection procedure scores (rather than mere use of validated procedures) on collective outcomes such as unit performance and retention?
<u>Developmental decisions</u> • What are the tradeoffs between synchronous and asynchronous video interviews? • What are the effects of providing contextual information on selection procedures? • Can we further clarify how method factors affect the extent to which selection procedures capture cognitive ability and other (un)intended constructs? • How can organizations best conduct and improve sensitivity reviews? • Are there prospects for identifying item features linked to DIF? • Can we build on recent insights into preventing (e.g., via forced-choice methods) and detecting (e.g., via real-time warnings) faking? • What are the tradeoffs of using game-based selection procedures in terms of validity, subgroup differences, applicant reactions, and cost? • What design factors exert the most influence on selection procedure evaluation criteria?
<u>Administration decisions</u> • How do different forms of applicant preparation affect evaluation criteria such as validity and subgroup differences? • Under what conditions does revealing KSAOs affect criterion-related validity? • What are the long-term effects of retesting on evaluation criteria such as validity and subgroup differences?

Scoring decisions

- What is the criterion-related validity of AI-based scores compared to traditional scores?
- To what extent does AI-based scoring produce subgroup differences and what underlies those differences?
- Does AI-based scoring result in differential prediction?
- How do various stakeholders (e.g., applicants, hiring managers) perceive AI-based selection procedures?
- Does anthropomorphizing selection procedures improve the applicant experience?
- To what extent do organizations use banding and how are they doing it?
- What are the effects of banding on criterion-related validity?
- How do organizations select gig workers?
- What KSAOs are important for gig and remote work?
- Is the selection process of gig workers prone to biases?