Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection Lee Kong Chian School Of Business

Lee Kong Chian School of Business

1-2008

Personnel selection

Paul R. SACKETT
University of Minnesota - Twin Cities

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg
DOI: https://doi.org/10.1146/annurev.psych.59.103006.093716

Follow this and additional works at: https://ink.library.smu.edu.sg/lkcsb_research

    Part of the Human Resources Management Commons, and the Organizational Behavior and Theory Commons

Citation

Personnel Selection

Paul R. Sackett[1] and Filip Lievens[2]

1Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455;

email: psackett@umn.edu

2Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000, Ghent,

Belgium; email: filip.lievens@ugent.be

Submitted version

Abstract

We review developments in personnel selection since the previous review by Hough & Oswald (2000) in the Annual Review of Psychology. We organize the review around a taxonomic structure of possible bases for improved selection, which includes (a) better understanding of the criterion domain and criterion measurement, (b) improved measurement of existing predictor methods or constructs, (c) identification and measurement of new predictor methods or constructs, (d) improved identification of features that moderate or mediate predictor-criterion relationships, (e) clearer understanding of the relationship between predictors or between predictors and criteria (e.g., via meta-analytic synthesis), (f) identification and prediction of new outcome variables, (g) improved ability to determine how well we predict the outcomes of interest, (h) improved understanding of subgroup differences, fairness, bias, and the legal defensibility, (i) improved administrative ease with which selection systems can be used, (j) improved insight into applicant reactions, and (k) improved decision-maker acceptance of selection systems.

Key Words

job performance, testing, validity, adverse impact, ability, personality

INTRODUCTION

Personnel selection has a long history of coverage in the Annual Review of Psychology. This is the first treatment of the topic since 2000, and length constraints make this a selective review of work since the prior article by Hough & Oswald (2000). Our approach to this review is to focus more on learning (i.e., How has our understanding of selection and our ability to select effectively changed?) than on documenting activity (i.e., What has been done?). We envision a reader who vanished from the scene after the Hough & Oswald review and reappears now, asking, "Are we able to do a better job of selection now than we could in 2000?" Thus, we organize this review around a taxonomic structure of possible bases for improved selection, which we list below.

1.   Better prediction of traditional outcome measures, as a result of:

a.   better understanding of the criterion domain and criterion measurement

b.   improved measurement of existing predictor methods or constructs

c.   identification and measurement of new predictor methods or constructs

d.   improved identification of features that moderate or mediate predictor-criterion relationships

e.   clearer understanding of the relationship between predictors or between predictors and criteria (e.g., via meta-analytic synthesis)

2.   Identification and prediction of new outcome variables

3.   Improved ability to determine how well we predict the outcomes of interest (i.e., improved techniques for estimating validity)

4.   Improved understanding of subgroup differences, fairness, bias, and the legal defensibility of our selection systems

5.   Improved administrative ease with which selection systems can be used

6.   Improved methods obtaining more favorable applicant reactions and better insight into consequences of applicant reactions

7.   Improved decision-maker acceptance of selection systems

Although our focus is on new research findings, we note that there are a number of professional developments important for anyone interested in the selection field. The Important Professional Developments sidebar briefly outlines these developments.

IMPORTANT PROFESSIONAL DEVELOPMENTS

Updated versions of the Standards for Educational and Psychological Testing (Am. Educ. Res. Assoc., Am. Psychol. Assoc., Natl. Counc. Meas. Educ. 1999) and the Principles for the Validation and Use of Personnel Selection Procedures (Soc. Ind. Organ. Psychol. 2003) have appeared. Jeanneret (2005) provides a useful summary and comparison of these documents.

Guidance on computer and Internet-based testing is provided in an American Psychological Association Task Force report (Naglieri et al. 2004) and in guidelines prepared by the International Test Commission (Int. Test Comm. 2006).

Edited volumes on a number of selection-related themes have appeared in the Society for Industrial and Organizational Psychology's two book series, including volumes on the management of selection systems (Kehoe 2000), personality in organizations (Barrick & Ryan 2003), discrimination at work (Dipboye & Colella 2004), employment discrimination litigation (Landy 2005a), and situational judgment tests (Weekley & Ployhart 2006). There are also edited volumes on validity generalization (Murphy 2003), test score banding (Aguinis 2004), emotional intelligence (Murphy 2006), and the Army's Project A (Campbell & Knapp 2001).

Two handbooks offering broad coverage of the industrial/organizational (I/O) field have been published; both containing multiple chapters examining various aspects of the personnel selection process (Anderson et al. 2001, Borman et al. 2003).

CAN WE PREDICT TRADITIONAL OUTCOME MEASURES BETTER BECAUSE OF BETTER UNDERSTANDING OF THE CRITERION DOMAIN AND CRITERION MEASUREMENT?

Conceptualization of the Criterion Domain

Research continues an ongoing trend of moving beyond a single unitary construct of job performance to a more differentiated model. Campbell's influential perspective on the dimensionality of performance (e.g., Campbell et al. 1993) and the large-scale demonstrations in the military's Project A (Campbell & Knapp 2001) of differential relationships between predictor constructs (e.g., ability, personality) and criterion constructs (e.g., task proficiency, effort, maintaining personal discipline) contributed to making this a major focus of contemporary research on predictor-criterion relationships.

Two major developments in understanding criterion dimensions are the emergence of extensive literature on organizational citizenship behavior (Podsakoff et al. 2000; see also Borman & Motowidlo 1997, on the closely related topic of contextual performance) and on counterproductive work behavior (Sackett & Devore 2001, Spector & Fox 2005). Dalal (2005) presents a meta-analysis of relationships between these two domains; the modest correlations (mean $r = -0.32$ corrected for measurement error) support the differentiation of these two, rather than the view that they are merely opposite poles of a single continuum. Rotundo & Sackett (2002) review and integrate a number of perspectives on the dimensionality of job performance and offer task performance, citizenship performance, and counterproductive work behavior as the three major domains of job performance. With cognitively loaded predictors as generally the strongest correlates of task performance and noncognitive predictors as generally the best predictors in the citizenship and counterproductive behavior domains, careful attention

to the criterion of interest to the organization is a critical determinant of the eventual makeup and success of a selection system.

Predictor-Criterion Matching

Related to the notion of criterion dimensionality, there is increased insight into predictor-criterion matching. This elaborates on the notion of specifying the criterion of interest and selecting predictors accordingly. We give several examples. First, Bartram (2005) offered an eight-dimensional taxonomy of performance dimensions for managerial jobs, paired with a set of hypotheses about specific ability and personality factors conceptually relevant to each dimension. He then showed higher validities for the hypothesized predictor-criterion combinations. Second, Hogan & Holland (2003) sorted criterion dimensions based on their conceptual relevance to various personality dimensions and then documented higher validity for personality dimensions when matched against these relevant criteria; see Hough & Oswald (2005) for additional examples in the personality domain. Third, Lievens et al. (2005a) classified medical schools as basic science–oriented versus patient care–oriented, and found an interpersonally oriented situational judgment test predictive of performance only in the patient care–oriented schools.

The Role of Time in Criterion Measurement

Apart from developments in better understanding the dimensionality of the criterion and in uncovering predictors for these different criterion dimensions, important progress has also been made in understanding the (in)stability of the criterion over time. Sturman et al. (2005) developed an approach to differentiating between temporal consistency, performance stability, and test-retest reliability. Removing the effects of temporal instability from indices of performance consistency is needed to understand the degree to which change in measured performance over time is a result of error in the measurement of performance versus real change in performance.

Predicting Performance Over Time

This renewed emphasis in the dynamic nature of the criterion has also generated studies that aim to predict change in the criterion construct. Studies have examined whether predictors of job performance differ across job stages. The transitional job stage, where there is a need to learn new things, is typically contrasted to the more routine maintenance job stage (Murphy 1989). Thoresen et al. (2004) found that the Big Five personality factor of Openness was related to performance and performance trends in the transition stage but not to performance at the maintenance stage. Stewart (1999) showed that the dependability aspects of the Conscientiousness factor (e.g., self-discipline) were related to job performance at the transitional stage, whereas the volitional facets of Conscientiousness (e.g., achievement motivation) were linked to job performance at the maintenance stage. Also worthy of note in understanding and predicting performance over time is Stewart & Nandleolyar's (2006) comparison of interindividual and intraindividual variation in performance over time. In a sales sample, they find greater intraindividual variation than interindividual variation in week-to-week performance. These intraindividual differences were further significantly determined by whether people benefited from situational opportunities (i.e., adaptability). There was also evidence that particular personality traits enabled people to increase their performance by effectively adapting to changes in the environment. Sales people high in Conscientiousness were better able to benefit from situational opportunities when they saw these opportunities as goals to achieve (task pursuit). The reverse was found for people high in Openness, who might be more effective in task revision situations.

Criterion Measurement

Turning from developments in the conceptualization of criteria to developments in criterion measurement, a common finding in ratings of job performance is a pattern of relatively high correlations among dimensions, even in the presence of careful scale development designed to maximize differentiation among scales. A common explanation offered for this is halo, as single raters commonly rate an employee on all dimensions. Viswesvaran et al. (2005) provided useful insights into this issue by meta-analytically comparing correlations between performance dimensions made by differing raters with those made by the same rater. Although ratings by the same rater were higher (mean interdimension r = 0.72) than those from different raters (mean r = 0.54), a strong general factor was found in both. Thus, the finding of a strong general factor is not an artifact of rater-specific halo.

With respect to innovations in rating format, Borman et al. (2001) introduced the computerized adaptive rating scale (CARS). Building on principles of adaptive testing, they scaled a set of performance behaviors and then used a computer to present raters with pairs of behaviors differing in effectiveness. The choice of which behavior best describes the ratee drives the selection of the next pair of behaviors, thus honing in on the ratee's level of effectiveness. Using ratings of videotaped performance episodes to compare CARS with graphic scales and behaviorally anchored scales, they reported higher reliability, validity, accuracy, and more favorable user reactions for CARS. Although this study is at the initial demonstration stage, it does suggest a potential route to higher-quality performance measures.

CAN WE PREDICT TRADITIONAL OUTCOME MEASURES BETTER BECAUSE OF IMPROVED MEASUREMENT OF EXISTING PREDICTOR METHODS OR CONSTRUCTS?

The prediction of traditional outcomes might be increased by improving the measurement of existing selection procedures. We outline five general strategies that have been pursued in attempting to improve existing selection procedures. Although there is research on attempts to improve measurement of a variety of constructs, much of the work focuses on the personality domain. Research in the period covered by this review continues the enormous surge of interest in personality that began in the past decade. Our sense is that a variety of factors contribute to this surge of interest, including (a) the clear relevance of the personality domain for the prediction of performance dimensions that go beyond task performance (e.g., citizenship and counterproductive behavior), (b) the potential for incremental validity in the prediction of task performance, (c) the common finding of minimal racial/ethnic group differences, thus offering the prospect of reduced adverse impact, and (d) some unease about the magnitude of validity coefficients obtained using personality measures. There seems to be a general sense that personality "should" fare better than it does. Our sense is that what is emerging is that there are sizable relationships between variables in the personality domain and important work outcomes, but that the pattern of relationships is complex. We believe the field "got spoiled" by the relatively straightforward pattern of findings in the ability domain (e.g., relatively high correlations between different attempts to measure cognitive ability and consistent success in relating virtually any test with a substantial cognitive loading to job performance measures). In the personality domain, mean validity coefficients for single Big Five traits are indeed relatively small (e.g., the largest corrected validity, for Conscientiousness, is about 0.20), leading to some critical views of the use of personality measures (e.g., Murphy & Dzieweczynski 2005). However, overall performance is predicted much better by compound traits and by composites of Big Five measures. In addition, more specific performance dimensions (e.g., citizenship, counterproductive work behavior) are better predicted by carefully selected measures that may be subfacets of broad Big Five traits (Hough & Oswald 2005, Ones et al. 2005). As the work detailed below indicates, the field does not yet have a complete understanding of the role of personality constructs and personality measures.

Measure the Same Construct with Another Method

The first strategy is to measure the same construct with another method. This strategy recognizes that the constructs being measured (such as conscientiousness, cognitive ability, manual dexterity) should be distinguished from the method of measurement (such as self-report inventories, tests, interviews, work samples). In the personality domain, there have been several attempts at developing alternatives to traditional self-report measures. One has been to explicitly structure interviews around the Five-Factor Model (Barrick et al. 2000, Van Iddekinge et al. 2005) instead of using self-report personality inventories. Even in traditional interviews, personality factors (35%) and social skills (28%) are the most frequently measured constructs (Huffcutt et al. 2001a). Another is to develop implicit measures of personality. One example of this is Motowidlo et al.'s (2006) development of situational judgment tests designed to tap an individual's implicit trait theories. They theorize, and then offer evidence, that individual personality shapes individual judgments of the effectiveness of behaviors reflecting high to low levels of the trait in question. Thus, it may prove possible to make inferences about personality from an individual's judgments of the effectiveness of various behaviors. Another approach to implicit measurement of personality is conditional reasoning (James et al. 2005) based on the notion that people use various justification mechanisms to explain their behavior, and that people with varying dispositional tendencies will employ differing justification mechanisms. The basic paradigm is to present what appear to be logical reasoning problems, in which respondents are asked to select the response that follows most logically from an initial statement. In fact, the alternatives reflect various justification mechanisms. James et al. (2005) present considerable validity evidence for a conditional reasoning measure of aggression. Other research found that a conditional reasoning test of aggression could not be faked, provided that the real purpose of the test is not disclosed (LeBreton et al. 2007).

Improve Construct Measurement

A second strategy is to improve the measurement of the constructs underlying existing selection methods. For instance, in the personality domain, it has been argued that the validity of scales that were originally developed to measure the Five-Factor Model of personality will be higher than scales categorized in this framework post hoc. Evidence has been mixed: Salgado (2003) found such effects for Conscientiousness and Emotional Stability scales; however, Hurtz & Donovan (2000) did not. Another example is Schmit et al.'s (2000) development of a personality instrument based in broad international input, thus avoiding idiosyncrasies of a single nation or culture in instrument content. Apart from using better scales, one might also experiment with other response process models as a way of improving the quality of construct measurement in personality inventories. Existing personality inventories typically assume that candidates use a dominance response process. Whereas such a dominance response process is clearly appropriate for cognitive ability tests, ideal point process models seem to provide a better fit of candidates' responses to personality test items than do the dominance models (even though these personality inventories were developed on the basis of a dominance model; Stark et al. 2006).

These attempts to improve the quality of construct measurement are not limited to the personality domain. Advances have been made in unraveling the construct validity puzzle in assessment centers. Although there is now relative consensus that assessment center exercises are more important than assessment center constructs (Bowler & Woehr 2006, Lance et al. 2004), we have a better understanding of which factors affect the quality of construct measurement in assessment centers. First, well-designed assessment centers show more construct validity evidence (Arthur et al. 2000, Lievens & Conway 2001). For instance, there is better construct validity when fewer dimensions are used and when assessors are psychologists. High interassessor reliability is important; otherwise, variance due to assessors will be confounded with variance due to exercises because assessors typically rotate through the various exercises (Kolk et al. 2002). Third, various studies (Lance et al. 2000, Lievens 2002) identified the nature

of candidate performance as another key factor. Construct validity evidence was established only for candidates whose performances varied across dimensions and were relatively consistent across exercises.

Increase Contextualization

A third strategy is to increase the contextualization of existing selection procedures. We commonly view predictors on a continuum from sign to sample. General ability tests and personality inventories are then typically categorized as signs because they aim to measure decontextualized abilities and predispositions that signal or forecast subsequent workplace effectiveness. Conversely, assessment center exercises and work samples are considered to be samples because they are based on behavioral consistency between behavior during the selection procedure and job behavior. Increasing the contextualization of a personality inventory makes this distinction less clear. In particular, it has been argued that the common use of noncontextualized personality items (e.g., "I pay attention to details") is one reason for the relatively low criterion-related validities of personality scales. Because of the ambiguous nature of such items, a general frame of reference (how do I behave across a variety of situations) may be the basis for an individual's response for one item, whereas work behavior or some other frame of reference might serve as basis for completing another item. Contextualized personality inventories aim to circumvent these interpretation problems by using a specific frame of reference (e.g., "I pay attention to details at work"). Recent studies have generally found considerable support for the use of contextualized personality scales as a way of improving the criterion-related validity of personality scales (Bing et al. 2004, Hunthausen et al. 2003).

Reduce Response Distortion

A fourth strategy is to attempt to reduce the level of response distortion (i.e., faking). This approach seems especially useful for noncognitive selection procedures that are based on self-reports (e.g., personality inventories, biodata) rather than on actual behaviors. Recent research has compared different noncognitive selection procedures in terms of faking. A self-report personality measure was more prone to faking than a structured interview that was specifically designed to measure the same personality factors (Van Iddekinge et al. 2005). However, structured interviews themselves were more prone to impression management than were assessment center exercises that tapped interpersonal skills (McFarland et al. 2005). So, these results suggest that faking is most problematic for self-report personality inventories, followed by structured interviews and then by assessment centers.

Although social desirability corrections are still the single most used response-distortion reduction technique [used by 56% of human resource (HR) managers, according to a survey by Goffin & Christiansen (2003)], research has shown that this strategy is ineffective. For example, Ellingson et al. (1999) used a within-subjects design and determined that scores obtained under faking instructions could not be corrected to match scores obtained under instructions to respond honestly. Building on the conclusion that corrections are not effective, Schmitt & Oswald (2006) examined the more radical strategy of removing applicants with high faking scores from consideration for selection and found this had small effects on the mean performance of those selected.

One provocative finding that has emerged is that important differences appear to exist between instructed faking and naturally occurring faking. Ellingson et al. (1999) found that the multidimensional structure of a personality inventory collapsed to a single factor under instructed faking conditions; in contrast, Ellingson et al. (2001) found that the multidimensional structure was retained in operational testing settings, even among candidates with extremely high social desirability scores. This suggests the

possibility that instructed faking results in a different response strategy (i.e., consistent choice of the socially desirable response across all items), whereas operational faking is more nuanced. Note also that instructed faking studies vary in terms of whether they focus on a specific job, on the workplace in general, or on a nonspecified context. This merits additional attention, given the extensive reliance on instructed faking as a research strategy.

We note that most research modeling the effects of faking have focused on top-down selection when using personality measures. However, in many operational settings, such measures are used with a relatively low fixed cutoff as part of initial screening. In such a setting, faking may result in an undeserving candidate succeeding in meeting the threshold for moving on to the next stage, but that candidate does not supplant a candidate who responds honestly on a rank order list, as in the case of top-down selection. The issue of unfairness to candidates responding honestly is less pressing here than in the case of top-down selection. In this vein, Mueller-Hanson et al. (2003) showed that faking reduced the validity of a measure of achievement motivation at the high end of the distribution but not at the low end, suggesting that faking may be less of an obstacle to screen-out uses of noncognitive measures than to screen-in uses.

Given these poor results for social desirability corrections, it seems important to redirect our attention to other interventions for reducing deliberate response distortion. So far, success has been mixed. A first preventive approach is to warn candidates that fakers can be identified and will be penalized. However, the empirical evidence shows only meager effects (around 0.25 standard deviation, or SD) for a combination of identification-only and consequences-only warnings on predictor scores and faking scale scores (Dwight & Donovan 2003). A second approach requires candidates to provide a written elaboration of their responses. This strategy seems useful only when the items are verifiable (e.g., biodata items). Elaboration lowered mean biodata scores but had no effect on criterion-related validity (Schmitt & Kunce 2002, Schmitt et al. 2003). Third, the use of forced response formats has received renewed attention. Although a multidimensional forced-choice response format was effective for reducing score inflation at the group level, it was affected by faking to the same degree as a traditional Likert scale at the individual-level analysis (Heggestad et al. 2006).

Impose Structure

A fifth strategy is to impose more structure on existing selection procedures. Creating a more structured format for evaluation should increase the level of standardization and therefore reliability and validity. Highly structured employment interviews constitute the best-known example of this principle successfully being put into action. For example, Schmidt & Zimmerman (2004) showed that a structured interview administered by one interviewer obtains the same level of validity as three to four independent unstructured interviews. The importance of question-and-response scoring standardization in employment interviews was further confirmed by the beneficial effects of interviewing with a telephone-based script (Schmidt & Rader 1999) and of carefully taking notes (Middendorf & Macan 2002).

Although increasing the level of structure has been especially applied to interviews, there is no reason why this principle would not be relevant for other selection procedures where standardization might be an issue. Indeed, in the context of reference checks, Taylor et al. (2004) found that reference checks significantly predicted supervisory ratings (0.36) when they were conducted in a structured and telephone-based format. Similarly, provision of frame-of-reference training to assessors affected the

construct validity of their ratings, even though criterion-related validity was not affected (Lievens 2001, Schleicher et al. 2002).

Thus, multiple strategies have been suggested and examined with the goal of improving existing predictors. The result is several promising lines of inquiry.

CAN WE PREDICT TRADITIONAL OUTCOME MEASURES BETTER BECAUSE OF IDENTIFICATION AND MEASUREMENT OF NEW PREDICTOR METHODS OR CONSTRUCTS?

Emotional Intelligence

In recent years, emotional intelligence (EI) is probably the new psychological construct that has received the greatest attention in both practitioner and academic literature. It has received considerable critical scrutiny from selection researchers as a result of ambiguous definition, dimensions, and operationalization, and also as a result of questionable claims of validity and incremental validity (Landy 2005b, Matthews et al. 2004, Murphy 2006; see also Mayer et al. 2008). A breakthrough in the conceptual confusion around EI is the division of EI measures into either ability or mixed models (Côté & Miners 2006, Zeidner et al. 2004). The mixed (self-report) EI model views EI as akin to personality. A recent meta-analysis showed that EI measures based on this mixed model overlapped considerably with personality trait scores but not with cognitive ability (Van Rooy et al. 2005). Conversely, EI measures developed according to an EI ability model (e.g., EI as ability to accurately perceive, appraise, and express emotion) correlated more with cognitive ability and less with personality. Note too that measures based on the two models correlated only 0.14 with one another. Generally, EI measures (collapsing both models) produce a meta-analytic mean correlation of 0.23 with performance measures (Van Rooy & Viswesvaran 2004). However, this included measures of performance in many domains beyond job performance, included only a small number of studies using ability-based EI instruments, and included a sizable number of studies using self-report measures of performance. Thus, although clarification of the differing conceptualizations of EI sets the stage for further work, we are still far from being at the point of rendering a decision as to the incremental value of EI for selection purposes.

Situational Judgment Tests

Interest has recently surged in the class of predictors under the rubric of situational judgment tests (SJTs). Although not a new idea, they were independently reintroduced under differing labels and found a receptive audience. Motowidlo et al. (1990) framed them as "low fidelity simulations," and Sternberg and colleagues (e.g., Wagner & Sternberg 1985) framed them as measures of "tacit knowledge" and "practical intelligence." Sternberg presented these measures in the context of a generally critical evaluation of the use of general cognitive ability measures (see Gottfredson 2003 and McDaniel & Whetzel 2005 for responses to these claims), while current use in I/O generally views them as a potential supplement to ability and personality measures. An edited volume by Weekley & Ployhart (2006) is a comprehensive treatment of current developments with SJTs.

McDaniel et al. (2001) meta-analyzed 102 validity coefficients (albeit only 6 predictive validity coefficients) and found a mean corrected validity of 0.34. Similarly, SJTs had incremental validity over cognitive ability, experience, and personality (Chan & Schmitt 2002, Clevenger et al. 2001). With this regard, there is also substantial evidence that SJTs have value for broadening the type of skills measured in college admission (Lievens et al. 2005a, Oswald et al. 2004).

Now that SJTs have established themselves as valid predictors in the employment and education domains, attention has turned to better understanding their features. The type of response instructions seems to be a key factor, as it has been found to affect the cognitive loading and amount of response distortion in SJTs (McDaniel et al. 2007, Nguyen et al. 2005). Behavioral-tendency instructions (e.g., "What are you most likely to do?") exhibited lower correlations with cognitive ability and lower adverse impact but higher faking than knowledge-based instructions (e.g., "What is the best answer?"). The amount of fidelity appears to be another factor. For example, changing an existing video-based SJT to a written format (keeping content constant) substantially reduced the criterion-related validity of the test (Lievens & Sackett 2006). We also need to enhance our understanding of why SJTs predict work behavior. Recently, procedural knowledge and implicit trait policies have been advocated as two plausible explanations (Motowidlo et al. 2006). These might open a window of possibilities for more theory-based research on SJTs.

CAN WE PREDICT TRADITIONAL OUTCOME MEASURES BETTER BECAUSE OF IMPROVED IDENTIFICATION OF FEATURES THAT MODERATE OR MEDIATE RELATIONSHIPS?

In recent years, researchers have gradually moved away from examining main effects of selection procedures ("Is this selection procedure related to performance?") and toward investigating moderating and mediating effects that might explain when (moderators) and why (mediators) selection procedures factors are or are not related to performance. Again, most progress has been made in increasing our understanding of possible moderators and mediators of the validity of personality tests.

Situation-Based Moderators

With respect to situation-based moderators, Tett & Burnett's (2003) person-situation interactionist model of job performance provided a huge step forward because it explicitly focused on situations as moderators of trait expression and trait evaluation. Hence, this model laid the foundation for specifying the conditions under which specific traits will predict job performance. This model goes much further than the earlier distinction between weak and strong situations. Its main hypothesis states that traits will be related to job performance in a given setting when (a) employees vary in their level on the trait, (b) trait expression is triggered by various situational (task, social, and organizational) cues, (c) trait-expressive behavior contributes to organizational effectiveness, and (d) the situation is not so strong as to override the expression of behavior. The model also outlines specific situational features (demands, distracters, constraints, and releasers) at three levels (task, social, and organizational); thus, it might serve as a welcome taxonomy to describe situations and interpret personality-performance relationships. Its value to understanding behavioral expression/evaluation as triggered by situations is not limited to personality but has also been fruitfully used in sample-based predictors such as assessment centers (Lievens et al. 2006).

Person-Based Moderators

The same conceptual reasoning runs through person-based moderators of personality-performance relationships. Similar to situational features, specific individual differences variables might constrain the behavior exhibited, in turn limiting the expression of underlying traits. For example, the relation between Big Five traits such as Extraversion, Emotional Stability, and Openness to Experience and interpersonal performance was lower when self-monitoring was high because people who are high in self-monitoring seem to be so motivated to adapt their behavior to environmental cues that it restricts their behavioral expressions (Barrick et al. 2005). Similar interactions between Conscientiousness and Agreeableness (Witt et al. 2002), Conscientiousness and Extraversion (Witt 2002), and Conscientiousness and social skills (Witt & Ferris 2003) have been discovered. In all of these cases, high levels of Conscientiousness

coupled with either low levels of Agreeableness, low levels of Extraversion, or inadequate social skills were detrimental for performance. These results also have practical relevance. For example, they highlight that selecting people high in Conscientiousness but low in Agreeableness for jobs that require frequent collaboration reduces validities to zero.

A literature is emerging on retesting as a moderator of validity. Hausknecht et al. (2002, 2007) showed that retesters perform better and are less likely to turn over, holding cognitive ability constant, which they attribute to higher commitment to the organization. In contrast, Lievens et al. (2005) reported lower-criterion performance among individuals who obtained a given score upon retesting than among those obtaining the same score on the first attempt. They also reported within-person analyses showing higher validity for a retest than an initial test, suggesting that score improvement upon retesting reflects true score change rather than artifactual observed score improvement.

Mediators

Finally, in terms of mediators, there is some evidence that distal measures of personality traits relate to work behavior through more proximal motivational intentions (Barrick et al. 2002). Examples of such motivation intentions are status striving, communion striving, and accomplishment striving. A distal trait such as Agreeableness is then related to communion striving, Conscientiousness to accomplishment striving, and Extraversion to status striving. However, the most striking result was that Extraversion was linked to work performance through its effect on status striving. Thus, we are starting to dig deeper into personality-performance relationships in search of the reasons why and how these two are related.

CAN WE PREDICT TRADITIONAL OUTCOME MEASURES BETTER BECAUSE OF CLEARER UNDERSTANDING OF RELATIONSHIPS BETWEEN PREDICTORS AND CRITERIA?

In this section, we examine research that generally integrates and synthesizes findings from primary studies. Such work does not enable better prediction per se, but rather gives us better insight into the expected values of relationships between variables of interest. Such findings may affect the quality of eventual selection systems to the degree that they aid selection system designers in making a priori design choices that increase the likelihood that selected predictors will prove related to the criteria of interest. Much of the work summarized here is meta-analytic, but we note that other meta-analyses are referenced in various places in this review as appropriate.

Incremental Validity

A significant trend is a new focus in meta-analytic research on incremental validity. There are many meta-analytic summaries of the validity of individual predictors, and recent work focuses on combining meta-analytic results across predictors in order to estimate the incremental validity of one predictor over another. The new insight is that if one has meta-analytic estimates of the relationship between two or more predictors and a given criterion, one needs one additional piece of information, namely meta-analytic estimates of the correlation between the predictors. Given a complete predictor-criterion intercorrelation matrix, with each element in the matrix estimated by meta-analysis, one can estimate the validity of a composite of predictors as well as the incremental validity of one predictor over another. The prototypic study in this domain is Bobko et al.'s (1999) extension of previous work by Schmitt et al. (1997) examining cognitive ability, structured interviews, conscientiousness, and biodata as predictors of overall job performance. They reported considerable incremental validity when the additional predictors are used to supplement cognitive ability and relatively modest incremental validity of cognitive ability over the other three predictors. We note that this work focuses on observed validity coefficients, and if

some predictors (e.g., cognitive ability) are more range restricted than others, the validity of the restricted predictor will be underestimated and the incremental validity of the other predictors will be overestimated. Thus, as we address in more detail elsewhere in this review, careful attention to range restriction is important for future progress in this area.

Given this interest in incremental validity, there are a growing number of meta-analyses of intercorrelations among specific predictors, including interview-cognitive ability and interview-personality relationships (Cortina et al. 2000; Huffcutt et al. 2001b), cognitive ability-situational judgment test relationships (McDaniel et al. 2001), and personality-situational judgment tests relationships (McDaniel & Nguyen 2001). A range of primary studies has also examined the incremental contribution to validity of one or more newer predictors over one or more established predictors (e.g., Clevenger et al. 2001, Lievens et al. 2003, Mount et al. 2000). All of these efforts are aimed at a better understanding of the nomological network of relationships among predictors and dimensions of job performance. However, a limitation of many of these incremental validity studies is that they investigated whether methods (i.e., biodata, assessment center exercises) added incremental validity over and beyond constructs (i.e., cognitive ability, personality). Thus, these studies failed to acknowledge the distinction between content (i.e., constructs measured) and methods (i.e., the techniques used to measure the specific content). When constructs and methods are confounded, incremental validity results are difficult to interpret.

Individual Predictors

There is also meta-analytic work on the validity of individual predictors. One particularly important finding is a revisiting of the validity of work sample tests by Roth et al. (2005). Two meta-analytic estimates appeared in 1984: an estimate of 0.54 by Hunter & Hunter (1984) and an estimate of 0.32 by Schmitt et al. (1984), both corrected for criterion unreliability. The 0.54 value has subsequently been offered as evidence that work samples are the most valid predictor of performance yet identified. Roth et al. (2005) documented that the Hunter & Hunter (1984) estimate is based on a reanalysis of a questionable data source, and they report an updated meta-analysis that produces a mean validity of 0.33, highly similar to Schmitt et al.'s (1984) prior value of 0.32. Thus, the validity evidence for work samples remains positive, but the estimate of their mean validity needs to be revised downward.

Another important finding comes from a meta-analytic examination of the validity of global measures of conscientiousness compared to measures of four conscientiousness facets (achievement, dependability, order, and cautiousness) in the prediction of broad versus narrow criteria (Dudley et al. 2006). Dudley and colleagues reported that although broad conscientiousness measures predict all criteria studied (e.g., overall performance, task performance, job dedication, and counterproductive work behavior), in all cases validity was driven largely by the achievement and/or dependability facets, with relatively little contribution from cautiousness and order. Achievement receives the dominant weight in predicting task performance, whereas dependability receives the dominant weight in predicting job dedication and counterproductive work behavior. For job dedication and counterproductive work behavior, the narrow facets provided a dramatic increase in variance accounted for over global conscientiousness measures. This work sheds light on the issue of settings in which broad versus narrow trait measures are preferred, and it makes clear that the criterion of interest leads to different decisions as to the predictor of choice.

In the assessment center domain, Arthur et al. (2003) reported a meta-analysis of the validity of final dimension ratings. They focused on final dimension ratings instead of on the overall assessment rating

(OAR). Although the OAR is practically important, it is conceptually an amalgam of evaluations on a variety of dimensions in a diverse set of exercises. Several individual dimensions produced validities comparable to the OAR, and a composite of individual dimensions outperformed the OAR. Problem solving accounted for the most variance, followed by influencing others. In the cognitive ability domain, a cross-national team of researchers (Salgado et al. 2003a,b) reaffirmed U.S. findings of the generalizability of the validity of cognitive ability tests in data from seven European countries. Two meta-analyses addressed issues of "fit," with Arthur et al. (2006) focusing on person-organization fit and with Kristof-Brown et al. (2005) dealing more broadly with person-job, person-group, and person-organization fit. Both examined relationships with job performance and turnover, and both discussed the potential use of fit measures in a selection context. Correlations were modest, and there is virtually no information about the use of such measures in an actual selection context (with the exception of the interview; see Huffcutt et al. 2001a). Thus, this remains a topic for research rather than for operational use in selection.

IDENTIFICATION AND PREDICTION OF NEW OUTCOME VARIABLES

A core assumption in the selection paradigm is the relative stability of the job role against which the suitability of applicants is evaluated. However, rapidly changing organizational structures (e.g., due to mergers, downsizing, team-based work, or globalization) have added to job instability and have challenged personnel selection (Chan 2000, Kehoe 2000). As noted above, there has been a renewed interest in the notion of dynamic performance. In addition, a growing amount of studies have aimed to shed light on predictors (other than cognitive ability) related to the various dimensions of the higher-order construct of adaptability (Pulakos et al. 2000, 2002).

Creatively solving problems, dealing with uncertain work situations, cross-cultural adaptability, and interpersonal adaptability are adaptability dimensions that have been researched in recent years. With respect to the dimension of creatively solving problems, George & Zhou (2001) showed that creative behavior of employees was highest among those high on Openness to Experience and when the situation created enough opportunities for this trait to be manifested (e.g., unclear means, unclear ends, and positive feedback). Openness also played an important role in facilitating handling uncertain work situations. Judge et al. (1999) showed that Openness was related to coping with organizational change, which in turn was associated with job performance. Regarding cross-cultural adaptability, Lievens et al. (2003) found that cross-cultural training performance was predicted by Openness, cognitive ability, and assessment center ratings of adaptability, teamwork, and communication. Viewing desire to terminate an international assignment early as the converse of adaptability, Caligiuri (2000) found that Emotional Stability, Extraversion, and Agreeableness had significant negative relationships with desire to prematurely terminate the assignment in a concurrent validity study. Finally, interpersonal adaptability (measured by individual contextual performance of incumbents in team settings) was linked to structured interview ratings of social skills, Conscientiousness, Extraversion, and team knowledge (Morgeson et al. 2005).

Clearly, many of these results have practical ramifications for broadening and changing selection practices in new contexts. For instance, they suggest that a selection process for international personnel based on job knowledge and technical competence should be broadened. Yet, these studies also share some limitations, as they were typically concurrent studies with job incumbents. However, an even more important drawback is that they do not really examine predictors of change. They mostly examined different predictors in a new context. To be able to identify predictors of performance in a changing task and organization, it is necessary to include an assessment of people "unlearning" the old task and then "relearning" the new task. Along these lines, Le Pine et al. (2000) focused on adaptability in decision

making and found evidence of different predictors before and after the change. Prechange task performance was related only to cognitive ability, whereas adaptive performance (postchange) was positively related to both cognitive ability and Openness to Experience and negatively to Conscientiousness.

IMPROVED ABILITY TO ESTIMATE PREDICTOR-CRITERION RELATIONSHIPS

The prototypic approach to estimating predictor-criterion relationships in the personnel selection field is to (a) estimate the strength of the linear relationship by correlating the predictor with the criterion, (b) correct the resulting correlation for measurement error in the criterion measure, and (c) further correct the resulting correlation for restriction of range. [The order of (b) and (c) is reversed if the reliability estimate is obtained on an unrestricted sample rather than from the selected sample used in step (a).] There have been useful developments in these areas.

Linearity

First, although relationships between cognitive ability and job performance have been found to be linear (Coward & Sackett 1990), there is no basis for inferring that this will generalize to noncognitive predictors. In fact, one can hypothesize curvilinear relationships in the personality domain (e.g., higher levels of conscientiousness are good up to a point, with extremely high levels involving a degree of rigidity and inflexibility resulting in lower performance). Investigations into nonlinear relationships are emerging, with mixed findings [e.g., no evidence for nonlinear conscientiousness-performance relationships in research by Robie & Ryan (1999), but evidence of such relationships in research by LaHuis et al. (2005)]. Future research needs to attend to a variety of issues, including power to detect nonlinearity, the possibility that faking masks nonlinear effects, and the conceptual basis for positing departures from linearity for a given job-attribute combination.

Meta-Analysis

Second, an edited volume by Murphy (2003) brings together multiple perspectives and a number of new developments in validity generalization. Important new developments include the development of new maximum likelihood estimation procedures (Raju & Drasgow 2003) and empirical bayesian methods (Brannick & Hall 2003). These bayesian methods integrate meta-analytic findings with findings from a local study to produce a revised estimate of local validity. This is an important reframing: In the past, the question was commonly framed as, "Should I rely on meta-analytic findings or on a local validity estimate?" These bayesian methods formally consider the uncertainty in a local study and in meta-analytic findings and weight the two accordingly in estimating validity.

Range Restriction

Third, there have been important new insights into the correction of observed correlations for range restriction. One is the presentation of a taxonomy of ways in which range restriction can occur and of methods of correction (Sackett & Yang 2000). Eleven different range restriction scenarios are treated, expanding the issue well beyond the common distinction between direct and indirect restriction. Another is an approach to making range restriction corrections in the context of meta-analysis. Sackett et al. (2007) note that common practice in meta-analysis is to apply a direct range restriction correction to the mean observed intercorrelations among predictors, which in effect applies the same correction to each study. They offer an approach in which studies are categorized based on the type of restriction present (e.g., no restriction versus direct versus indirect), with appropriate corrections made within each category.

The most significant development regarding range restriction is Hunter et al.'s (2006) development of a new approach to correcting for indirect range restriction. Prior approaches are based on the assumption that the third variable on which selection is actually done is measured and available to the researcher. However, the typical circumstance is that selection is done on the basis of a composite of measured and unmeasured variables (e.g., unquantified impressions in an interview), and that this overall selection composite is unmeasured. Hunter et al. (2006) developed a correction approach that does not require that the selection composite is measured. Schmidt et al. (2006) apply this approach to meta-analysis, which has implicitly assumed direct range restriction, and show that applying a direct restriction correction when restriction is actually indirect results in a 21% underestimate of validity in a reanalysis of four existing meta-analytic data sets.

We also note that two integrative reviews of the use of synthetic validation methods have appeared (Scherbaum 2005, Steel et al. 2006), which offer the potential for increased use of this family of methods for estimating criterion-related validity.

IMPROVED UNDERSTANDING OF SUBGROUP DIFFERENCES, FAIRNESS, BIAS, AND THE LEGAL DEFENSIBILITY OF OUR SELECTION SYSTEMS

Group differences by race and gender remain important issues in personnel selection. The heavier the weight given in a selection process to a predictor on which group mean differences exist, the lower the selection rate for members of the lower-scoring group. However, a number of predictors that fare well in terms of rated job relevance and criterion-related validity produce substantial subgroup differences (e.g., in the domain of cognitive ability for race/ethnicity and in the domain of physical ability for gender). This results in what has been termed the validity-adverse impact tradeoff, as attempts to maximize validity tend to involve giving heavy weight to predictors on which group differences are found, and attempts to minimize group differences tend to involve giving little or no weight to some potentially valid predictors (Sackett et al. 2001). This creates a dilemma for organizations that value both a highly productive and diverse workforce. This also has implications for the legal defensibility of selection systems, as adverse impact resulting from the use of predictors on which differences are found is the triggering mechanism for legal challenges to selection systems. Thus, there is interest in understanding the magnitude of group differences that can be expected using various predictors and in finding strategies for reducing group differences in ways that do not compromise validity. Work in this area has been very active since 2000, with quite a number of important developments. Alternatives to the U.S. federal government's four-fifths rule for establishing adverse impact have been proposed, including a test for the significance of the adverse impact ratios (Morris & Lobsenz 2000) and pairing the adverse impact ratio with a significance test (Roth et al. 2006). The issue of determining minimum qualifications (e.g., the lowest score a candidate can obtain and still be eligible for selection) has received greater attention because of court rulings (Kehoe & Olson 2005).

Subgroup Mean Differences

There have been important efforts at consolidating what is known about the magnitude of subgroup differences on various predictors. A major review by Hough et al. (2001) summarized the evidence for differences by race/ethnicity, gender, and age for a broad range of predictors, including cognitive abilities, personality, physical ability, assessment centers, biodata, interviews, and work samples. One theme emerging from that review is that there is considerable variation within subfacets of a given construct. For example, racial group differences on a number of specific abilities are smaller than differences on general

cognitive ability, and race and gender differences vary within subfacets of the Big Five personality dimensions. A more focused review by Roth et al. (2001a) focused on differences by race/ethnicity on measures of cognitive ability. Roth and colleagues (2001a) add considerable nuance to the often-stated summary finding of white-black standardized mean differences of about 1.0 SD, noting (a) larger differences in applicant samples than incumbent samples, (b) larger differences in broad, pooled samples than in job-specific samples, and (c) larger differences in applicant samples for low-complexity jobs than for high-complexity jobs. The effects of range restriction mechanisms on subgroup differences were further explored by Roth et al. (2001b) in the context of multistage selection systems. Additional studies examined mean differences on other predictors, such as grade point average (Roth & Bobko 2000), educational attainment (Berry et al. 2006), and structured interviews (Roth et al. 2002). Two meta-analyses examined race differences in performance measures (McKay & McDaniel 2006, Roth et al. 2003); both reported overall uncorrected white-black mean differences of about 0.25 SD.

Mechanisms for Reducing Differences

There is new insight into several hypothesized mechanisms for reducing subgroup differences. Sackett et al. (2001) reviewed the cumulative evidence and concluded that several proposed mechanisms are not, in fact, effective in reducing differences. These include (a) using differential item functioning analysis to identify and remove items functioning differently by subgroup, (b) providing coaching programs (these may improve scores for all, but group differences remain), (c) providing more generous time limits (which appears to increase group differences), and (d) altering test taking motivation. The motivational approach receiving most attention is the phenomenon of stereotype threat (Steele et al. 2002). Although this research shows that the way a test is presented to students in laboratory settings can affect their performance, the limited research in employment settings does not produce findings indicative of systematic effects due to stereotype threat (Cullen et al. 2004, 2006). Finally, interventions designed to reduce the tendency of minority applicants to withdraw from the selection process are also not a viable approach for reducing subgroup differences because they were found to have small effects on the adverse impact of selection tests (Tam et al. 2004).

Sackett et al. (2001) did report some support for expanding the criterion as a means of reducing subgroup differences. The relative weight given to task, citizenship, and counterproductive behavior in forming a composite criterion affects the weight given to cognitive predictors. Sackett et al. cautioned against differential weighting of criteria solely as a means of influencing predictor subgroup differences; rather, they argued that criterion weights should reflect the relative emphasis the organization concludes is appropriate given its business strategy. They also reported some support for expanding the range of predictors used. The strategy of supplementing existing cognitive predictors with additional predictors outside the cognitive domain can reduce the overall subgroup differences in some circumstances. This strategy has received considerable attention because broadening the range of predictors has the potential to both reduce subgroup differences and increase validity.

There has been considerable activity regarding test score banding, including an integrative review featuring competing perspectives (Campion et al. 2001) and an edited volume (Aguinis 2004). Although banding is not advocated only as a device for reducing adverse impact, the potential for impact reduction is a key reason for the interest in banding. A clearer picture is emerging of the circumstances under which banding does or does not affect minority-hiring rates, with key features including the width of the band and the basis for selection within a band.

Forecasting Validity and Adverse Impact

New methods exist for forecasting the likely effects of various ways of combining multiple predictors on subsequent performance and on adverse impact. Although combining predictors via multiple regression is statistically optimal in terms of predicting performance, there may be alternative ways of combining that fare better in terms of adverse impact at what is judged to be an acceptably small reduction in validity. De Corte et al. (2007) applied the concept of pareto optimality and provided a computer program that shows the set of predictor weights that give the lowest possible degree of subgroup difference at any given degree of reduction in validity. In other words, the procedure estimates the reduction in subgroup differences that would be attainable should the decision maker be willing to accept, say, a 1%, or a 5%, or a 10% reduction in validity. Thus, it makes the validity-diversity tradeoff very explicit. In another study, De Corte et al. (2006) offered a computer program for examining the effects of different ways of sequencing predictors in a multistage selection system to achieve intended levels of workforce quality, workforce diversity, and selection cost. Also, Aguinis & Smith (2007) offered a program for examining the effect of the choice of selection ratio on mean criterion performance and adverse impact.

Differential Prediction

New methods have been developed for examining differential prediction (i.e., differences in slopes and intercepts between subgroups). Johnson et al. (2001) applied the logic of synthetic validity to pool data across jobs, thus making such analyses feasible in settings where samples within jobs are too small for adequate power. Sackett et al. (2003) showed that omitted variables that are correlated with both subgroup membership and the outcome of interest can bias attempts to estimate slope and intercept differences and offer strategies for addressing the omitted variables problem.

IMPROVED ADMINISTRATIVE EASE WITH WHICH SELECTION SYSTEMS CAN BE USED

To increase the efficiency and consistency of test delivery, many organizations have implemented Internet technology in their selection systems. Benefits of Internet-based selection include cost and time savings because neither the employer nor the applicants have to be present at the same location. Further, organizations' access to larger and more geographically diverse applicant pools is expanded. Finally, it might give organizations a "high-tech" image.

Lievens & Harris (2003) reviewed current research on Internet recruiting and testing. They concluded that most research has focused on either applicants' reactions or measurement equivalence with traditional paper-and-pencil testing. Two forms of the use of the Internet in selection have especially been investigated, namely proctored Internet testing and videoconference interviewing.

With respect to videoconferencing interviews (and other technology-mediated interviews such as telephone interviews or interactive voice-response telephone interviews), there is evidence that their increased efficiency might also lead to potential drawbacks as compared with face-to-face interviews (Chapman et al. 2003). Technology-mediated interviews might result in less favorable reactions and loss of potential applicants. However, it should be emphasized that actual job pursuit behavior was not examined.

The picture for Internet-based testing is somewhat more positive. With regard to noncognitive measures, the Internet-based format generally leads to lower means, larger variances, more normal distributions, and larger internal consistencies. The only drawback seems to be the somewhat higher-scale intercorrelations

(Ployhart et al. 2003). In within-subjects designs (Potosky & Bobko 2004), similar acceptable cross-mode correlations for noncognitive tests were found. However, this is not the case for timed tests. For instance, cross-mode equivalence of a timed spatial-reasoning test was as low as 0.44 (although there were only 30 minutes between the two administrations). On the one hand, the loading speed inherent in Internet-based testing might make the test different from its paper-and-pencil counterpart. In the Internet format, candidates also cannot start by browsing through the test to gauge the time constraints and type of items (Potosky & Bobko 2004, Richman et al. 1999). On the other hand, the task at hand (spatial reasoning) is also modified by the administration format change because it is not possible to make marks with a pen.

One limitation of existing Internet-based selection research is that explanations are seldom provided for why equivalence was or was not established. At a practical level, the identification of conditions that moderate measurement equivalence would also be insightful (see the aforementioned example of the spatial-reasoning test). More fundamentally, we believe that the current research on Internet testing is essentially conservative. Although an examination of equivalence is of key psychometric and legal importance, it does not advance our understanding of the new test administration format. That is, adapting traditional tests to the new technology is different from using the new technology to change existing tests/test administration and to enhance prediction. So, equivalence research per definition does not take the opportunity to improve the quality of assessment. Roznowski et al. (2000) offered an illustration of the use of cognitive processing measures that explicitly build on the possibilities of computerization to go beyond the type of measurement possible with paper-and-pencil testing and show incremental validity over a general cognitive measure in predicting training performance.

Unproctored Internet testing is a controversial example of the Internet radically changing the test administration process. Unproctored Internet testing might lead to candidate identification and test security concerns. Although test security problems might be partly circumvented by item banking and item-generation techniques (Irvine & Kyllonen 2002), user identification seems to be a deadlock (unless sophisticated techniques such as retinal scanning become widely available). To date, there seems to be relative consensus that unproctored testing is advisable only in low-stakes selection (Tippins et al. 2006). However, empirical evidence about the equivalence of proctored and unproctored testing in a variety of contexts is lacking.

Finally, it is striking that no evidence is available as to how Internet-based administration affects the utility of the selection system. So, we still do not know whether Internet selection affects the quantity and quality of the applicant pool and the performance of the people hired. However, utility studies of Internet-based selection seem necessary as recent surveys show that technology-based solutions are not always a panacea for organizations (e.g., Chapman & Webster 2003). Frequently mentioned complaints included the decreasing quality of the applicant pool, the huge dependency on a costly and ever-changing technology, and a loss of personal touch.

IMPROVED MEASUREMENT OF AND INSIGHT INTO CONSEQUENCES OF APPLICANT REACTIONS

Consequences of Applicant Reactions

Since the early 1990s, a growing number of empirical and theoretical studies have focused on applicants' perceptions of selection procedures, the selection process, and the selection decision, and their effects on individual and organizational outcomes. Hausknecht et al.'s (2004) meta-analysis found that perceived procedural characteristics (e.g., face validity, perceived predictive validity) had moderate relationships

with applicant perceptions. Person characteristics (e.g., age, gender, ethnic background, personality) showed near-zero correlations with applicant perceptions. In terms of selection procedures, work samples and interviews were perceived more favorably than were cognitive ability tests, which were perceived more positively than personality inventories.

This meta-analysis also yielded conclusions that raise some doubts about the added value of the field of applicant reactions. Although applicant perceptions clearly show some link with self-perceptions and applicants' intentions (e.g., job offer acceptance intentions), evidence for a relationship between applicant perceptions and actual behavioral outcomes was meager and disappointing. In fact, in the meta-analysis, there were simply too few studies to examine behavioral outcomes (e.g., applicant withdrawal, job performance, job satisfaction, and organizational citizenship behavior). Looking at primary studies, research shows that applicant perceptions play a minimal role in actual applicant withdrawal (Ryan et al. 2000, Truxillo et al. 2002). This stands in contrast with introductions to articles about applicant reactions that typically claim that applicant reactions have important individual and organizational outcomes. So, in applicant perception studies it is critical to go beyond self-reported outcomes (see also Chan & Schmitt 2004).

Methodological Issues

The Hausknecht et al. (2004) meta-analysis also identified three methodological factors that moderated the results found. First, monomethod variance was prevalent, as the average correlations were higher when both variables were measured simultaneously than when they were separated in time. Indeed, studies that measured applicant reactions longitudinally at different points in time (e.g., Schleicher et al. 2006, Truxillo et al. 2002, Van Vianen et al. 2004) are scarce and demonstrate that reactions differ contingent upon the point in the selection process. For example, Schleicher et al. (2006) showed that opportunity to perform became an even more important predictor of overall procedural fairness after candidates received negative feedback. Similarly, Van Vianen et al. (2004) found that pre-feedback fairness perceptions were affected by different factors than were post-feedback fairness perceptions. Second, large differences between student samples and applicant samples were found. Third, correlations differed between hypothetical and authentic contexts. The meta-analysis showed that the majority of applicant reactions studies were not conducted with actual applicants (only 36.0%), in the field (only 48.8%), and in authentic contexts (only 38.4%); thus, these methodological factors suggest that some of the relationships found in the meta-analysis might be either under- or overestimated (depending on the issue at hand). Even among actual applicants, it is important that the issue examined is meaningful to applicants. This is nicely illustrated by Truxillo & Bauer (1999). They investigated applicants' reactions to test score banding in three separate actual applicant samples. Race differences in applicants' reactions to banding were found only in a sample wherein participants were really familiar with banding.

Influencing Applicant Reactions

Despite these critical remarks, the field of applicant reactions also made progress. New ways of obtaining more favorable applicant reactions were identified. In a longitudinal study, Truxillo et al. (2002) demonstrated that the provision of information to candidates prior to the selection process might be a practical and inexpensive vehicle to improve applicant reactions. Applicants who were given information about a video-based test perceived this test as fairer both at the time of testing and one month later, upon receiving their test results. However, more distal behavioral measures were not affected by the pretest information. The provision of an explanation for selection decisions was identified as another practical intervention for promoting selection procedure fairness (Gilliland et al. 2001, Ployhart et al. 1999). Although no one ideal explanation feature to reduce applicants' perceptions of unfairness was identified,

explanations seemed to matter. It was noteworthy that Gilliland et al. (2001) even found evidence for a relationship between the type of explanation provided and actual reapplication behavior of applicants of a tenure-track faculty position.

Measurement of Applicant Reactions

Another important positive development in this field is improved measurement of applicants' perceptions and attitudes. Unidimensional and study-specific measures were replaced by newer multidimensional and theory-driven measures that have the potential to be used across many studies. Three projects were most noteworthy. First, Bauer et al. (2001) developed the selection procedural justice scale. This scale was based on procedural justice theory and assessed 11 procedural justice rules. Second, Sanchez et al. (2000) used expectancy theory to develop a multifaceted measure of test motivation, called the Valence, Instrumentality, and Expectancy Motivation Scale. This measure proved to be a more theory-driven way of structuring and measuring the construct of test motivation as compared with the extant unidimensional motivation scale of the Test Attitude Scale (Arvey et al. 1990). Third, McCarthy & Goffin (2004) undertook a similar effort as they tried to improve on the unidimensional test-anxiety subscale of the Test Attitude Scale (Arvey et al. 1990). Specifically, they focused on anxiety in employment interviews and developed the Measure of Anxiety in Selection Interviews. To this end, McCarthy & Goffin (2004) borrowed on separate streams of anxiety research and conceptualized interview anxiety as consisting of five dimensions: communication anxiety, appearance anxiety, social anxiety, performance anxiety, and behavioral anxiety. Results confirmed that this context-specific multidimensional anxiety measure had a consistent negative relationship with interview performance and explained additional variance over and above noncontextualized anxiety scales.

In short, the field of applicant reactions has made strides forward in terms of better measuring applicant reactions as several multidimensional and theory-driven improvements over existing measures were developed. Some progress was also made in terms of devising ways of obtaining more favorable applicant reactions (i.e., through the use of pretest information and posttest explanations). Yet, we highlighted the meager evidence of a relationship between applicant perceptions and key individual and organizational consequences (e.g., actual withdrawal from the selection process, test performance, job satisfaction, and organizational citizenship behavior) as the Achilles heel of this field.

IMPROVED DECISION-MAKER ACCEPTANCE OF SELECTION SYSTEMS

Research findings outlined above have applied value only if they find inroads in organizations. However, this is not straightforward, as psychometric quality and legal defensibility are only some of the criteria that organizations use in selection practice decisions. Given that sound selection procedures are often either not used or are misused in organizations (perhaps the best known example being structured interviews), we need to better understand the factors that might impede organizations' use of selection procedures.

Apart from broader legal, economic, and political factors, some progress in uncovering additional factors was made in recent years. One factor identified was the lack of knowledge/awareness of specific selection procedures. For instance, the two most widely held misconceptions about research findings among HR professionals are that conscientiousness and values both are more valid than general mental ability in predicting job performance (Rynes et al. 2002). An interesting complement to Rynes et al's (2002) examination of beliefs of HR professionals was provided by Murphy et al.'s (2003) survey of I/O psychologists regarding their beliefs about a wide variety of issues regarding the use of cognitive ability

measures in selection. I/O psychologists are in relative agreement that such measures have useful levels of validity, but in considerable disagreement about claims that cognitive ability is the most important individual-difference determinant of job and training performance.

In addition, use of structured interviews is related to participation in formal interviewer training (Chapman & Zweig 2005, Lievens & De Paepe 2004). Another factor associated with selection practice use was the type of work practices of organizations. Organizations use different types of selection methods contingent upon the nature of the work being done (skill requirements), training, and pay level (Wilk & Cappelli 2003). Finally, we also gained some understanding of potential operating factors in the international selection area. In that context, the issue of gaining acceptance for specific selection procedures is even more complicated due to tensions between corporate requirements of streamlined selection practices and local desires of customized ones. A 20-country study showed that national differences accounted for considerable variance in selection practice, whereas differences grounded in cultural values (uncertainty avoidance and power distance) explained only some of the variability (Ryan et al. 1999).

Taken together, this handful of studies produced a somewhat better understanding of potential factors (e.g., knowledge, work practices, and national differences) related to acceptance of selection procedures. Yet, there is still a long way to go. All of these studies were descriptive accounts. We need prescriptive studies that produce specific strategies for gaining acceptance of selection practices or successfully introducing new ones. Along these lines, Muchinsky (2004) presented an interesting case study wherein he used a balancing act (combining strategies of education, shared responsibility, negotiation, respect, and recognition of available knowledge of all stakeholders) to successfully implement psychometrically straightforward test development principles of a job knowledge test in an organizational context.

CONCLUSION

We opened with a big question: "Can we do a better job of selection today than in 2000?" Our sense is that we have made substantial progress in our understanding of selection systems. We have greatly improved our ability to predict and model the likely outcomes of a particular selection system, as a result of developments such as more and better meta-analyses, better insight into incremental validity, better range restriction corrections, and better understanding of validity-adverse impact tradeoffs. Thus, someone well informed about the research base is more likely to attend carefully to determining the criterion constructs of interest to the organization, more likely to select trial predictors with prior conceptual and empirical links to these criteria, more likely to select predictors with incremental validity over one another, and less likely to misestimate the validity of a selection system due to use of less-than-optimal methods of estimating the strength of predictor-criterion relationships.

We have identified quite a number of promising leads with the potential to improve the magnitude of predictor-criterion relationships should subsequent research support initial findings. These include contextualization of predictors and the use of implicit measures. We also have new insights into new outcomes and their predictability (e.g., adaptability), better understanding of the measurement and consequences of applicant reactions, and better understanding of impediments of selection system use (e.g., HR manager misperceptions about selection systems). Overall, relative to a decade ago, at best we are able to modestly improve validity at the margin, but we are getting much better at modeling and predicting the likely outcomes (validity, adverse impact) of a given selection system.

LITERATURE CITED

Aguinis H, ed. 2004. Test-Score Banding in Human Resource Selection: Legal, Technical, and Societal Issues. Westport, CT: Praeger

Aguinis H, Smith MA. 2007. Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. Pers. Psychol. 60:165–99

Am. Educ. Res. Assoc., Am. Psychol. Assoc., Natl. Counc. Meas. Educ. 1999. Standards for Educational and Psychological Testing. Washington, DC: Am. Educ. Res. Assoc.

Anderson N, Ones DS, Sinangil HK, Viswesvaran C, eds. 2001. Handbook of Industrial, Work and Organizational Psychology: Personality Psychology. Thousand Oaks, CA: Sage

Arthur W, Bell ST, Villado AJ, Doverspike D. 2006. The use of person-organization fit in employment decision making: an assessment of its criterion-related validity. J. Appl. Psychol. 91:786–801

Arthur W, Day EA, McNelly TL, Edens PS. 2003. A meta-analysis of the criterion-related validity of assessment center dimensions. Pers. Psychol. 56:125–54

Arthur W, Woehr DJ, Maldegen R. 2000. Convergent and discriminant validity of assessment center dimensions: a conceptual and empirical reexamination of the assessment confer construct-related validity paradox. J. Manag. 26:813–35

Arvey RD, Strickland W, Drauden G, Martin C. 1990. Motivational components of test taking. Pers. Psychol. 43:695–716

Barrick MR, Parks L, Mount MK. 2005. Self-monitoring as a moderator of the relationships between personality traits and performance. Pers. Psychol. 58:745–67

Barrick MR, Patton GK, Haugland SN. 2000. Accuracy of interviewer judgments of job applicant personality traits. Pers. Psychol. 53:925–51

Barrick MR, Ryan AM, eds. 2003. Personality and Work: Reconsidering the Role of Personality in Organizations. San Francisco, CA: Jossey-Bass

Barrick MR, Stewart GL, Piotrowski M. 2002. Personality and job performance: test of the mediating effects of motivation among sales representatives. J. Appl. Psychol. 87:43–51

Bartram D. 2005. The great eight competencies: a criterion-centric approach to validation. J. Appl. Psychol. 90:1185–203

Bauer TN, Truxillo DM, Sanchez RJ, Craig JM, Ferrara P, Campion MA. 2001. Applicant reactions to selection: development of the Selection Procedural Justice Scale (SPJS). Pers. Psychol. 54:387–419

Berry CM, Gruys ML, Sackett PR. 2006. Educational attainment as a proxy for cognitive ability in selection: effects on levels of cognitive ability and adverse impact. J. Appl. Psychol. 91:696–705

Bing MN, Whanger JC, Davison HK, Van Hook JB. 2004. Incremental validity of the frame-of-reference effect in personality scale scores: a replication and extension. J. Appl. Psychol. 89:150–57

Bobko P, Roth PL, Potosky D. 1999. Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. Pers. Psychol. 52:561–89

Borman WC, Buck DE, Hanson MA, Motowidlo SJ, Stark S, Drasgow F. 2001. An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. J. Appl. Psychol. 86:965–73

Borman WC, Ilgen DR, Klimoski RJ, Weiner IB, eds. 2003. Handbook of Psychology. Vol. 12: Industrial and Organizational Psychology. Hoboken, NJ: Wiley

Borman WC, Motowidlo SJ. 1997. Organizational citizenship behavior and contextual performance. Hum. Perform. 10:67–69

Bowler MC, Woehr DJ. 2006. A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. J. Appl. Psychol. 91:1114–24

Brannick MT, Hall SM. 2003. Validity generalization: a Bayesian perspective. See Murphy 2003, pp. 339–64

Caligiuri PM. 2000. The Big Five personality characteristics as predictors of expatriate's desire to terminate the assignment and supervisor-rated performance. Pers. Psychol. 53:67–88

Campbell JP, Knapp DJ, eds. 2001. Exploring the Limits in Personnel Selection and Classification. Mahwah, NJ: Erlbaum

Campbell JP, McCloy RA, Oppler SH, Sager CE. 1993. A theory of performance. In Personnel Selection in Organizations, ed. N Schmitt, WC Borman, pp. 35–70. San Francisco, CA: Jossey Bass

Campion MA, Outtz JL, Zedeck S, Schmidt FL, Kehoe JF, et al. 2001. The controversy over score banding in personnel selection: answers to 10 key questions. Pers. Psychol. 54:149–85

Chan D. 2000. Understanding adaptation to changes in the work environment: integrating individual difference and learning perspectives. Res. Pers. Hum. Resour. Manag. 18:1–42

Chan D, Schmitt N. 2002. Situational judgment and job performance. Hum. Perform. 15:233–54

Chan D, Schmitt N. 2004. An agenda for future research on applicant reactions to selection procedures: a construct-oriented approach. Int. J. Select. Assess. 12:9–23

Chapman DS, Uggerslev KL, Webster J. 2003. Applicant reactions to face-to-face and technology-mediated interviews: a field investigation. J. Appl. Psychol. 88:944–53

Chapman DS, Webster J. 2003. The use of technologies in the recruiting, screening, and selection processes for job candidates. Int. J. Select. Assess. 11:113–20

Chapman DS, Zweig DI. 2005. Developing a nomological network for interview structure: antecedents and consequences of the structured selection interview. Pers. Psychol. 58:673–702

Clevenger J, Pereira GM, Wiechmann D, Schmitt N, Harvey VS. 2001. Incremental validity of situational judgment tests. J. Appl. Psychol. 86:410–17

Cortina JM, Goldstein NB, Payne SC, Davison HK, Gilliland SW. 2000. The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. Pers. Psychol. 53:325–51

Cote S, Miners CTH. 2006. Emotional intelligence, cognitive intelligence, and job performance. Adm. Sci. Q. 51:1–28

Coward WM, Sackett PR. 1990. Linearity of ability performance relationships—a reconfirmation. J. Appl. Psychol. 75:297–300

Cullen MJ, Hardison CM, Sackett PR. 2004. Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. J. Appl. Psychol. 89:220–30

Cullen MJ, Waters SD, Sackett PR. 2006. Testing stereotype threat theory predictions for math-identified and nonmath-identified students by gender. Hum. Perform. 19:421–40

Dalal RS. 2005. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. J. Appl. Psychol. 90:1241–55

De Corte W, Lievens F, Sackett PR. 2006. Predicting adverse impact and multistage mean criterion performance in selection. J. Appl. Psychol. 91:523–37

De Corte W, Lievens F, Sackett PR. 2007. Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. J. Appl. Psychol. In press Dipboye RL, Colella A, eds. 2004. Discrimination at Work: The Psychological and Organizational Bases. Mahwah, NJ: Erlbaum

Dudley NM, Orvis KA, Lebiecki JE, Cortina JM. 2006. A meta-analytic investigation of conscientiousness

in the prediction of job performance: examining the intercorrelations and the incremental validity of narrow traits. J. Appl. Psychol. 91:40–57

Dwight SA, Donovan JJ. 2003. Do warnings not to fake reduce faking? Hum. Perform. 16:1–23

Ellingson JE, Sackett PR, Hough LM. 1999. Social desirability corrections in personality measurement: issues of applicant comparison and construct validity. J. Appl. Psychol. 84:55–166

Ellingson JE, Smith DB, Sackett PR. 2001. Investigating the influence of social desirability on personality factor structure. J. Appl. Psychol. 86:122–33

George JM, Zhou J. 2001. When openness to experience and conscientiousness are related to creative behavior: an interactional approach. J. Appl. Psychol. 86:513–24

Gilliland SW, Groth M, Baker RC, Dew AF, Polly LM, Langdon JC. 2001. Improving applicants' reactions to rejection letters: an application of fairness theory. Pers. Psychol. 54:669–703

Goffin RD, Christiansen ND. 2003. Correcting personality tests for faking: a review of popular personality tests and an initial survey of researchers. Int. J. Select. Assess. 11:340–44

Gottfredson LS. 2003. Dissecting practical intelligence theory: its claims and evidence. Intelligence 31:343–97

Hausknecht JP, Day DV, Thomas SC. 2004. Applicant reactions to selection procedures: an updated model and meta-analysis. Pers. Psychol. 57:639–83

Hausknecht JP, Halpert JA, Di Paolo NT, Moriarty MO. 2007. Retesting in selection: a metaanalysis of practice effects for tests of cognitive ability. J. Appl. Psychol. 92:373–85

Hausknecht JP, Trevor CO, Farr JL. 2002. Retaking ability tests in a selection setting: implications for practice effects, training performance, and turnover. J. Appl. Psychol. 87:243–54

Heggestad ED, Morrison M, Reeve CL, McCloy RA. 2006. Forced-choice assessments of personality for selection: evaluating issues of normative assessment and faking resistance. J. Appl. Psychol. 91:9–24

Hogan J, Holland B. 2003. Using theory to evaluate personality and job-performance relations: a socioanalytic perspective. J. Appl. Psychol. 88:100–12

Hough LM, Oswald FL. 2000. Personnel selection: looking toward the future—remembering the past. Annu. Rev. Psychol. 51:631–64

Hough LM, Oswald FL. 2005. They're right . . . well, mostly right: research evidence and an agenda to rescue personality testing from 1960's insights. Hum. Perform. 18:373–87

Hough LM, Oswald FL, Ployhart RE. 2001. Determinants, detection and amelioration of adverse impact in personnel selection procedures: issues, evidence and lessons learned. Int. J. Select. Assess. 9:152–94

Huffcutt AI, Conway JM, Roth PL, Stone NJ. 2001a. Identification and meta-analytic assessment of psychological constructs measured in employment interviews. J. Appl. Psychol. 86:897–913

Huffcutt AI, Weekley JA, Wiesner WH, Degroot TG, Jones C. 2001b. Comparison of situational and behavior description interview questions for higher-level positions. Pers. Psychol. 54:619–44

Hunter JE, Hunter RF. 1984. Validity and utility of alternative predictors of job performance. Psychol. Bull. 96:72–98

Hunter JE, Schmidt FL, Le H. 2006. Implications of direct and indirect range restriction for meta-analysis methods and findings. J. Appl. Psychol. 91:594–612

Hunthausen JM, Truxillo DM, Bauer TN, Hammer LB. 2003. A field study of frame-of-reference effects on personality test validity. J. Appl. Psychol. 88:545–51

Hurtz GM, Donovan JJ. 2000. Personality and job performance: the Big Five revisited. J. Appl. Psychol. 85:869–79

Int. Test Comm. (ITC). 2006. International guidelines on computer-based and Internet-delivered testing. http://www.intestcom.org. Accessed March 2, 2007

Irvine SH, Kyllonen PC, eds. 2002. Item Generation for Test Development. Mahwah, NJ: Erlbaum

James LR, McIntyre MD, Glisson CA, Green PD, Patton TW, et al. 2005. A conditional reasoning measure for aggression. Organ. Res. Methods 8:69–99

Jeanneret PR. 2005. Professional and technical authorities and guidelines. See Landy 2005, pp. 47–100

Johnson JW, Carter GW, Davison HK, Oliver DH. 2001. A synthetic validity approach to testing differential prediction hypotheses. J. Appl. Psychol. 86:774–80

Judge TA, Thoresen CJ, Pucik V, Welbourne TM. 1999. Managerial coping with organizational change: a dispositional perspective. J. Appl. Psychol. 84:107–22

Kehoe JF, ed. 2000. Managing Selection in Changing Organizations: Human Resource Strategies. San Francisco, CA: Jossey-Bass

Kehoe JF, Olson A. 2005. Cut scores and employment discrimination litigation. See Landy 2005, pp. 410–49

Kolk NJ, Born MP, van der Flier H. 2002. Impact of common rater variance on construct validity of assessment center dimension judgments. Hum. Perform. 15:325–37

Kristof-Brown AL, Zimmerman RD, Johnson EC. 2005. Consequences of individuals' fit at work: a meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. Pers. Psychol. 58:281–342

LaHuis DM, Martin NR, Avis JM. 2005. Investigating nonlinear conscientiousness-job performance relations for clerical employees. Hum. Perform. 18:199–212

Lance CE, Lambert TA, Gewin AG, Lievens F, Conway JM. 2004. Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. J. Appl. Psychol. 89:377–85

Lance CE, Newbolt WH, Gatewood RD, Foster MR, French NR, Smith DE. 2000. Assessment center exercise factors represent cross-situational specificity, not method bias. Hum. Perform. 13:323–53

Landy FJ, ed. 2005a. Employment Discrimination Litigation: Behavioral, Quantitative, and Legal Perspectives. San Francisco, CA: Jossey Bass

Landy FJ. 2005b. Some historical and scientific issues related to research on emotional intelligence. J. Organ. Behav. 26:411–24

LeBreton JM, Barksdale CD, Robin JD, James LR. 2007. Measurement issues associated with conditional reasoning tests of personality: deception and faking. J. Appl. Psychol. 92:1–16

LePine JA, Colquitt JA, Erez A. 2000. Adaptability to changing task contexts: effects of general cognitive ability conscientiousness, and openness to experience. Pers. Psychol. 53:563–93

Lievens F. 2001. Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. J. Appl. Psychol. 86:255–64

Lievens F. 2002. Trying to understand the different pieces of the construct validity puzzle of assessment centers: an examination of assessor and assessee effects. J. Appl. Psychol. 87:675–86

Lievens F, Buyse T, Sackett PR. 2005a. The operational validity of a video-based situational judgment test for medical college admissions: illustrating the importance of matching

predictor and criterion construct domains. J. Appl. Psychol. 90:442–52

Lievens F, Buyse T, Sackett PR. 2005b. The effects of retaking tests on test performance and validity: an examination of cognitive ability, knowledge, and situational judgment tests. Pers. Psychol. 58:981–1007

Lievens F, Chasteen CS, Day EA, Christiansen ND. 2006. Large-scale investigation of the role

of trait activation theory for understanding assessment center convergent and discriminant validity. J. Appl. Psychol. 91:247–58

Lievens F, Conway JM. 2001. Dimension and exercise variance in assessment center scores: a large-scale evaluation of multitrait-multimethod studies. J. Appl. Psychol. 86:1202–22

Lievens F, De Paepe A. 2004. An empirical investigation of interviewer-related factors that discourage the use of high structure interviews. J. Organ. Behav. 25:29–46

Lievens F, Harris MM. 2003. Research on Internet recruiting and testing: current status and future directions. In International Review of Industrial and Organizational Psychology, ed. CL Cooper, IT Robertson, pp. 131–65. Chichester, UK: Wiley

Lievens F, Harris MM, Van Keer E, Bisqueret C. 2003. Predicting cross-cultural training performance: the validity of personality, cognitive ability, and dimensions measured by an assessment center and a behavior description interview. J. Appl. Psychol. 88:476–89

Lievens F, Sackett PR. 2006. Video-based vs. written situational judgment tests: a comparison in terms of predictive validity. J. Appl. Psychol. 91:1181–88

Matthews G, Roberts RD, Zeidner M. 2004. Seven myths about emotional intelligence. Psychol. Inq. 15:179–96

Mayer JD, Roberts RD, Barsade SG. 2007. Emerging research in emotional intelligence. Annu. Rev. Psychol. 2008: In press

McCarthy J, Goffin R. 2004. Measuring job interview anxiety: beyond weak knees and sweaty palms. Pers. Psychol. 57:607–37

McDaniel MA, Hartman NS, Whetzel DL, Grubb WL III. 2007. Situational judgment tests, response instructions and validity: a meta-analysis. Pers. Psychol. 60:63–91

McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. 2001. Use of situational judgment tests to predict job performance: a clarification of the literature. J. Appl. Psychol. 86:730–40

McDaniel MA, Nguyen NT. 2001. Situational judgment tests: a review of practice and constructs assessed. Int. J. Select. Assess. 9:103–13

McDaniel MA, Whetzel DL. 2005. Situational judgment test research: informing the debate on practical intelligence theory. Intelligence 33:515–25

McFarland LA, Yun GJ, Harold CM, Viera L, Moore LG. 2005. An examination of impression management use and effectiveness across assessment center exercises: the role of competency demands. Pers. Psychol. 58:949–80

McKay PF, McDaniel MA. 2006. A reexamination of black-white mean differences in work performance: more data, more moderators. J. Appl. Psychol. 91:538–54

Middendorf CH, Macan TH. 2002. Note-taking in the employment interview: effects on recall and judgments. J. Appl. Psychol. 87:293–303

Morgeson FP, Reider MH, Campion MA. 2005. Selecting individuals in team settings: the importance of social skills, personality characteristics, and teamwork knowledge. Pers. Psychol. 58:583–611

Morris SB, Lobsenz RE. 2000. Significance tests and confidence intervals for the adverse impact ratio. Pers. Psychol. 53:89–111

Motowidlo SJ, Dunnette MD, Carter GW. 1990. An alternative selection procedure: the low-fidelity simulation. J. Appl. Psychol. 75:640–47

Motowidlo SJ, Hooper AC, Jackson HL. 2006. Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. J. Appl. Psychol. 91:749–61

Mount MK, Witt LA, Barrick MR. 2000. Incremental validity of empirically keyed biodata scales over GMA and the five factor personality constructs. Pers. Psychol. 53:299–323

Muchinsky PM. 2004. When the psychometrics of test development meets organizational realities: a conceptual framework for organizational change, examples, and recommendations. Pers. Psychol. 7:175–209

Mueller-Hanson R, Heggestad ED, Thornton GC. 2003. Faking and selection: considering the use of personality from select-in and select-out perspectives. J. Appl. Psychol. 88:348–55

Murphy KR. 1989. Is the relationship between cognitive ability and job performance stable over time? Hum. Perform. 2:183–200

Murphy KR, ed. 2003. Validity Generalization: A Critical Review. Mahwah, NJ: Erlbaum

Murphy KR, ed. 2006. A Critique of Emotional Intelligence: What Are the Problems and How Can They Be Fixed? Mahwah, NJ: Erlbaum

Murphy KR, Cronin BE, Tam AP. 2003. Controversy and consensus regarding the use of cognitive ability testing in organizations. J. Appl. Psychol. 88:660–71

Murphy KR, Dzieweczynski JL. 2005. Why don't measures of broad dimensions of personality perform better as predictors of job performance? Hum. Perform. 18:343–57

Naglieri JA, Drasgow F, Schmit M, Handler L, Prifitera A, et al. 2004. Psychological testing on the Internet: new problems, old issues. Am. Psychol. 59:150–62

Nguyen NT, Biderman MD, McDaniel MA. 2005. Effects of response instructions on faking a situational judgment test. Int. J. Select. Assess. 13:250–60

Ones DS, Viswesvaran C, Dilchert S. 2005. Personality at work: raising awareness and correcting misconceptions. Hum. Perform. 18:389–404

Oswald FL, Schmit N, Kim BH, Ramsay LJ, Gillespie MA. 2004. Developing a biodata measure and situational judgment inventory as predictors of college student performance. J. Appl. Psychol. 89:187–207

Ployhart RE, Ryan AM, Bennett M. 1999. Explanations for selection decisions: applicants' reactions to informational and sensitivity features of explanations. J. Appl. Psychol. 84:87–106

Ployhart RE, Weekley JA, Holtz BC, Kemp C. 2003.Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? Pers. Psychol. 56:733–52

Podsakoff PM, MacKenzie SB, Paine JB, Bachrach DG. 2000. Organizational citizenship behaviors: a critical review of the theoretical and empirical literature and suggestions for future research. J. Manag. 26:513–63

Potosky D, Bobko P. 2004. Selection testing via the Internet: practical considerations and exploratory empirical findings. Pers. Psychol. 57:1003–34

Pulakos ED, Arad S, Donovan MA, Plamondon KE. 2000. Adaptability in the workplace: development of a taxonomy of adaptive performance. J. Appl. Psychol. 85:612–24

Pulakos ED, Schmitt N, Dorsey DW, Arad S, Hedge JW, Borman WC. 2002. Predicting adaptive performance: further tests of a model of adaptability. Hum. Perform. 15:299–323

Raju NS, Drasgow F. 2003. Maximum likelihood estimation in validity generalization. See Murphy 2003, pp. 263–85

Richman WL, Kiesler S, Weisband S, Drasgow F. 1999. A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. J. Appl. Psychol. 84:754–75

Robie C, Ryan AM. 1999. Effects of nonlinearity and heteroscedasticity on the validity of conscientiousness in predicting overall job performance. Int. J. Select. Assess. 7:157–69

Roth PL, Bevier CA, Bobko P, Switzer FS,Tyler P. 2001a. Ethnic group differences in cognitive ability in employment and educational settings: a meta-analysis. Pers. Psychol. 54:297–330

Roth PL, Bobko P. 2000. College grade point average as a personnel selection device: ethnic group differences and potential adverse impact. J. Appl. Psychol. 85:399–406

Roth PL, Bobko P, McFarland LA. 2005. A meta-analysis of work sample test validity: updating and integrating some classic literature. Pers. Psychol. 58:1009–37

Roth PL, Bobko P, Switzer FS. 2006. Modeling the behavior of the 4/5ths rule for determining adverse impact: reasons for caution. J. Appl. Psychol. 91:507–22

Roth PL, Bobko P, Switzer FS, Dean MA. 2001b. Prior selection causes biased estimates of standardized ethnic group differences: simulation and analysis. Pers. Psychol. 54:591–617

Roth PL, Huffcutt AI, Bobko P. 2003. Ethnic group differences in measures of job performance: a new meta-analysis. J. Appl. Psychol. 88:694–706

Roth PL, Van Iddekinge CH, Huffcutt AI, Eidson CE, Bobko P. 2002. Corrections for range restriction in structured interview ethnic group differences: the values may be larger than researchers thought. J. Appl. Psychol. 87:369–76

Rotundo M, Sackett PR. 2002. The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: a policy-capturing approach. J. Appl. Psychol. 87:66–80

Roznowski M, Dickter DN, Hong S, Sawin LL, Shute VJ. 2000. Validity of measures of cognitive processes and general ability for learning and performance on highly complex computerized tutors: Is the g factor of intelligence even more general? J. Appl. Psychol. 85:940–55

Ryan AM, McFarland L, Baron H, Page R. 1999. An international look at selection practices: nation and culture as explanations for variability in practice. Pers. Psychol. 52:359–91

Ryan AM, Sacco JM, McFarland LA, Kriska SD. 2000. Applicant self-selection: correlates of withdrawal from a multiple hurdle process. J. Appl. Psychol. 85:163–79

Rynes SL, Colbert AE, Brown KG. 2002. HR professionals' beliefs about effective human resource practices: correspondence between research and practice. Hum. Res. Manag. 41:149–74

Sackett PR, Devore CJ. 2001. Counterproductive behaviors at work. See Anderson et al. 2001, pp. 145–64

Sackett PR, Laczo RM, Lippe ZP. 2003. Differential prediction and the use of multiple predictors: the omitted variables problem. J. Appl. Psychol. 88:1046–56

Sackett PR, Lievens F, Berry CM, Landers RN. 2007. A cautionary note on the effects of range restriction on predictor intercorrelations. J. Appl. Psychol. 92:538–44

Sackett PR, Schmitt N, Ellingson JE, Kabin MB. 2001. High-stakes testing in employment, credentialing, and higher education—prospects in a postaffirmative-action world. Am. Psychol. 56:302–18

Sackett PR, Yang H. 2000. Correction for range restriction: an expanded typology. J. Appl. Psychol. 85:112–18

Salgado JF. 2003. Predicting job performance using FFM and non-FFM personality measures. J. Occup. Organ. Psychol. 76:323–46

Salgado JF, Anderson N, Moscoso S, Bertua C, de Fruyt F. 2003a. International validity generalization of GMA and cognitive abilities: a European community meta-analysis. Pers. Psychol. 56:573–605

Salgado JF, Anderson N, Moscoso S, Bertua C, de Fruyt F, Rolland JP. 2003b. A metanalytic study of general mental ability validity for different occupations in the European community. J. Appl. Psychol. 88:1068–81

Sanchez RJ, Truxillo DM, Bauer TN. 2000. Development and examination of an expectancy-based measure of test-taking motivation. J. Appl. Psychol. 85:739–50

Scherbaum CA. 2005. Synthetic validity: past, present, and future. Pers. Psychol. 58:481–515

Schleicher DJ, Day DV, Mayes BT, Riggio RE. 2002. A new frame for frame-of-reference training: enhancing the construct validity of assessment centers. J. Appl. Psychol. 87:735–46

Schleicher DJ, Venkataramani V, Morgeson FP, Campion MA. 2006. So you didn't get the job . . . now what do you think? Examining opportunity-to-perform fairness perceptions. Pers. Psychol. 59:559–90

Schmidt FL, Oh IS, Le H. 2006. Increasing the accuracy of corrections for range restriction: implications for selection procedure validities and other research results. Pers. Psychol. 59:281–305

Schmidt FL, Rader M. 1999. Exploring the boundary conditions for interview validity: metanalytic validity findings for a new interview type. Pers. Psychol. 52:445–64

Schmidt FL, Zimmerman RD. 2004. A counterintuitive hypothesis about employment interview validity and some supporting evidence. J. Appl. Psychol. 89:553–61

Schmit MJ, Kihm JA, Robie C. 2000. Development of a global measure of personality. Pers. Psychol. 53:153–93

Schmitt N, Gooding RZ, Noe RA, Kirsch M. 1984. Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. Pers. Psychol. 37:407–22

Schmitt N, Kunce C. 2002. The effects of required elaboration of answers to biodata questions. Pers. Psychol. 55:569–87

Schmitt N, Oswald FL. 2006. The impact of corrections for faking on the validity of noncognitive measures in selection settings. J. Appl. Psychol. 91:613–21

Schmitt N, Oswald FL, Kim BH, Gillespie MA, Ramsay LJ, Yoo TY. 2003. Impact of elaboration on socially desirable responding and the validity of biodata measures. J. Appl. Psychol. 88:979–88

Schmitt N, Rogers W, Chan D, Sheppard L, Jennings D. 1997. Adverse impact and predictive efficiency of various predictor combinations. J. Appl. Psychol. 82:719–30

Soc. Ind. Organ. Psychol. 2003. Principles for the Validation Use of Personnel Selection Procedures. Bowling Green, OH: Soc. Ind. Organ. Psychol.

Spector PE, Fox S. 2005. A model of counterproductive work behavior. In Counterproductive Workplace Behavior: Investigations of Actors and Targets, ed. S Fox, PE Spector, pp. 151–74. Washington, DC: APA

Stark S, Chernyshenko OS, Drasgow F, Williams BA. 2006. Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? J. Appl. Psychol. 91:25–39

Steel PDG, Huffcutt AI, Kammeyer-Mueller J. 2006. From the work one knows the worker: a systematic review of the challenges, solutions, and steps to creating synthetic validity. Int. J. Select. Assess. 14:16–36

Steele C, Spencer SJ, Aronson J. 2002. Contending with group image: the psychology of stereotype threat. In Advances in Experimental Social Psychology, ed. MP Zanna, pp. 379–440. San Diego, CA: Academic

Stewart GL. 1999. Trait bandwidth and stages of job performance: assessing differential effects for conscientiousness and its subtraits. J. Appl. Psychol. 84:959–68

Stewart GL, Nandkeolyar AK. 2006. Adaptation and intraindividual variation in sales outcomes: exploring the interactive effects of personality and environmental opportunity. Pers. Psychol. 59:307–32

Sturman MC, Cheramie RA, Cashen LH. 2005. The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. J. Appl. Psychol. 90:269–83

Tam AP, Murphy KR, Lyall JT. 2004. Can changes in differential dropout rates reduce adverse impact? A computer simulation study of a multi-wave selection system. Pers. Psychol. 57:905–34

Taylor PJ, Pajo K, CheungGW,Stringfield P. 2004. Dimensionality and validity of a structured telephone reference check procedure. Pers. Psychol. 57:745–72

Tett RP, Burnett DD. 2003. A personality trait-based interactionist model of job performance. J. Appl. Psychol. 88:500–17

Thoresen CJ, Bradley JC, Bliese PD, Thoresen JD. 2004. The Big Five personality traits and individual job performance growth trajectories in maintenance and transitional job stages. J. Appl. Psychol. 89:835–53

Tippins NT, Beaty J, Drasgow F, Gibson WM, Pearlman K, et al. 2006. Unproctored Internet testing in employment settings. Pers. Psychol. 59:189–225

Truxillo DM, Bauer TN. 1999. Applicant reactions to test score banding in entry-level and promotional contexts. J. Appl. Psychol. 84:322–39

Truxillo DM, Bauer TN, Campion MA, Paronto ME. 2002. Selection fairness information and applicant reactions: a longitudinal field study. J. Appl. Psychol. 87:1020–31

Van Iddekinge CH, Raymark PH, Roth PL. 2005. Assessing personality with a structured employment interview: construct-related validity and susceptibility to response inflation. J. Appl. Psychol. 90:536–52

Van Rooy DL, Viswesvaran C. 2004. Emotional intelligence: a meta-analytic investigation of predictive validity and nomological net. J. Vocat. Behav. 65:71–95

Van Rooy DL, Viswesvaran C, Pluta P. 2005. An evaluation of construct validity: What is this thing called Emotional Intelligence? Hum. Perform. 18:45–62

Van Vianen AEM, Taris R, Scholten E, Schinkel S. 2004. Perceived fairness in personnel selection: determinants and outcomes in different stages of the assessment procedure. Int. J. Select. Assess. 12:149–59

Viswesvaran C, Schmidt FL, Ones DS. 2005. Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. J. Appl. Psychol. 90:108–31

Wagner RK, Sternberg RJ. 1985. Practical intelligence in real world pursuits: the role of tacit knowledge. J. Personal. Soc. Psychol. 49:436–58

Weekley JA, Ployhart RE, eds. 2006. Situational Judgment Tests: Theory, Measurement and Application. San Francisco, CA: Jossey Bass

Wilk SL, Cappelli P. 2003. Understanding the determinants of employer use of selection methods. Pers. Psychol. 56:103–24

Witt LA. 2002. The interactive effects of extraversion and conscientiousness on performance. J. Manag. 28:835–51

Witt LA, Burke LA, Barrick MR, Mount MK. 2002. The interactive effects of conscientiousness and agreeableness on job performance. J. Appl. Psychol. 87:164–69

Witt LA, Ferris GR. 2003. Social skill as moderator of the conscientiousness-performance relationship: convergent results across four studies. J. Appl. Psychol. 88:809–20

Zeidner M, Matthews G, Roberts RD. 2004. Emotional intelligence in the workplace: a critical review. Appl. Psychol. Int. Rev. 53:371–99