Personnel Selection: Looking Toward the Future—Remembering the Past

ARTICLE in ANNUAL REVIEW OF PSYCHOLOGY · FEBRUARY 2000


2 AUTHORS, INCLUDING:

Frederick L Oswald
Rice University

105 PUBLICATIONS 2,183 CITATIONS

Available from: Frederick L Oswald
PERSONNEL SELECTION: Looking Toward the Future—Remembering the Past

Leaetta M. Hough and Frederick L. Oswald
The Dunnette Group, St. Paul, Minnesota 55102; e-mail: Leaetta@msn.com
Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907; e-mail: oswald@psych.purdue.edu

Key Words  job performance, assessment, validity, adverse impact, personality

Abstract  This chapter reviews personnel selection research from 1995 through 1999. Areas covered are job analysis; performance criteria; cognitive ability and personality predictors; interview, assessment center, and biodata assessment methods; measurement issues; meta-analysis and validity generalization; evaluation of selection systems in terms of differential prediction, adverse impact, utility, and applicant reactions; emerging topics on team selection and cross-cultural issues; and finally professional, legal, and ethical standards. Three major themes are revealed: (a) Better taxonomies produce better selection decisions; (b) The nature and analyses of work behavior are changing, influencing personnel selection practices; (c) The field of personality research is healthy, as new measurement methods, personality constructs, and compound constructs of well-known traits are being researched and applied to personnel selection.

CONTENTS

Introduction ................................................................. 632
Job and Work Analysis .................................................. 632
Criteria ................................................................. 633
   Taxonomic Issues .................................................. 633
   Measurement Issues .............................................. 634
   Dynamic Criteria ................................................. 634
Predictors ............................................................... 634
   Cognitive Abilities and Job Knowledge ...................... 634
   Personality .......................................................... 636
   Multiple Predictor Domains ...................................... 640
Assessment Methods .................................................... 641
   Interview .......................................................... 641
   Assessment Centers .............................................. 642
   Biodata ............................................................. 643
Measurement Issues and Validation Strategies .................. 644
   Measurement Issues .............................................. 644
INTRODUCTION


JOB AND WORK ANALYSIS

Recognizing the changing nature of work, many researchers and practitioners conduct “work” analysis, focusing on tasks and cross-functional skills of workers, rather than “job” analysis with its focus on static jobs (Cascio 1995, Nelson 1997, Pearlman 1997, Sanchez & Levine 1999). Perhaps the O*NET, the computerized delivery system for the Dictionary of Occupational Titles (DOT), has made the greatest operational strides in addressing this shift (Peterson et al 1999). The O*NET is a flexible database containing occupational information structured around a “content model” linking work behaviors to worker attributes, much in line with Dunnette’s call to bridge these “two worlds of behavioral taxonomies” (1976, p. 477). As information on work (e.g. jobs, organizational contexts, and work characteristics) and the worker (e.g. knowledge, skills, interests, and motivation) changes, the computerized nature of the O*NET allows an equally responsive change in its database. Collecting and using such information to select individuals for jobs, tasks, or roles is becoming more critical than ever (Campbell 1999).
Work/job analysis now includes personality variables alongside traditional cognitive, behavioral, and situational variables. The National Skill Standards Board (K Pearlman, unpublished manuscript) and O*NET (Peterson et al 1999) both incorporate personality-based work requirements. One personality-based job analysis instrument generates job profiles along seven personality scales (Hogan & Rybicki 1998). Another personality-based job analysis evaluated 260 jobs, meaningfully distinguishing 12 DOT-based job clusters (Raymark et al 1997).

Sixteen potential cognitive and social sources of rating inaccuracy in job analysis may influence different dimensions and psychometric properties of job analysis (Morgeson & Campion 1997). Type of job data and method for clustering jobs affect similarities and distinctions between jobs (Colihan & Burger 1995), and Q-factor analysis of rated task importance for two job titles shows meaningful within-title variation (Sanchez et al 1998). Future job analysis studies might identify substantive sources of variance attributable to types of raters, workers, or both.

**CRITERIA**

**Taxonomic Issues**

Job performance constitutes all measurable work behaviors relevant to organizational goals and within the individual’s control (Campbell et al 1996). Job performance is complex, dynamic, and multidimensional, and consequently personnel selection systems might predict individual differences for several types of job performance (e.g. task proficiency and leadership behaviors). Models incorporating multiple predictors and multiple criteria first apply rational weights to performance criteria and then derive least-squares optimal weights for the predictors. Meta-analytic correlations between ability, conscientiousness (predictors), individual task performance, and organizational citizenship (criteria) illustrate how validity can vary greatly depending on criterion weights (Murphy & Shiarella 1997). Absent criterion data for establishing regression weights, rational weighting of a selection battery increases appropriateness and legal defensibility. Weights could multiply job analysis ratings of importance, time spent, consequences of errors, and time-to-proficiency (Arthur et al 1996).

Contextual performance, or organizational citizenship behavior, is a relatively new and multifaceted job performance construct (Borman & Motowidlo 1997). Coleman & Borman (1999) classified organizational citizenship behaviors into three broad categories: interpersonal citizenship behavior (benefiting employees), organizational citizenship performance (benefiting organizations), and job/task conscientiousness (benefiting work itself). Hierarchical regression analyses suggest that interpersonal facilitation is a part of contextual performance, but job dedication (similar to job/task conscientiousness within the organizational citizenship framework) is a part of task performance (Van Scotter & Motowidlo...
Organizations clearly require both task and contextual performance (Kiker & Motowidlo 1999).

**Measurement Issues**

Regarding criterion reliability, interrater reliability coefficients are more appropriate to use than intrarater (coefficient alpha) or test-retest reliabilities, and interrater reliabilities of supervisory ratings of overall and dimensional job performance are higher than peer ratings (Viswesvaran et al 1996). A greater understanding of “unreliable” criterion variance is needed to address biases in measures of constructs (Schmitt et al 1995). Similarly, criterion range restriction (or enhancement) may be a legitimate organizational phenomenon affecting criterion-related validities and not merely a statistical artifact (James et al 1992). Organizational climate, for instance, does not always attenuate correlations between procedural fairness and customer-perceived performance relationships (Burke et al 1996). Further research should investigate how individual differences, job types, and their interactions influence the mean and variance of criterion measures (see Hattrup & Jackson 1996).

**Dynamic Criteria**

Job performance and the relative contributions of its determinants (job knowledge, skill, and motivation) change, calling for longitudinal models of reliability and validity (Tisak & Tisak 1996). The nature of performance change (e.g. systematic vs random change or reversible vs irreversible change) and how constructs relate between individual and group levels are critical theoretical and methodological issues (Chan 1998a,b). Nonlinear mixed-effects models simultaneously estimate individual and group levels of change (Cudeck 1996), accommodating missing data, prespecified error structures (see DeShon et al 1998a), and individuals not measured at the same time points. For a review and tutorial of quadratic and linear models of longitudinal change, see Chan (1998a).

Personality items predicted change in eight consecutive quarters of securities sales performance, with a curvilinear group mean increase over time (i.e. greater increase initially) and different rates of increase for each individual (Ployhart & Hakel 1998). Psychomotor ability predicted initial piece-rate performance of sewing machine operators, and cognitive ability predicted performance change. Individuals with less experience and lower levels of initial performance changed more (Deadrick et al 1997).

**PREDICTORS**

**Cognitive Abilities and Job Knowledge**

Self-selection on cognitive ability may precede personnel selection. Job seekers may select into or “gravitate” toward jobs with ability requirements commensurate with the seekers’ own general cognitive ability (Wilk & Sackett 1996).
Ability self-evaluations might lead to seeking coaching and practice on ability tests. Firefighter applicants with lower cognitive ability scores are likely to attend a free test preparation program, although the program’s effects on raising ability test scores (reading, listening, and spelling) are minimal (Ryan et al 1998b).

**General Cognitive Ability** Various cognitive ability tests (e.g. verbal, numerical, and spatial tests) intercorrelate positively, and the common variance often operationalizes \( g \), a single general cognitive ability factor. For many jobs and practical work outcomes (job knowledge acquisition, training performance, and job performance), \( g \) predicts well (e.g. Levine et al 1996). For prediction in less complex jobs or in later stages of complex learning, \( g \) is less useful but rarely useless (Gottfredson 1997). Greater understanding of \( g \) is needed (Campbell 1996), as its determinants and theoretical meaning are debated (see Lubinski 2000). Individuals higher in \( g \) show lower intercorrelations between specific abilities (Legree et al 1996). Matching individuals’ specific abilities (or ability profiles) to particular jobs may therefore be especially important for individuals with higher \( g \) (Lubinski & Benbow 1999). Item response theory and computerized adaptive testing have clarified relationships between \( g \) and specific abilities contributing to it (Sands et al 1997, Segall 1999). Developers of good cognitive ability selection tests cannot rely on knowledge of \( g \) alone.

**Ability and Job Knowledge** Multiple-ability test battery data from 3000 Air Force enlistees supported a hierarchical ability structure (Carretta & Ree 1996) that fit the data well, considerably better than two nonhierarchical models or a \( g \)-only model (see also Carretta et al 1998). Men and women show similar hierarchical ability structures (Carretta & Ree 1997).

Carretta & Doub (1998) tested the mediating effect of prior mechanical and electrical knowledge on the relationship between \( g \) and subsequent job knowledge of Air Force trainees. Comparing racial groups (White, Black, and Hispanic) resulted in little moderating effect. Comparing gender groups prior job knowledge mediated subsequent job knowledge for males but not for females. The effect of \( g \) was weaker in individuals with more prior job knowledge; conversely, \( g \) was stronger in individuals with less prior job knowledge. It seems that \( g \) predicts the rate of job knowledge acquisition, which in turn has a larger direct influence on performance ratings than the indirect effect of \( g \) (Ree et al 1995).

**Academic Achievement and Language Proficiency** Meta-analysis found that undergraduate college grade point average (GPA) predicted job performance across many types of organizations, especially for job performance measured closer in time to the GPA (Roth et al 1996). Another meta-analysis reported substantial criterion-related validities for aptitude tests predicting GPA in graduate school. Subject-specific tests (possibly a job knowledge analog) had higher validities than verbal and mathematical tests, but both were of useful magnitude (Kuncel et al 1999).
Increasing immigration and concomitant workforce diversity suggest measuring English language proficiency when selecting for certain jobs. For entry-level meat trimmers \((N = 87)\) whose native language was not English, a written and spoken English proficiency test was internally consistent and clearly linked to job analysis information. The test predicted supervisory ratings of overall job performance (Chan et al 1999). Spoken English proficiency assessment is now possible via real-time computer processing and analysis of human speech (Bernstein 1999).

**Adverse Impact** Cognitive ability measures tend to show nontrivial racial-group mean differences. Helms’ (1992) hypothesis states that White-Black mean differences might be reduced by couching ability test content within a social context. An expert panel modified abstract ability items to reflect everyday organizational, social, and life situations. Contrary to the hypothesis, marked White-Black differences remained under the new test format, even under large-sample replication and parallel test forms (DeShon et al 1998b). Research needs to expand the number and types of items, explore different administration formats, and examine other specific abilities.

Short-term memory tests (digit span and digit-symbol substitution) show promise as an alternative or supplement to traditional ability tests, with lower adverse impact and good validities with job performance. Meta-analysis estimated White-Black mean differences on short-term memory tests at 0.48 SD, about half the 1.0 SD difference typically found in general cognitive ability tests. Short-term memory tests are reliable and correlate with training performance and job performance \([r \approx 0.45\) (Verive & McDaniel 1996)]. Lower adverse impact combined with respectable overall criterion-related validity encourages future research in this area.

The Armed Services Vocational Aptitude Battery (ASVAB) and an experimental ability test battery predicted training performance criteria in 17 military jobs (Sager et al 1997), showing some necessary tradeoffs. One cannot completely (a) minimize adverse impact for all subgroups compared (see Hoffman & Thornton 1997 for this issue in utility context), (b) maximize both criterion-related validity within jobs and classification efficiency across jobs, or (c) satisfy both (a) and (b). A selection strategy aimed at minimizing adverse impact may differ somewhat from a selection strategy aimed at maximizing mean predicted performance (Sackett & Roth 1996).

**Personality**

**Personality Taxonomies and Constructs** The Five Factor Model (FFM), consisting of Extraversion, Agreeableness, Conscientiousness, Neuroticism (Adjustment), and Openness to Experience (see Wiggins & Trapnell 1997), enjoys considerable support. Factor analysis supports the robustness and generalizability of the FFM across different theoretical frameworks, assessments, rating sources,
and cultures (see Hogan & Ones 1997b, Saucier & Goldberg 1998, Wiggins & Trapnell 1997). The model is useful for summarizing information and guiding theory and research (e.g. Mount & Barrick 1995, Tokar et al 1998).

The FFM yields information about the higher-order factor structure of personality; however, it ignores, confounds, or otherwise obscures understanding of variables combined into five broad factors (Hough 1997, 1998b; Hough & Schneider 1996). FFM factors contain facets with high and low criterion-related validities, diluting the criterion-related validity of the factors. A review of meta-analyses concluded the FFM factors do not correlate highly with job performance (Matthews 1997). As alternatives, researchers are turning to nonhierarchical models such as the circumplex (Plutchik & Conte 1997) and other hierarchies. Hough (1997, 1998b) argues for a more refined taxonomy, distinguishing achievement from conscientiousness and extraversion and affiliation from extraversion. Meta-analyses demonstrate the importance of these distinctions for predicting managerial performance (Hough et al 1998) and sales performance (Vinchur et al 1998). Ghiselli’s (1966) personality framework was compared with the FFM framework in the Barrick & Mount (1991) meta-analysis: the median uncorrected validity in the Ghiselli meta-analysis was 0.24; the highest mean uncorrected validity for an FFM variable in the Barrick & Mount (1991) study was 0.15 (Hough 1997). Several important personality constructs not within the FFM have been used for predicting work behavior:

**Emotionality** Emotionality, or affectivity, consists of two bipolar dimensions at the most general level: negative-positive and aroused-unaroused (Averill 1997, Russell & Carroll 1999). A state measure of emotionality did not correlate with job performance, but a dispositional measure did (Wright & Staw 1999). In social welfare workers, negative affectivity correlated positively with emotional exhaustion or “burnout,” and emotional exhaustion correlated negatively with job performance (Wright & Cropanzano 1998).

**Social Competence** Social competence is a compound variable consisting of social insight, social maladjustment, social appropriateness, social openness, social influence, warmth, and extraversion (Schneider et al 1996). Reliable self-report measures of social insight (e.g. Gough 1968) and empathy (e.g. Hogan 1968) have a long history, as do situational judgment measures of social intelligence (e.g. Moss et al 1955). Variables subsumed under social competence might increment predictive validity for criteria emphasizing interpersonal effectiveness.

**Conscientiousness** Many claim that conscientiousness, a FFM factor, is a valid predictor across organizations, jobs, and situations (Hogan & Ones 1997b, Mount & Barrick 1995, Salgado 1997a, 1998). Others question this wholesale conclusion (Hough 1997, 1998b; Robertson & Callinan 1998). Whether or in what direction conscientiousness predicts performance obviously depends on the criterion construct and how conscientiousness is defined and operationalized. Based on the
Hogan & Ones’ (1997) definition of conscientiousness as conformity and socially prescribed impulse control, conscientiousness would likely not predict performance across organizations, jobs, or situations in which creativity or innovation is important (Hough 1997, 1998b; Hough et al 1998).

**Integrity Tests** Meta-analyses of relations between integrity tests and FFM variables indicate that integrity tests are compound variables consisting primarily of conscientiousness, agreeableness, and adjustment (Ones & Viswesvaran 1998b). Four themes account for most of the variance in overt and personality-based integrity tests: punitive attitudes, admissions of illegal drug use, reliability, and admissions of theft (Hogan & Brinkmeyer 1997). Importantly, integrity tests differ from tests of deception (Murphy & Luther 1997). Sackett & Wanek (1996) provided an insightful and thorough review of integrity testing that dealt with construct- and criterion-related validity evidence; moderator variables; social desirability and applicant reactions; and legal, professional, and governmental evaluations. Meta-analysis indicates that integrity and conscientiousness tests usefully supplement general cognitive ability tests when predicting overall job performance (Schmidt & Hunter 1998). Converging evidence exists for the construct, criterion-related, and incremental validity of integrity tests (Miner & Capps 1996, Ones & Viswesvaran 1998b), but considerable variability may accompany the overall findings.

For example, criteria often used in integrity-testing research are problematic. Self-report or admission of counterproductive behavior confounds reliability with validity for overt integrity tests and underestimates the extent of counterproductive work behaviors, as do more direct measures (e.g. detected theft). Moreover, counterproductivity is not a unitary construct (Ashton 1998, Sackett & Wanek 1996). Two recent meta-analyses summarized correlations between integrity test scores and two facets of counterproductivity: reported drug abuse \( r = 0.21 \) (Schmidt et al 1997) and number of job-related accidents \( r = 0.52 \) (Ones & Viswesvaran 1998b). Workplace violence, a facet of counterproductivity, tends to be better predicted by narrow measures such as aggression and violence scales than by broad honesty tests. Meta-analysis indicates that the validity for predicting workplace violence is higher for violence scales than for integrity tests \( r = 0.48 \) vs 0.26 (Ones et al 1994). A physical-aggression measure predicted aggressive penalty minutes \( r = 0.33 \) in high school hockey games, but not nonaggressive penalty minutes \( r = 0.04 \) (Bushman & Wells 1998).

**Customer Service Orientation** Meta-analysis finds that customer service orientation is a compound variable consisting of agreeableness \( r = 0.70 \), adjustment \( r = 0.58 \), and conscientiousness \( r = 0.43 \) (Ones & Viswesvaran 1996), and customer service scales correlate with performance in customer service jobs \( r = 0.31 \) (Frei & McDaniel 1998).
Core Self-Evaluation  Core self-evaluation is a compound variable consisting of self-esteem, generalized self-efficacy, locus of control, and emotional stability (Judge et al. 1998). Meta-analysis estimated the validity of core self-evaluation for predicting job performance at 0.30 (TA Judge & JE Bono, submitted for publication). A much larger meta-analysis suggests that self-efficacy, a facet of core self-evaluation, correlates higher with job performance than does core self-evaluation as a whole \([r = 0.38\) (Stajkovic & Luthans 1998)].

Other Meta-Analyses  Meta-analysis of the Five Factor Model validities in U.S. studies from 1992 through 1997 produced results similar to past U.S. meta-analyses (Anderson & Viswesvaran 1998), but meta-analyses involving only European samples produced somewhat different results (Salgado 1997a, 1998). Both conscientiousness and emotional stability correlated positively with job performance across occupational groups, and both contributed incremental variance beyond general mental ability in predicting overall job performance. Other meta-analyses of validities of FFM factors indicate that agreeableness, as well as conscientiousness and emotional stability, predicts performance in jobs involving interpersonal interaction (Mount et al. 1998). Managerial potential scales predict overall managerial job performance \([r \approx 0.40\) (Ones et al. 1998)], although meta-analytic validities tend to be low for FFM factors predicting overall managerial performance (Hough et al. 1998). Some FFM facets had much higher validities, shedding light on how facet-level variables might combine to form managerial potential scales with high criterion-related validity. Many of these meta-analytic researchers corrected study correlations for predictor range restriction by using national norm SDs from personality test manuals. This practice of using norm SDs appears warranted, because SDs of job applicants on personality measures are about 2%–9% less than those based on national norms (Ones & Viswesvaran 1999).

Conditional Reasoning  Pioneered by James (1998, 1999), conditional reasoning assumes that individuals' personalities are differentiated by the type of logical reasoning used to justify their actions. For example, people who score high on achievement motivation tend to attribute success to internal rather than external sources and consider demanding tasks challenging rather than frustrating. An achievement conditional reasoning scale correlated positively with scholastic criteria, in-basket performance, and other achievement scales (James 1998, Migetz et al. 1999a, Smith et al. 1995). An aggression conditional reasoning scale correlated negatively with overall job performance (Hornick et al. 1999, James 1998) and positively with counterproductive work behavior (Burroughs et al. 1999, Migetz et al. 1999b, Patton et al. 1999). This approach appears to overcome many problems related to intentional distortion.

Intentional Distortion  Not surprisingly, meta-analysis shows large mean-score differences between honest and directed-faking conditions (Viswesvaran & Ones
The amount of distortion in naturally occurring applicant settings is uncertain. Rosse et al. (1998) found that applicants (N = 197) scored on average 0.69 SD higher than incumbents (N = 73) on FFM facet-level scales. In contrast, three separate samples involving over 40,500 applicants and over 1700 incumbents found significantly less distortion on similar scales (Hough 1998a). Ability and motivation to fake may be key determinants in the amount of distortion found in applicant settings (Snell et al. 1999). Meta-analysis indicates that explicit warnings not to distort do reduce distortion [0.23 SD (Dwight & Donovan 1998)].

A slew of recent studies has investigated intentional distortion effects on criterion-related validities. Many assert that distortion does not tend to moderate, mediate, suppress, or attenuate the criterion-related validities of personality scales (Barrick & Mount 1996; Hogan 1998; Hough 1997, 1998a,b; Ones & Viswesvaran 1998b,c; Ones et al. 1996). Others, such as Douglas et al. (1996), Snell & McDaniel (1998), and Zickar & Drasgow (1996), contend that distortion seriously reduces criterion-related validity. Hough (1998a) resolved the apparent conflict by stratifying results by employment setting. In directed-faking settings, self-report scale scores have dramatically lower criterion-related validities than those obtained in applicant or incumbent settings; in applicant settings, self-report scale scores have the same or slightly lower criterion-related validities than those obtained from job incumbents in research-only settings. Similarly, construct validity may be negatively affected in directed-faking studies (Ellingson et al. 1999a), but the effect does not seem to be as serious in applicant settings (Collins & Gleaves 1998, Ellingson et al. 1999b, Ones & Viswesvaran 1998c).

Coaching individuals on personality tests potentially threatens the effectiveness of traditional social desirability scales. Subtle items can be more resistant to coaching and distortion than obvious items (Alliger et al. 1996). Theory-driven approaches to scale development and validity data to refine items produce subtle items resistant to distortion and with excellent validity (Gough 1994).

**Race and Ethnic Background** Similar personality factor structures for Blacks and Whites are found (Collins & Gleaves 1998). Meta-analyses of White, Black, Hispanic, and Native American groups indicate minimal group mean differences for three overt integrity tests (Ones & Viswesvaran 1998a) and for FFM factors, although Hispanics scored 0.60 SD higher than Whites on social desirability scales (Hough 1998b). Personality variables have little adverse impact against minorities, if any. Score correction strategies using social desirability scales to correct distortion in content scale scores might affect Hispanics more than others.

**Multiple Predictor Domains**

Applied psychology has long postulated that ability and motivation interact in predicting job performance: High performers must have the requisite ability and effort to do the job; neither ability nor effort alone suffices. However, regression
analyses of data across job samples, performance criteria, and different ability and motivation measures have yielded nonexistent or very slight incremental ability-motivation interaction effects (Sackett et al 1998). A combined meta-analysis of ability-personality correlations and review of the empirical findings on ability-interest and personality-interest relationships have produced an integrated model identifying four categories or “trait complexes”: science/math, clerical/conventional, social, and intellectual/cultural (Ackerman & Heggestad 1997). A vocational interest structure was empirically linked to Gottfredson’s (1986) ability-based job classification framework, yielding four similar categories (Oswald & Ferstl 1999). Linking different predictor domains to a common job classification framework may clarify the constellations of predictors that are useful for selection into various types of jobs (see Arthur & Bennett 1995, Hattrup et al 1998, Johnson et al 1997, Mael et al 1996b, Vinchur et al 1998) and may improve synthetic validity efforts.

ASSESSMENT METHODS

Interview

**Interview Structure** Compared with unstructured interviews, structured employment interviews define content more explicitly. Their successes are therefore more likely to replicate, and they are better analyzed and meta-analyzed to determine their transportability to other jobs and work settings. Nonetheless, organizations still prefer unstructured interviews by a wide margin (Graves & Karren 1996). A comprehensive review indicates that various components of structured interviews influence the interview’s psychometric properties, legal defensibility, and applicant/interviewer reactions (Campion et al 1997). Recent research offers at least three other compelling reasons for structuring the core of the interview (see also the Dipboye 1997 review).

**Reliability** Structured interviews tend to have higher interrater reliability than unstructured interviews. Meta-analysis reports average interrater reliabilities of 0.67 for high structure vs 0.34 for low structure (Conway et al 1995).

**Standardization** Standardized interviews place more burden on the instrument than any particular interviewer’s interviewing and assessment skills. A highly standardized situational interview, in which applicants respond to hypothetical critical work incidents, can be less susceptible to rating biases (Kataoka et al 1997). Computerized phone interviews have efficiently obtained standardized applicant information, with validity coefficients similar to those for traditional interviews (Schmidt & Rader 1999). Interviewer experience and training further standardize the interview (Conway et al 1995, Campion et al 1997, Huffcutt & Woehr 1999). Training in note taking improves attention, encoding, recall, and
evaluation of interview information focusing on work behaviors (Burnett et al 1998).

**Fairness** Structured interviews treat applicants in a consistent manner. Mean differences by race are more likely reduced in highly structured interviews containing content related to noncognitive constructs, especially for high-complexity jobs (Huffcutt & Roth 1998). Court outcomes on disparate impact and disparate treatment have favored organizations high on three interview characteristics: standardized administration, high job relatedness, and multiple raters (Williamson et al 1997).

**Interpersonal and Nonverbal Behavior** Interviewee characteristics (e.g. gaze, hand movement, and physical attractiveness) can predict several dimensions of managerial effectiveness (leadership, teamwork, and planning/organization), even when characteristics are coded independently of the content (Burnett & Motowidlo 1998; see also Motowidlo & Burnett 1995). The convergent and discriminant validities between interpersonal behavior in the interview and different job performance criteria should be considered.

**Assessment Centers**

Assessment centers (ACs) have long been haunted by evidence of content- and criterion-valid ratings lacking construct validity (Arthur et al 1999, Spychalski et al 1997, Woehr & Arthur 1999). Confusion about the constructs being measured, rating errors, type and form of rating procedures, and participant inconsistencies in behavior across exercises are possible explanations (Arthur & Tubre 1999, Guion 1998). Features improving AC ratings include having (a) only a few conceptually distinct constructs, (b) concrete, job-related construct definitions, (c) frame-of-reference assessor training with evaluative standards, (d) cross-exercise assessment, and (e) several psychology-trained assessors (Lievens 1998, Woehr & Arthur 1999). An AC designed and implemented on the basis of research and professional-practice guidelines (see Task Force on Assessment Center Guidelines 1989) produced construct-valid AC ratings. Generalizability theory facets associated with individuals and constructs accounted for 60% of the total variance, and facets associated with assessors and exercises accounted for 11% of the total variance (Arthur et al 1999).

ACs are expensive and prone to cost-benefit comparisons with other predictors. AC ratings have significant incremental validity over personality variables, and vice versa, when predicting managerial performance (Goffin et al 1996, with N = 68). AC ratings also have incremental validity over cognitive ability, although some AC exercise validities are founded primarily on their cognitive component (Goldstein et al 1998). By using policy capturing and meta-analysis, overall AC ratings have been predicted from both cognitive ability and personality variables \( R = 0.77 \) (Collins et al 1999). The key question is whether policy-
captured predictors predict job performance better than AC ratings (Howard 1999). Black-White mean exercise score differences have ranged from 0.03 to 0.40 SDs, with Blacks scoring lower (Goldstein et al 1999). Exercises emphasizing interpersonal skills more than cognitive ability have resulted in less or no adverse impact for Blacks (Goldstein et al 1999, Bobrow & Leonards 1997).

Biodata

Several researchers have focused on much-needed construct-oriented approaches to biodata (biographical information). CN MacLane (submitted for publication) refined the federal government’s Individual Achievement Record biodata scales, which measure social and cognitive abilities (see Gandy et al 1994). Personality-based biodata scales predict leadership (Stricker & Rock 1998) and life insurance sales (McManus & Kelly 1999). A review of eleven studies examined the validity of biodata scales based on Mumford & Stokes’ (1992) rigorous construct-oriented item-generation procedures. Scales were content and construct valid, with criterion-related validities similar to those for traditional empirical keying (Mumford et al 1996).

Biodata theory relies heavily on the principle that past behavior is the best predictor of future job performance (i.e. the “consistency” principle). Failure or negative life experiences also explain why biodata predict performance (Russell 1999). Moxie (i.e. courage or “ego-resiliency”) may moderate how negative life experiences influence development and subsequent job performance (Dean et al 1999, Muchinsky 1999). “Negative” is often in the eye of the beholder, however. Both positive and negative responses to elements of a broader life-events taxonomy may be needed.

Conclusions about the effectiveness of rational, empirical-keying, and factor-analytic biodata scale development strategies are inconsistent. Rational scales have predicted sales performance at least as well as empirically keyed and factor-analytic scales (Stokes & Searcy 1999). Factor-analytic and rational scales have predicted several customer service criteria much better than empirical keying (Schoenfeldt 1999). A meta-analysis found similar levels of criterion-related cross-validities across all three scale construction strategies (Hough & Paullin 1994). Rational, empirical-keying, and factor-analytic scale strategies need not be executed and compared separately; the strategies may iteratively inform one another.

Regarding the generalizability of biodata, the reliability, factor structure, and validity of biodata keys appear stable across two English-speaking countries (Dalesio et al 1996). Validity for a biodata inventory predicting managerial progress generalized across organizations and educational levels (Carlson et al 1999). “Contemporary” items tend to be more valid than “future/hypothetical” or “historical” items, and items that ask respondents about others’ opinions of them are more valid than direct self-report items (Lefkowitz et al 1999). Meta-analysis found that the amount and task-level specificity of work experience correlated
most highly with job performance (Quiñones et al. 1995). Rational biodata scales may produce inadequate levels of validity for separate racial/ethnic groups, but empirical item analysis can be used to produce a scale valid across groups (Schmitt & Pulakos 1998, Schmitt et al. 1999). Whitney & Schmitt (1997) discovered differential item functioning between racial subgroups in about one quarter of the biodata items they examined.

MEASUREMENT ISSUES AND VALIDATION STRATEGIES

Measurement Issues

Many published studies in personnel selection continue to suffer from low statistical power due to small sample sizes (Mone et al. 1996, Salgado 1998). Confidence intervals directly convey the impact of sample size on the accuracy of statistics (Hunter 1997). A computer program calculates confidence intervals on correlation coefficients corrected for measurement unreliability and range restriction (Salgado 1997b). Formulas are accurate in the large-sample case; for the small-sample case, the bootstrap method (i.e. calculating a distribution of correlations by resampling data with replacement) has generated accurate confidence intervals for ability-training performance validities (Russell et al. 1998). Instead of confidence intervals, Murphy & Myors (1999) provide noncentral $F$ tables and real-world examples that test minimum-effect null hypotheses for $t$-tests, correlations, and ANOVAs.

Measurement error variance can distort patterns of research results and mislead conclusions. Circumstances in 26 applied-research situations show when to correct validities for such error by using the appropriate reliability coefficient (Schmidt & Hunter 1996). Structural equation modeling can test the statistical significance of corrected correlation coefficients and the difference between two such correlations (Hancock 1997). Recent studies considered maximizing the reliability of a linear composite by weighting the constituent variables as a function of their reliabilities (Cliff & Caruso 1998, Li et al. 1996, Raykov 1997, Wang 1998).

Estimation formulas for the population validity (expected prediction in the entire population) and cross-validity (expected prediction in other independent samples) were thoroughly reviewed (Raju et al. 1997), and formulas were compared in a Monte Carlo study with data from a large sample of Air Force enlistees (Raju et al. 1999). Generally, the Ezekiel, Smith, and Wherry procedures all provided good squared population validity estimates, and Burkett’s formula best estimated the squared population cross-validity. In stepwise regression, results showed that the sample size-to-predictor ratio had to be relatively large (10:1 at least) to yield good cross-validity estimates. All cross-validity estimation formulas performed similarly well; none was clearly superior (Schmitt & Ployhart...
Instead of stepwise regression, researchers often judge the relative size of regression weights to decide on the important variables within a particular model. A new type of relative importance weights enhances the interpretability of regression results when predictors are highly intercorrelated (Johnson 1999).

Finally, personnel selection research must often deal with missing data. Generally, pairwise deletion is better than listwise deletion, and estimating missing scores via regression is better than substituting missing scores with their unconditional mean (Roth et al 1996). Personnel selection would profit from understanding substantive processes underlying “missingness.” Job promotions or transfers, emotional exhaustion from work, and organizational redesign all may reflect different types of longitudinal and cross-sectional attrition processes.

Meta-Analysis and Validity Generalization

Meta-analysis has had a far-reaching impact on policymaking, real-world application, and academic research (Hunter & Schmidt 1996) and is a useful quantitative tool for summarizing large bodies of personnel selection research (Murphy 1997). The meta-analytic mean effect size across studies (e.g. mean correlation coefficient) tends to be fairly accurate. In contrast, estimates of the variance of effect sizes (after correcting for statistical artifacts) can deviate from their actual population values by practically significant amounts (Oswald & Johnson 1998), which may affect meta-analytic conclusions about selection research. Variance estimates can be downwardly biased if one ignores the fact that individuals were selected on a variable correlated with the variables in the meta-analysis [i.e. incidental range restriction (Aguinis & Whitehead 1997)]. Statistical homogeneity tests of the variance have low statistical power and tend to discourage the search for moderator effects (Sánchez-Meca & Marín-Martínez 1997).

Meta-analysis is one of many lines of evidence supporting the use of a selection test. In addition to meta-analytically averaging correlations, researchers might consider what predicts group or organizational mean differences on predictors and criteria (Ostroff & Harrison 1999). Hierarchical linear modeling (HLM) has been advocated for this purpose; Hofmann (1997) presented some practical organizational examples. Synthetic validity evidence may provide additional validity information (Hoffman & McPhail 1998).

EVALUATION OF SELECTION SYSTEMS

Differential Prediction

A selection test with equal regression slopes across subgroups (e.g. race, gender, and age groups) does not necessarily measure a latent construct the same way across subgroups. Only data conforming to a special set of mathematical constraints will show a lack of differential prediction and latent construct equivalence simultaneously (Millsap 1995, 1997). To the extent that these constraints
are violated, this finding may challenge previous selection research suggesting a lack of differential prediction (e.g., cognitive ability research).

Aguinis & Stone-Romero (1997) examined range restriction effects on the statistical power of moderated multiple regression for detecting differential prediction. Range restriction often mistakenly led to concluding no differential prediction when validity differences were moderate (0.4 correlation units). Larger validity differences were detected, and smaller differences were not, regardless of range restriction. Computer software for estimating the statistical power of differential prediction in moderated multiple regression is available (Aguinis & Pierce 1998b). Future software could incorporate violations of the assumption of equal error variances between subgroup regression models. Assumption violations occur when larger subgroup sample sizes are paired with the smaller subgroup validity coefficient (e.g., Aguinis & Pierce 1998a, DeShon & Alexander 1996), but for some organizational data, violations are not severe enough to affect statistical inferences from moderated multiple regression (Oswald et al. 1999).

Error variance in the independent variables affects differential prediction (Terris 1997), and errors-in-variables regression addresses this problem. Errors-in-variables regression detected differential prediction more accurately than moderated multiple regression when reliability coefficients were >0.65 and sample sizes were >250 (Anderson et al. 1996).

Adverse Impact

Common sense might assert that combining low-adverse-impact predictors with a high-adverse-impact predictor improves adverse impact over using the high-adverse-impact predictor alone. However, Sackett & Ellingson (1997) presented tables illustrating reduced subgroup differences for some composites but increased differences for others. Composite measures invariably reduce subgroup differences less than expected. In particular, the “four-fifths rule” set out in the Uniform Guidelines (U.S. Equal Employment Opportunity Commission et al., 1978) is usually met only under very high selection ratios (≥0.90) or very slight composite group-mean differences ($d \leq 0.20$).

Similar conclusions came from meta-analytic estimates of criterion-related validities for a cognitive ability measure, a noncognitive composite (interview, biodata, and conscientiousness), and a composite of both, all independently predicting overall job performance. Selection batteries excluding cognitive ability almost always satisfied the four-fifths rule; batteries that included cognitive ability alone or in a composite almost never satisfied the rule (Schmitt et al. 1997). Findings are echoed in two large-sample studies of firefighter and police officer job applicants selected on cognitive ability and personality measures (Ryan et al. 1998a). Bobko et al (1999) updated the Schmitt et al (1997) meta-analytic matrix, discovering that even a noncognitive composite could violate the four-fifths rule when selection ratios were ≤50%. Group-mean differences on noncognitive composites should be determined before assuming they reduce adverse impact. Also,
adverse impact is partly a function of the criteria chosen and how they are weighted (Hattrup et al 1997).

Alternative forms of administration potentially reduce adverse impact and increase overall validity. Videotaped versions of situational judgment test material can reduce adverse impact and have greater face validity than paper-and-pencil versions. Presumably, validity improves because the content in the video format is preserved, and irrelevant variance related to reading comprehension is removed (Chan & Schmitt 1997). Similar conclusions come from comparing a video-based ability test to a traditional ability test (Pulakos & Schmitt 1996).

Banding
Given subgroup differences on selection tests (most notably cognitive ability tests), statistical banding can fulfill important goals such as maintaining workforce diversity and improving perceptions of process and outcome fairness in a selection procedure (Truxillo & Bauer 1999). Linearly transforming bands on predicted criterion scores into bands on predictor scores (Aguinis et al 1998) may improve banding because criterion differences tend to matter more to organizations than predictor differences (see Campbell 1996, Gottfredson 1999). Further advances in banding might consider that reliable predictors tend to produce smaller statistical bandwidths, in which differences larger than the band may still not be practical differences or translate into practical criterion differences.

Utility
A few recent studies focus on how utility information communicates the effectiveness of a selection system to organizational stakeholders. Whyte & Latham (1997) replicated the counterintuitive results of Latham & Whyte (1994), discovering that communicating positive utility information can actually decrease managers’ intentions to use a selection system. Utility information may prove beneficial as supplementary information, not delivered face-to-face as a “hard sell” (Cronshaw 1997). Communicating utility in terms of multiple outcomes (e.g. dollars, job performance, and organizational effectiveness) may lead to greater acceptance of the utility message by different stakeholders (Roth & Bobko 1997). Considering and balancing all particular stakeholder positions is difficult but beneficial to organizations (Austin et al 1996).

Applicant Reactions
Positive applicant reactions increase the chances of hiring the best applicants, facilitate the ability to recruit effectively, avoid the possibility of costly litigation, and contribute to the organization’s reputation (Gilliland & Steiner 1999, Ryan & Greguras 1998, Schmitt & Chan 1999). Selection systems can be viewed as socialization mechanisms imparting job information to applicants and affecting their work-related thoughts, attitudes, and behaviors (Anderson & Ostroff 1997). Managing applicant reactions does not imply making the organization attractive to all individuals; accurate perceptions can lead to applicant withdrawal. Black-
White test-taking attitude differences may not affect group differences in applicant withdrawal (Schmit & Ryan 1997), but they might, which would add to adverse-impact concerns (Chan 1997).

Chan et al (1998) discovered that (a) pretest reactions affected cognitive ability test performance, and test performance affected posttest reactions, and (b) personality test performance was unrelated to either pretest or posttest reactions. Integrity test results paralleled the personality test conclusions, and the overt integrity test had greater perceived job relatedness than the personality-based test (Whitney et al 1999).

Test outcome (passing or failing) contributes strongly to subsequent applicant reactions. Applicants passing a test for a clerical position rated organizational attractiveness, intentions to work for the organization, and test fairness higher than their initial reactions (Bauer et al 1998). Test outcome affects the perceived fairness of the hiring decision much more than the selection ratio (Thorsteinson & Ryan 1997). Ployhart et al (1999) determined that fairness perceptions of cognitive and job knowledge tests increased with a positive selection outcome and with sensitively conveyed personal and procedural information regardless of outcome. Selected applicants’ self-perception improved with personal or procedural information; rejected applicants’ self-perception declined. Sensitive explanations amplified this result, implying that providing information about the selection procedure would increase fairness and organizational perceptions but be counterbalanced by lower self-perceptions for rejected applicants.

Invasiveness of personnel selection measures has been investigated. Verifiable, impersonal, and face-valid biodata items tend to be perceived as less invasive, especially for individuals who understand the general purpose of biodata (Mael et al 1996a). Ways to obtain potentially invasive information without violating applicants’ needs for privacy have been offered [e.g. explaining the job relevance of the item (Mael 1998)].

EMERGING TOPICS

Team Member Selection

Organizations increasingly use team-based structures for organizing, motivating, and performing work (see Guzzo & Salas 1995, Howard 1995, Kehoe 1999, Klimoski & Zukin 1999, Kraut & Korman 1999b, O’Neil 1997, Sundstrom 1999). Much has been learned about factors affecting team performance and effectiveness (see reviews by Cohen & Bailey 1997, and West & Allen 1997), but more work remains for personnel selection. Selection systems need to consider differences between team selection and traditional selection methods, particular work team circumstances (task type, role differentiation, and resources), and selection into new versus preexisting work teams.
Individual characteristics and types of tasks interact within a team to influence team performance and effectiveness. On a creative problem-solving task, a mid-range of extraverts appears best; too many or too few depress performance ratings slightly. Conscientiousness did not predict team performance on the creative task (Barry & Stewart 1997), as Hough (1992) found at the individual level of analysis. In other team studies (LePine et al 1997, Barrick et al 1998), conscientiousness did predict task performance. Given a designated leader and group members each with unique expertise, team decision-making accuracy over time tends to be best when all members are high in conscientiousness and general cognitive ability (LePine et al 1997). Barrick et al (1998) found that conscientiousness, general cognitive ability, and extraversion all predicted overall team performance ratings in manufacturing work teams, where team members contributed independently to the outcome. Team agreeableness predicted teamwork for those tasks in which intergroup conflict was possible. One disagreeable team member was often enough to disrupt team performance, indicating the importance of interpersonal skills when selecting for some teams. Negotiation, an interpersonal skill, was validly measured with a simulation exercise (O’Neil et al 1997a,b). Significant advances in team selection research await good taxonomies of “team difference” variables (the analog of individual-difference variables for individuals) and situational variables relevant to teams.

Cross-Cultural Selection Issues

With their expanding global markets, culturally diverse work teams, and expatriate work assignments, international and multinational organizations place new demands on selection processes and measurement tools. Validities of domestic selection instruments may not generalize to international sites, because different predictor and criterion constructs may be relevant, or, if the constructs are the same, the behavioral indicators may differ. Interpersonal skill, open-mindedness, and adaptability are important factors for expatriate success, and family situation is the most commonly cited reason for failure (Arthur & Bennett 1995, Nyfield & Baron 1999). The vast majority of companies base their expatriate selection decisions on technical competence alone (Aryee 1997), so finding a very high failure rate among expatriates is unsurprising [between 15% and 40% (Shackleton & Newell 1997)]. A clear need for improving expatriate selection exists.

Several personality inventories originally developed in English have demonstrated similar psychometric properties across languages and cultures (see Katigbak et al 1996, McCrae & Costa 1997, Nyfield & Baron 1999). The International Committee on Test Standards produced a set of stringent standards for translating tests into another language (see Hambleton 1999). Psychologists from many different cultures might be involved in all phases of inventory development and validation, a strategy used to develop the Global Personality Inventory (Schmit et al 1999). Cultural variables likely moderate the validity of selection procedures. The House et al (1997) review concluded that Hofstede’s (1980) four constructs
HOUGH ■ OSWALD

(power distance, uncertainty avoidance, individualism vs collectivism, and masculinity vs femininity) described and differentiated cultures most usefully.

PROFESSIONAL, LEGAL, AND ETHICAL STANDARDS

Professional

Three initiatives sponsored by the American Psychological Association provide guidelines and policies regarding test-taker rights and responsibilities, test standards, and test-user qualifications. First, the American Psychological Association is in the final stages of approving 10 test-taker rights and corresponding responsibilities (Joint Committee on Testing Practices 1999). Second, the “Standards for Educational and Psychological Testing” (unpublished manuscript) revises and updates standards for psychological tests and for psychological measurement in general. All parties involved in employment decisions involving psychological assessment should become familiar with these new standards. Third, the “Test User Qualifications” document is currently under review (Fox 1999). These three documents will impact personnel selection practices and have significant legal and ethical implications.

Legal

Over the past 20 years or so, more employment litigation has been brought under common-law torts than under federal or state equal employment opportunity statutes (Highberger 1996). Nevertheless, legal challenges to personnel selection decisions are often based on the Civil Rights Acts, Americans with Disabilities Act (ADA), and Age Discrimination in Employment Act. In the vast majority of these cases, the Equal Employment Opportunity Commission is typically not involved (Sharf & Jones 1999), although commission guidelines provide important compliance information for the public and the courts. Unstructured interviews account for the majority of federal court cases involving selection tools, followed by cognitive ability tests, and physical ability tests; together, they were judged to be discriminatory in about 40% of the cases, with cognitive tests faring somewhat better (Terpstra et al 1999). For important practical guidance and discussions of the many issues, risks, and myths regarding fair employment as well as trends in employment litigation, see Barrett (1996, 1998), Jeanneret (1998), and Sharf & Jones (1999).

The U.S. Supreme Court significantly limited the scope of ADA, ruling that impairments should be evaluated in their corrected or “mitigated” state (Sutton vs United Airlines 1999, Murphy vs United Parcel Service 1999). The U.S. Equal Employment Opportunity Commission (1999) answered 46 frequently asked questions pertaining to employers’ legal obligations and the rights of the disabled in both the application and employment settings. Tippins (1999) provided practical ADA guidance for several testing scenarios, and Bruyère (1999) outlined
the psychologist’s role in upholding ADA provisions in all phases of the employment process.

Affirmative Action and Reverse Discrimination Although individuals differ greatly in their perceptions of affirmative action programs (Kravitz et al 1997), negative consequences consistently occur when employees or applicants believe hiring is based on group membership rather than merit (Heilman 1996, Heilman et al 1998, Kravitz et al 1997, Stanush et al 1998). Affirmative action programs are associated with slight improvement in employment conditions for women and racial minorities and appear to have virtually no effect on organizational effectiveness (Kravitz et al 1997).

No general agreement exists on how to prevent discrimination or remedy past discrimination (Campbell 1996). Reverse-discrimination court cases have clarified that race or other job-irrelevant class membership cannot be used when making employment-related decisions. Therein lies a conflict: The Uniform Guidelines indicate that organizations should seek out valid non- or less-discriminating predictors, yet developing a selection system with such measures requires attention to class membership. In *Hayden vs County of Nassau* (1999), the claim that an entrance exam designed to minimize discriminatory impact on minority job candidates necessarily discriminated against nonminority job candidates was ruled to be without merit. This case sets a precedent in affirming the reasonableness of designing selection systems to minimize adverse impact against protected groups.

Ethical

Ethical issues in personnel selection are complex, context-specific, and relative to each concerned party. We refer the reader to two important new sources in the field. Lowman (1998) authored an updated ethics casebook for human resource professionals practicing within organizations, and Jeanneret (1998) discussed ethical issues involved in individual assessment, detailing the responsibilities for both assessors and organizations.

PARTING REMARKS

New areas in personnel selection are unfolding, and traditional areas continue to improve. *(a)* Greater conceptual and methodological attention has been devoted to understanding and predicting how organizationally relevant criteria might change over time. Given the present and future state of rapid change in the world of work, this line of research is critically important for improving personnel selection and overall organizational effectiveness. *(b)* Personality theory and measurement within a personnel selection context have burgeoned. New personality constructs and compound constructs of well-known traits are being brought into
the fold. (c) Applicant reactions to personnel selection procedures have been energetically studied. (d) Team member and cross-cultural selection issues have drawn greater research attention. (e) Refined taxonomic structures are being developed across many different domains in personnel selection, both from the worker and work perspectives. We predict that selection systems will become more complex as a consequence of all this work; selection systems will mirror today’s realities and prove to be more effective and rewarding for individuals and organizations alike. We also predict that Guion’s (1998) book on personnel selection will fast become a classic. It is readable, practical, thoughtful, and thorough, and it assures its readers that much has been learned in personnel selection. Although the turn of the millennium marks a distinct ending and beginning, personnel selection theory and practice remain in constant process, looking toward the future while remembering the past.


LITERATURE CITED

Ackerman PL, Heggestad ED. 1997. Intelligence, personality, and interests: evidence for overlapping traits. Psychol. Bull. 121:219–45
Arthur W, Doverspike D, Barrett GV. 1996. Development of a job analysis-based procedure for weighting and combining con-
tent-related tests into a single test battery score. *Pers. Psychol.* 49:971–85


Bushman BJ, Wells GL. 1998. Trait aggressiveness and hockey penalties: predicting
hot tempers on the ice. J. Appl. Psychol. 83:969–74
Chan D. 1998a. The conceptualization and analysis of change over time: an integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). Organ. Res. Methods 1:421–83
Chan D. 1998b. Functional relations among constructs in the same content domain at different levels of analysis: a typology of composition models. J. Appl. Psychol. 83:234–46
Cohen SG, Bailey DE. 1997. What makes teams work: group effectiveness research from the shop floor to the executive suite. J. Manage. 23(3):239–90
Collins JM, Schmidt FL, Sanchez-Ku M, Thomas LE, McDaniel M. 1999. Predicting assessment center ratings from cognitive ability and personality. See Quiñones 1999
Conway JM, Jako RA, Goodman DF. 1995. A meta-analysis of interrater and internal con-
sistency reliability of selection interviews. J. Appl. Psychol. 80:565–79


Goldstein H, Riley Y, Yusko KP. 1999. Exploration of Black-White subgroup differences on interpersonal constructs. Subgroup dif-


Hayden vs County of Nassau. No. 98–6113, 1999. WL 373636 (2nd Cir June 9, 1999)


Hornick CW, Fox KA, Axton TR, Wyatt BS. 1999. The relative contribution of conditional reasoning and multiple intelligence measures in predicting firefighter and law enforcement officer job performance. See Williams & Burroughs 1999
Howard A. 1999. Discussant comments. See Quiñones 1999
Kiker S, Motowidlo S. 1999. Main and interaction effects of task and contextual performance on supervisory reward decisions. J. Appl. Psychol. 84:602–09
Legree PJ, Pifer ME, Grafton FC. 1996. Correlations among cognitive abilities are lower for higher ability groups. Intelligence 23:45–57
Migetz DZ, McIntyre M, James LR. 1999b. Measuring reliability among contingent workers. See Williams & Burroughs 1999
Murphy KR, Myors B. 1999. Testing the hypothesis that treatments have negligible effects: minimum-effect tests in the general linear model. J. Appl. Psychol. 84:234–47
Ones DS, Viswesvaran C. 1998a. Gender, age, and race differences on overt integrity tests: results across four large-scale job applicant data sets. J. Appl. Psychol. 83:35–42


Truxillo DM, Bauer TN. 1999. Applicant reactions to test score banding in entry-level


