Determinants, Detection and Amelioration of Adverse Impact in Personnel Selection Procedures: Issues, Evidence and Lessons Learned

Leaetta M. Hough, Frederick L. Oswald and Robert E. Ployhart*

Mean subgroup (gender, ethnic/cultural, and age) differences are summarized across studies for several predictor domains – cognitive ability, personality and physical ability – at both broadly and more narrowly defined construct levels, with some surprising results. Research clearly indicates that the setting, the sample, the construct and the level of construct specificity can all, either individually or in combination, moderate the magnitude of differences between groups. Employers using tests in employment settings need to assess accurately the requirements of work. When the exact nature of the work is specified, the appropriate predictors may or may not have adverse impact against some groups. The possible causes and remedies for adverse impact (measurement method, culture, test coaching, test-taker perceptions, stereotype threat and criterion conceptualization) are also summarized. Each of these factors can contribute to subgroup differences, and some appear to contribute significantly to subgroup differences on cognitive ability tests, where Black–White mean differences are most pronounced. Statistical methods for detecting differential prediction, test fairness and construct equivalence are described and evaluated, as are statistical/mathematical strategies for reducing adverse impact (test-score banding and predictor/criterion weighting strategies).

Introduction

When different groups (e.g., within ethnic/cultural, gender, age groups) score on average differently, adverse impact, or differential hiring rates within groups can result. Many hypotheses have been proposed as causes and explanations for adverse impact. In this article we summarize results and describe meta-analyses we conducted that identify subgroup mean-score differences according to individual-difference constructs (cognitive ability, personality and physical ability). Although the specific data and legal context are generally US-based, the results illustrate the types of differences in predictor construct scores, selection decisions and legal contexts (regarding employment discrimination) that one would encounter internationally. We review causes and strategies for reducing subgroup mean differences that apply in the broader international context. We discuss the research regarding the causes and explanations for group differences according to the following factors: measurement method, culture, test coaching, test-taker perceptions, others' perceptions (e.g., stereotype threat) and criterion composition. Finally, we discuss methodological and practical issues in applying and interpreting the weighting of predictors and criteria, test-score banding, adverse impact analyses, differential prediction analyses and construct equivalence methods. Throughout, we discuss lessons learned that raise our understanding of how to reduce mean-score differences between groups and the unfair adverse impact that results.

Group Mean-Score Differences in Individual Difference Variables

Taxonomic structures of individual differences useful for personnel selection exist (cf. Ackerman and Heggestad 1997; Fleishman and Quaintance 1984), yet a good deal of disagreement remains about the details of the taxons themselves. Take, for example, the various definitions and theories of intelligence. Many theorists tend to focus on cognitive abilities in their definition of intelligence (e.g., Jensen 1972; Spearman 1927; Thurstone 1938), whereas some theorists expand

Address for correspondence:
*Leaetta M. Hough, The Dunnette Group, Ltd., 370 Summit Avenue, Saint Paul, Minnesota 55102, USA. E-mail: leaetta@msn.com

© Blackwell Publishers Ltd 2001, 108 Cowley Road, Oxford OX4 1JF, UK and
350 Main Street, Malden, MA 02148, USA.

the definition of intelligence to include social abilities, musical abilities, and more (e.g., Gardner 1993; Goleman 1995; Guilford 1967; Mayer and Geher 1996; E. Thorndike 1920; R. Thorndike 1936). Indeed, when 24 prominent theorists were asked to define intelligence, they gave 24 different definitions with varying degrees of overlap (Sternberg and Detterman 1986).

Research has nonetheless resulted in several more-or-less agreed-upon general domains of skills and abilities. We summarize the literature regarding age, gender, and ethnic/cultural group differences for three major types of skills and abilities that industrial-organizational psychologists tend to use in personnel selection decisions: cognitive abilities, personality and physical abilities. Results are shown in Tables 1–4. International research may seek to discover other important subgroup differences for these skills and abilities, depending on the types of subgroups that would be relevant and applicable (e.g., religious affiliation in Northern Ireland, Aborigines vs. Australians of European descent).

Cognitive Abilities

Cognitive ability can be envisioned as a hierarchical structure with a general intelligence (*g*) at the apex, with more refined ability factors as one descends the hierarchy (cf. Carroll 1993; Gottfredson 1998; Gustafsson 1999; Lubinski and Dawis 1992; Lubinski 2000; Neisser *et al.* 1996). For example, Ree and Carretta (1994) factor analyzed the Armed Services Vocational Aptitude Battery (ASVAB) in a large sample of US high-school and college-age individuals, essentially replicating Vernon's (1950) hierarchical structure, with *g* as the general factor and Cognitive Speed, Verbal/Numerical and Technical Skills factors beneath. When more diverse batteries are factor analyzed, for example, batteries that contain additional measures of cognitive processes, somewhat different hierarchical structures emerge, although a general ability factor is at the apex (e.g., Roznowski, Dickter, Hong, Sawin and Shute 2000). We summarize the existing literature on group-mean differences at the level of *g* and at levels that are narrower theoretically but are practically useful (and used in practice) in employment testing.

General Intelligence

Ethnic/cultural group differences. Regarding general intelligence, the commonly accepted mean difference between Blacks and Whites is about one standard deviation, with Blacks scoring lower than Whites (Arvey *et al.* 1994; Herrnstein and Murray 1994; Hunter and Hunter 1984; Jensen 1980; Loehlin, Lindzey and Spuhler 1975; Neisser *et al.* 1996; Reynolds, Chastain, Kaufman and McLean 1987; Sackett and Wilk 1994; Williams and Ceci 1997). This may be an accurate general statement, but a closer inspection of the research data suggests some important underlying heterogeneity: Black–White group-mean differences differ across employment settings for a variety of reasons.

Table 1: Summary of standardized mean-score differences on cognitive ability constructs for ethnic, gender, and age groups

Comparison group	*g*	Verbal ability	Quantitative ability	Science achievement	Spatial ability	Memory	Mental processing speed
Blacks/Whites	−1.0	−.6	−.7	−1.0	−.7	−.5	−.3
Hispanics/Whites	−.5	−.4	−.3	− .6	–	–	−.4
East Asians/Whites	.2	–	–	–	–	–	–
Women/Men	.0	.1	−.2	−.2	−.4	.0	.0
Older/Younger Working Adults	− .4[2]	–	–	.0	−.2[3]	–	−.6[3,4]
		Crystallized intelligence			Fluid intelligence		

Notes: Approximate mean *d*[1]. All mean *d* values may have significant variation associated with the specific type of ability as well as other factors such as the type of sample, occupation, setting, and measure.

[1] Negative *d* values mean the group listed first scores lower.

[2] Assumes average 'younger' age is 30 and average 'older' age is 50. Calculations based on a Flynn effect of $d =$ .2 for every decade. (Hartigan and Wigdor (1989) found a GATB composite of verbal and quantitative ability tests correlated −.28 with age.)

[3] Hartigan and Wigdor (1989) found a GATB composite of spatial/perceptual speed and accuracy tests correlated −.45 with age.

[4] Hartigan and Wigdor (1989) found a GATB composite of dexterity tests correlated −.52 with age.

Roth, BeVier, Bobko, Switzer and Tyler (in press) meta-analyzed Black–White mean differences on measures of general intelligence and found a mean difference of one standard deviation tends to be accurate for *overall* analyses of job *applicants* in *corporate* settings, but this finding was moderated by two important factors. *Job complexity* may moderate the group-mean differences because of both self-selection effects and direct selection effects. Regarding self-selection effects, research finds that higher-ability individuals tend to gravitate toward more complex jobs, and conversely, lower-ability individuals tend to gravitate toward less complex jobs (Sackett and Wilk 1994; Desmarais and Sackett 1993; Wilk, Desmarais and Sackett 1995). Regarding direct selection effects, more complex jobs may select a smaller proportion of job applicants (i.e., have lower selection ratios) leading to smaller d values. In fact, this is what the Roth *et al.* meta-analysis found: the standardized mean-score difference (d) between Blacks and Whites applying for high-complexity jobs was .63, versus .86 for low-complexity jobs.

The *study design* was another important moderator. For within-job study designs, d = .83 for job applicants in industrial settings (similar to N. Schmitt *et al.* 1996), but across jobs d = 1.0 to 1.23 for applicants. As one might expect with selection effects, the d value was smaller in employee samples: .38 for within-job studies and .92 for across-job study designs.

Hispanic and White mean scores on general intelligence tests also differ. Research shows that Hispanics tend to score lower than Whites, although the mean difference is smaller than the difference between Blacks and Whites (Herrnstein and Murray 1994). N. Schmitt *et al.* (1996) found an average d = .48 across studies, Hispanics scoring lower than Whites. The Roth *et al.* (in press) meta-analysis found an average d = .72 across studies, but for only industrial tests of cognitive ability (excluding the Wonderlic) d = .58, closer to N. Schmitt *et al.*'s (1996) d = .48.

East Asians – Japanese, Chinese and perhaps Koreans as well – appear to have a somewhat higher level of measured general intelligence on average than do Whites (Herrnstein and Murray 1994; Lynn 1987, 1991). Average differences appear to range from zero to ten IQ points on general intelligence; Herrnstein and Murray (1994) found an average difference of three points (d = .20).

Gender differences. Most standard tests of general intelligence are designed so that men and women have the same overall mean score (Neisser *et al.* 1996). The variability (standard deviation) of test scores, however, is somewhat larger for men than for women (Herrnstein and Murray 1994). In other words, more men than women have scores at either the high end or the low end of the score distribution. Although men and women may have similar means on general intelligence, they have different means on specific abilities, as will be discussed later.

Age differences. During much of the twentieth century, the generally accepted wisdom of most intelligence researchers said that scores on general intelligence measures peaked in early adulthood (one's early twenties), and from that point declined gradually and steadily into old age (Botwinick 1967; Troll 1975). In the 1970s, researchers realized that although cross-sectional studies (studies with age-diverse samples at one point in time) showed a significant decrement across decades, longitudinal studies (studies tracking the same individuals across time points) tend not to show as strong a decline across decades (e.g., Botwinick 1977; Schaie and Strother 1968). Cross-sectional studies reveal important cohort effects – for example, increased educational opportunities may translate into average increases in measured ability over time. Longitudinal studies are more informative about patterns of intellectual decline; those studies indicate that general intelligence test scores tend to remain stable from early adulthood up through around age 50, at which time scores typically gradually decline into old age (Minton and Schneider 1980). The pattern of decline varies depending upon specific ability (see Baltes and Schaie 1976, for a different point of view). For example, fluid intelligence, such as mental processing speed, tends to decline whereas crystallized intelligence, such as verbal and quantitative ability, does not (Horn and Donaldson 1976).

A considerable amount of data suggests a worldwide rise in IQ scores of about three points (d = .20) per decade, the so-called *Flynn effect* (Flynn 1984, 1987, 1998; Teasdale and Owen 1989). In short, we tend to score appreciably higher on IQ tests than our grandparents did at a similar age. The cause of the Flynn effect is unknown, although researchers have offered some tentative causal hypotheses (cf. Neisser 1998). Regardless of its origins, one likely effect is adverse impact against older working adults when cognitive ability tests are used for personnel selection. The extent of adverse impact would not be as great as the typical Black–White effect for cognitive ability, but it is present, and we speculate that in some cases it may have an accentuating effect on race differences (e.g., older Black group vs. younger White group).

Using data from industrial settings, Hartigan and Wigdor (1989) calculated the correlation between mean age and mean General Aptitude

Test Battery (GATB) test scores in 715 studies ($N = 77141$). Ideally, they would have calculated the effect of age by computing correlation coefficients separately for people in different age groups; however, their data reported only mean age for each study. Although they did not have a measure of *g*, their general verbal/numerical composite correlated −.28 with age.

US employment law as it relates to personnel decision-making is less concerned with the stability of an individual's general intelligence over time as it is the finding that at any particular point in time (e.g., when job applicants take an ability test), older adults will tend to score lower than younger adults. Regardless of the cause, differences in mean scores can open the door to some amount of adverse impact with respect to age.

Summary. Age and ethnic/cultural differences on general cognitive ability tests exist, but by design, men and women on average score similarly. Of the ethnic/cultural groups reviewed, East Asians generally have the highest mean-score on general intelligence tests followed by Whites, Hispanics and Blacks. The magnitude of the difference depends upon the sample, such as military vs. industrial, job applicant vs. job incumbent, high vs. low complexity jobs, within-job vs. across-job study design.

Regarding age differences in general intelligence, general intelligence appears reasonably stable for people of employment age in most developed countries. Nonetheless, within an age-diverse group of people, general intelligence tests are likely to produce some adverse impact against those who are older.

Many studies provide evidence that for different ethnic and cultural groups *d* varies according to the specific cognitive ability measured (e.g., Herrnstein and Murray 1994; Hyde and McKinley 1997; Mullis, Dossey, Foertsch, Jones and Gentile 1991; Neisser *et al*. 1996; Roth *et al*., in press; N. Schmitt *et al*. 1996; Storfer 1990). Examining these differences provides important insights and suggests appropriate strategies for reducing adverse impact against protected groups.

Verbal Ability

For the increasing number of organizations that have an international focus in the employees they hire and in the clients with whom they deal, effective verbal skills and the criterion-related validity of particular verbal tests may require some rethinking. We speculate that perhaps broader verbal ability tests will be more relevant for these organizations, tests such as verbal communication may yield better criterion-related validities than more narrow verbal tests such as vocabulary or even verbal fluency, especially for those employees that are not native speakers of the dominant language of the organization. Another issue with verbal ability is of special concern in international employment testing: verbal ability contaminates a measure (differentially distorts true scores) when language proficiency varies in a sample, and the level of verbal ability needed to complete the test is greater than the level of verbal ability needed to perform on the job successfully (Sireci and Geisinger 1994; AERA, APA, NCME 1999, Standard 9.10).

Ethnic/cultural group differences. Across three different studies, Blacks on average scored lower than Whites on verbal tests: Mullis *et al*. (1991), using the National Assessment of Education Progress study data, found $d = .71$ for Black and White 17-year olds on the reading test (Hauser 1998, p. 234). Similar Black–White differences were found in the Roth *et al*. (in press) meta-analysis, which reports average $d = .76$ on industrial tests of verbal ability, and the N. Schmitt *et al*. (1996) meta-analysis, where $d = .55$ on verbal ability tests. For Hispanic–White verbal ability differences, Roth *et al*. (in press) meta-analyzed found average $d = .40$ on industrial tests of verbal ability, Hispanics scoring lower on average.

Gender differences. Maccoby and Jacklin's (1974) qualitative review on gender differences concluded that women tend to score higher than men on verbal ability measures. The Hyde (1981) meta-analysis of the Maccoby and Jacklin (1974) studies supports this conclusion, finding median $d = .24$, women scoring higher on average. More recent meta-analyses and large-sample analyses find smaller differences in the same direction (Hedges and Nowell 1995; Hyde and Linn 1988).

Regarding specific verbal abilities, Hyde and Linn (1988) conducted separate meta-analyses for different types of verbal ability, finding that women tended to score higher than men on anagrams ($d = .22$), speech production ($d = .33$), and general verbal ability ($d = .20$), whereas men tended to score higher on analogies ($d = .16$). A reanalysis of six large US adolescent data sets by Hedges and Nowell (1995) found that women on average scored only slightly higher than men on reading comprehension ($d = .09$) and vocabulary ($d = .06$). Others have found rather large effect sizes favoring women on tests of synonym generation and verbal fluency (similar to speech production), with effect sizes ranging from $d = .5$ to 1.2. In summary, men and women on average score similarly on verbal ability tests, except for speech production or verbal fluency tests, on which women score higher.

Age differences. On traditional verbal ability tests, older adults within normal working-age range score about the same as they did when they were younger (Botwinick 1967, 1977; Minton and Schneider, 1980). In fact, research suggests that verbal abilities are stable or actually increase up to late stages of adult development (e.g., Horn, 1968, 1982, 1989; Horn and Hofer, 1992). The so-called *Flynn effect*, or increase in intelligence of about .2 standard deviations per decade, is not well understood, but according to Flynn (1998), it seems to be due more to increases in fluid intelligence (e.g., working memory and spatial ability), than to increases in crystallized intelligence (e.g., verbal ability).

Quantitative Ability

Ethnic/cultural group differences. The National Assessment of Education Progress (NAEP) quantitative ability data on 17-year-olds (Mullis *et al*. 1991) found the Black–White d = .68, Blacks tending to score lower than Whites (Hauser 1998, p. 234); similarly, the N. Schmitt *et al*. (1996) meta-analysis found d = .64, and the Roth *et al*. (in press) meta-analysis found d = .76 for industrial tests. The Hispanic–White d = .28, Hispanics scoring lower than Whites on average.

Gender differences. Hyde (1981) meta-analyzed Maccoby and Jacklin (1974) study results on math ability tests and found median d = .43, with men tending to score higher. More recent meta-analyses using a greater number of studies resulted in much less difference between men and women on quantitative ability. Hyde, Fennema and Lamon (1990) found d = .15 and Hedges and Nowell (1995) found d = .16, men scoring higher than women on average in both studies.

Hyde *et al*. (1990) conducted separate meta-analyses on specific quantitative abilities: computation (simple, memorized mathematical facts), concepts (analysis or comprehension of mathematical ideas), and problem solving (extension of mathematical knowledge or its application to new situations). Overall, men and women scored essentially the same on computation and concepts, but men tended to score higher than women on problem solving (d = .32).

In samples of mathematically precocious youth (ages 12 to 13), Benbow (1988) found gender differences consistently favoring boys, d = .39. This d value was based on high-ability youths. Feingold's (1992) meta-analytic finding that the variance in the male distribution of math test scores is greater than the female variance on math tests implies a higher d value in groups of individuals with high overall ability.

Age differences. Similar to tests of verbal ability, adults within normal working-age range have stable scores on tests of numerical ability. Numerical ability appears to remain relatively stable until late adulthood (late sixties) after which notable decline occurs (Horn and Donaldson 1976).

Summary. On quantitative ability tests, men generally score somewhat higher than women – in the range of d = .15 – with little or no differences on computation and concepts but larger differences on problem solving (d = .32). Blacks score lower than Whites on average, d from .65-.75. For Hispanics compared with Whites, d is lower, in the high .20s. Older adults of working age score about the same as younger adults. Thus, for jobs requiring math skills, selecting high-ability individuals on math ability would tend to accentuate gender differences in hiring rates but not age differences.

Spatial Ability

Ethnic/cultural group differences. The N. Schmitt *et al*. (1996) meta-analysis reports a Black–White mean-score difference on spatial ability tests of d = .66, with Blacks scoring lower than Whites.

Gender differences. On spatial abilities, the Hyde (1981) meta-analysis of studies in Maccoby and Jacklin (1974) finds median d = .45, men scoring higher than women on average. Recent meta-analyses present more complex conclusions. Two sets of researchers (Linn and Petersen 1985; Voyer, Voyer and Bryden 1995) meta-analyzed mean-score differences on specific spatial abilities: mental rotation (mentally rotating a three-dimensional object depicted in two-dimensional space, e.g., Mental Rotations Test), spatial perception (determining horizontality or verticality, e.g., Water-Level Test), and spatial visualization (visually locating a simple figure within a complex one, e.g., Embedded Figures Test). Men scored higher than women on all three types of spatial ability tests, but the magnitude of the differences across abilities varied in a systematic way. Linn and Peterson (1985) found d = .73 for mental rotation, d = .44 for spatial perception, and d = .13 for spatial visualization; Voyer *et al*. (1995) found a similar pattern: d = .56, .44, and .19, respectively. The Masters and Sanders' (1993) meta-analysis also yielded large gender differences on tests of mental rotation. Spatiotemporal or dynamic spatial tasks (e.g., tracking a moving object in space) produce d values in the mid-.30s (Law, Pellegrino and Hunt 1993). Thus, mean differences by gender are large for tests of mental rotation but small-to-nonexistent for tests of spatial visualization.

Age differences. Longitudinal studies suggest that spatial ability improves up to age 40, plateaus through the mid-fifties, and declines thereafter, whereas cross-sectional studies suggest a gradual decline from the late twenties (Horn and Donaldson 1976). Spatial ability seems to be one ability contributing to the Flynn effect, the steady cohort increase in measured intelligence during the twentieth century (Flynn 1998).

The present authors computed d values comparing the mean-scores for working adults 40 years of age and older ($N = 1153$) and adults younger than 40 years of age ($N = 2281$) on a spatial ability test (cf. Hough, Carter, Dohm, Nelson and Dunnette 1993). The finding, $d = .24$, suggests that spatial ability test scores tend to favor younger working adults somewhat.

Summary. On average, Blacks score lower than Whites on spatial ability tests, d in the mid-.60s. Men tend to score higher than women on tests of spatial ability, though the magnitude of the differences varies depending upon the type of spatial ability: Men score significantly higher than women mental rotation tests ($d \sim .70$), higher on spatial perception tests ($d \sim .45$), higher on dynamic spatial tests ($d \sim .35$), but only modestly higher on spatial visualization tests ($d \sim .15$). Older adults tend to score lower than younger adults on spatial ability tests.

Memory

Primary memory (often called short-term memory) is the part of the memory system involved with material in conscious awareness (i.e., still being rehearsed by the individual). *Secondary memory* is the part of the memory system in which information is stored and from which information may be retrieved. Tests of primary memory require the person to remember information while carrying out an information-processing task (Salthouse 1991). Not all primary memory tests are tests of working memory. For example, memory-span tests, such as digit span and digit symbol substitution, are generally considered primary memory tests that, in fact, are not highly correlated with other measures of primary or secondary memory (Carroll 1993; Martin 1978). These are important distinctions because subgroup-mean score differences, as will be seen below, vary according to type of memory test.

Ethnic/cultural group differences. Verive and McDaniel (1996) meta-analyzed mean-score differences between Blacks and Whites on memory-span tests (e.g., digit span, digit symbol substitution), tests that are distinct from associative, free-recall, meaningful, and visual memory tests/factors (Carroll 1993). The tests used numbers as stimuli, lacked a study period, and required immediate recall. Estimated population $d = .48$ ($k = 31$ studies, total $N = 27973$), Whites scoring higher on average. This d value is less than the Black–White d value of 1.0 for traditional cognitive ability tests. In addition to lower mean-score differences, the corrected validity for predicting job performance was .41, suggesting that memory-span tests may yield reasonable criterion-related validities and less adverse impact for Blacks at the same time.

Gender differences. Overall, men and women manifest some differences on memory tests. Maccoby and Jacklin (1974) explained that at least some of the difference was due to the content of the test material. Namely, women scored better than men on memory tests containing verbal content, whereas minimal differences existed on memory tests containing objects and digits. In addition, women appear to have better memory for social stimuli; generally, women are shown to be better at remembering people's names and faces (Bahrick, Bahrick and Wittlinger 1974).

Age differences. Evidence clearly shows that memory capacity declines with age. This difference may relate more to a decline in secondary memory more than in primary memory (Craik 1977). Some research (Craik 1977; Reese 1976) suggests that the age-related memory deficit is not the result of failure to perceive the material but from difficulties in storing and/or retrieving the information, although these findings should be corroborated with job applicant data. It does appear, nonetheless, that working memory and secondary memory both show greater declines with age. Tests of very simple memory, such as digit span (memory span), may not produce as large a mean-score difference between older and younger working adults (Minton and Schneider 1980), but the criterion-related validity and utility of simple memory tests vs. traditional cognitive ability tests in predicting job performance for younger and older job applicants appears to be unknown. In general, memory tests tend to produce mean-score differences between older and younger working adults and thus may produce adverse impact.

Summary. When comparing Black and White mean-score differences, d values for memory-span tests are smaller than those for general intelligence tests. Thus, memory span tests are likely to have less adverse impact against Blacks than are tests of general cognitive ability, and meta-analytic evidence indicates that criterion-related validities of memory-span tests compare favorably with those of general cognitive ability tests ($r = .41$ in the study above vs. the meta-analytically corrected $r = .51$ reported in Schmidt and Hunter 1998).

Women tend to score somewhat higher than men do on memory tests. Most of the difference, however, depends upon the context of the stimulus items. Women have better memory for verbal and social content. Minimal differences between the sexes exist for memory for symbols and digits.

Older adults score lower on memory tests than younger adults, although the mean-score difference appears to be smaller for memory-span tests, e.g., digit span tests, than for other primary memory tests or tests of working memory or secondary memory. Thus, integrating the ethnic and age findings, including a short-term memory test in a test battery rather than a general intelligence test, may reduce adverse impact against Blacks while leaving mean-score differences between older and younger working adults unaffected.

Mental Processing Speed (Cognitive Speed and Decision Speed)

In Carroll's (1993) three-stratum theory, cognitive speed and decision speed are each at the same level as fluid and crystallized intelligence. In Carroll's model, tests of ideational fluency, figural fluency, association fluency, naming facility are indicators of cognitive speed, whereas simple reaction time, choice reaction time and comparison time are indicators of decision speed. Analyses by Roberts, Pallier and Goff (1999) distinguish decision time and movement time, although both correlate highly with clerical speed.

Ethnic/cultural group differences. N. Schmitt *et al.* (1996) meta-analyzed effect sizes of mean-score differences between Blacks and Whites on tests of clerical speed and accuracy, with $d = .15$. The type and size of the samples, however, were relatively small ($k = 2$, $N = 341$).

The present authors calculated effect sizes on a clerical speed and accuracy test administered to three large samples ($N = 15302$, with Black $N = 5527$, White $N = 9775$; see Dohm and Hough 1993; Hough *et al.* 1993). For Black–White differences, the sample-size-weighted $d = .35$, the Black mean lower than the White mean. The Hispanic–White sample-size-weighted was about the same, $d = .38$ ($N = 10270$, with Hispanic $N = 495$), Hispanics scoring lower than Whites.

Gender differences. Men tend to score higher than women on reaction-time tests (Anastasi 1958; Garai and Schneinfeld 1968). Women, on the other hand, tend to score higher on tests of manual dexterity; Maccoby and Jacklin (1974) find the advantage is focused on finger dexterity. Overall, women also score higher than men do on tests of clerical speed and accuracy (Minton and Schneider 1980).

The present authors computed standardized mean-score differences between men and women in the three large samples described above. In addition, the authors computed the d for the male and female sample means on the Visual Speed and Accuracy test reported in the technical manual of the *Employee Aptitude Survey* (Ruch and Ruch 1963). The sample-weighted $d = .12$, women scoring slightly higher than men (9 female samples and 34 male samples, total $N = 28091$).

Age differences. One of the more solid empirical research findings on age is that individuals experience a slowing down of speed in performing tasks as they get older. Older adults tend to score significantly lower than young adults on measures of choice reaction time (e.g., responding to each of several signals over time) and slightly lower on measures of simple reaction time (e.g., one signal and one response; see Welford 1977).

According to Welford (1977), the slowing is not due so much to muscular limitations as to the additional time needed to decide on and monitor the required sequence of movements. Similarly, according to Salthouse, many of the age-related decrements in fluid intelligence appear to be 'mediated by age-related reductions in working memory, which may in turn be largely mediated by age-related reductions in speed of executing simple processing operations' (1991, p. 179).

One of the large samples described above contained age data on the clerical speed and accuracy test, so the present authors computed d values between adults 40 years of age and older ($N = 1128$) and adults younger than 40 years of age ($N = 2281$). Younger working adults scored higher than older working adults, on average, $d = .63$. Hartigan and Wigdor (1989) calculated the correlation between mean age and mean GATB spatial/perceptual speed and accuracy test scores. Although this composite confounds differences that might exist separately for spatial ability and perceptual speed and accuracy, they found spatial/perceptual speed and accuracy test scores correlated -.45 with age ($k = 715$, $N = 77141$). Hartigan and Wigdor (1989) also found the GATB dexterity composite correlated -.52 with age ($k = 715$, $N = 77141$). In short, mental speed tests are likely to produce adverse impact effects against older working adults.

Summary. Older adults tend to score lower than younger adults on mental speed tests, as expected. Unexpectedly, women tend to score only marginally higher than men on clerical speed and accuracy tests. Men tend to score higher than women on reaction time tests. Blacks and Hispanics generally score lower than Whites on tests of clerical speed and accuracy, each with means about one-third of a standard deviation lower than Whites.

Science Achievement

Science achievement is a label with too much underlying heterogeneity to be considered a specific ability, although it distinguishes itself from other broad areas of achievement, such as verbal or artistic achievement. Several researchers have reported results on group mean-score differences on scientific achievement.

Ethnic/cultural group differences. Using the National Assessment of Education Progress study data on 17-year-olds, Mullis *et al*. (1991) found $d = 1.04$ for the science test section, Blacks tending to score a full standard deviation lower than Whites (Hauser 1998, p. 234). Roth *et al*. (in press) meta-analyzed ACT Science scores and found a Hispanic–White mean-score difference of $d = .58$, Hispanics scoring lower than Whites.

Gender differences. Fleming and Malone (1983) meta-analyzed studies of science performance for students ages 5 to 18 and found $d = .16$, male means somewhat higher than female means. Males and females score essentially the same in life sciences (e.g., biological sciences), $d = .02$, whereas in physical sciences (e.g., physics), males score higher than females, $d = .30$. Becker (1989) obtained a similar mean-score difference between males and females on science achievement, $d = .16$, males scoring higher than females, and she also found that d varied with the specific subject matter (e.g., $d = .35$ for physics, favoring males). The Hedges and Nowell (1995) reanalysis found somewhat larger d values for high school students, ds ranging from .11 to .50 on specific science areas, mean $d = .32$, male means higher than female means.

Age differences. Science achievement reflects both learning and experience in the science domain. As such, it reflects crystallized intelligence, an area in which older adults are likely to increase their scores well into adulthood. Thus, older working adults score higher than younger adults on general science knowledge and even more highly than younger adults on tests of highly specialized science knowledge (Ackerman and Rolfhus 1999).

Summary. Blacks, and to a lesser extent Hispanics, have tended to score lower than Whites on tests of science achievement ($d = 1.04$ and $d = .58$, respectively). In general, males score higher than females on tests of science, although the differences vary depending upon the specific subject matter. Males and females, on average, score essentially the same on life sciences (e.g., biological sciences), whereas males tend to score higher than females on the physical sciences (e.g., physics). Older working adults tend to score higher than younger working adults on tests of science knowledge, both specific and general science knowledge.

Crystallized Intelligence and Fluid Intelligence

Factor analyses of Thurstone's tests of primary mental abilities often produce two second-order factors: crystallized intelligence and fluid intelligence (Cattell 1971/1987; Horn and Cattell 1967; Horn 1970, 1975). *Crystallized intelligence* is measured primarily by tests of verbal comprehension, vocabulary, general factual information, and knowledge about academic subject matter (e.g., math, science, literature, social studies). Crystallized intelligence is dependent on the particular kinds of learning and training experiences to which an individual has been exposed (Horn 1976).

Fluid intelligence, by contrast, refers to reasoning facility and abstract relational skills (Horn 1976). Fluid intelligence tests involve figural and non-word symbolic materials, such as letter series, matrices, mazes, figure classifications, and word groupings, as well as the performance subtests of the Wechsler scales (e.g., block designs, picture arrangements, object assembly and picture completion). Tests of spatial ability, memory span, associate memory (e.g., recall in paired-associate learning), short-term memory and working memory are also considered part of fluid intelligence (Horn 1975; Horn and Donaldson 1976; Minton and Schneider 1980). Carroll (1993), however, in his three-stratum model of the structure of cognitive abilities separates spatial ability and memory tests from fluid intelligence and crystallized intelligence tests.[1]

Crystallized and fluid intelligence concepts are paralleled in several intelligence researchers' work: in Vernon's (1950) verbal/education and spatial/mechanical distinction (cf. Horn 1976), in Ackerman's (1996) intelligence-as-knowledge and intelligence-as-process, and Sternberg's (1985) content/knowledge components and information-processing components, respectively.

Age differences. Seemingly conflicting data regarding the relationship between age and cognitive ability become consistent and interpretable when cognitive abilities are grouped according to crystallized and fluid intelligence (Horn and Donaldson 1976). Starting early in adulthood, fluid intelligence measures show a steady decline, whereas crystallized intelligence measures increase or remain constant through adulthood (Horn 1975). Steady increases in average intelligence scores during the twentieth century have been primarily due to increases on fluid intelligence tests, not crystallized intelligence tests (Flynn 1998). Mean-score differences between older and younger adults tested at approximately the same time are likely to exist on tests of fluid intelligence, young working adults scoring higher than older working adults. Differences would be less pronounced for crystallized intelligence measures, unless those measures involve highly specialized knowledge, in which case older working adults tend to score more highly.

Lessons Learned

Mean-score differences between ethnic/cultural, gender, and age groups are summarized in Table 1. Perhaps the data shown there are no surprise. However, each mean *d* is a summary, and we want to emphasize that as a descriptive summary, it may not reflect what is found in any particular circumstance. The research described above clearly indicates that the setting, the sample, and the construct can all, either individually or in combination, moderate the magnitude of the differences between groups (i.e., meaningful heterogeneity exits across different variously defined 'situations'). We do not wish to revisit the 'situational specificity' debates in validity generalization (e.g., Schmidt and Hunter 1984); we only wish to say that informative and useful analyses exist at a 'medium' level of specificity, analyses that lie somewhere between (a) collapsing across all studies with different settings, samples, constructs, and measures; and (b) examining each study individually. Therefore, the amount of adverse impact observed in different settings can vary, and attempts to reduce adverse impact against one group (e.g., Blacks) may increase it for another (e.g., Women).

One of the important conclusions from the research described in this section is that the specific ability construct does moderate the magnitude of the effect size of mean differences between groups, often by amounts that are practically meaningful for personnel selection. For example, on tests of spatial ability, women score significantly lower than men on tests of mental rotation tasks but almost the same as men on tests of spatial visualization. Thus, the test user's choice of spatial ability test affects the extent of adverse impact against women. Similarly, measures of *g* have more or less adverse impact against older working adults depending on how much the measure weights fluid versus crystallized intelligence. Of course, the appropriate choice of predictor construct is largely dependent on the nature of the criterion; we discuss criterion issues in a later section.

Compared with scores on *g*, mean-score differences between Blacks and Whites on *specific* cognitive abilities are often less than one standard deviation lower than Whites. The fact that the Black–White mean difference on general cognitive ability tests has been about $d = 1.0$ should not lead to complacency in test developers and test users in selection settings. When the most appropriate cognitive ability for the job is included in a test battery, adverse impact against Blacks may actually be reduced.

Employers using tests in employment settings need to assess accurately the cognitive requirements of work. When the exact nature of the cognitive ability requirements of work is specified, the appropriate test of cognitive ability may not be a test of general intelligence. Instead, it may be that a specific cognitive ability, such as verbal ability, is the more appropriately measured cognitive ability. Especially for more complex jobs, hiring individuals on specific abilities becomes more important (Lubinski 2000).

Personality Constructs

Considerable research has examined the structure of personality traits. Many have embraced the Five-Factor Model (FFM), i.e., Extraversion, Agreeableness, Conscientiousness, Neuroticism (Adjustment), and Openness to Experience. A body of research finds these 'Big Five' personality factors to be stable and robust, covering the broad scope of normal personality (Saucier and Goldberg 1998; Wiggins and Trapnell 1997). Others, however, (e.g., Block 1995; Hough 1992; Loevinger 1994; McAdams 1992; Pervin 1994; R. Schneider and Hough 1995; Tellegen 1993) have criticized the FFM as too broad for theoretical understanding or for measuring personality traits for prediction purposes.

Hough (1997) critically investigated the adequacy and usefulness of the FFM for I/O psychology. Specifically, she presented data supporting significant heterogeneity in the extraversion and conscientiousness factors, heterogeneity that masked the differential prediction of more specific factors within FFM. Construct validity (i.e., understanding) and criterion-related validity (i.e., prediction) is clearer when Extraversion is separated into

Affiliation and Surgency (potency), and when Conscientiousness is separated into Achievement and Dependability. We summarize mean-score differences for ethnic/cultural, gender, and age groups for Neuroticism, Openness to Experience, Agreeableness, Affiliation, Surgency, Achievement, Dependability, as well as Hough's (1992) Rugged Individualism (masculinity/femininity). We also present *d* for social desirability scales and for integrity and managerial potential, two compound variables that combine FFM factors.

Several decision points and methodological steps were required to summarize the research literature. Correlations were converted from *r* to *d*. These *d* values were sample-size weighted to obtain mean *d* across studies at the facet level. We used the following strategy to obtain mean *d* for each Big-Five variable with studies reporting facet-level measures (e.g., Achievement and Dependability instead of Conscientiousness) as well as studies reporting global Big-Five measures (e.g., Conscientiousness). First, we sample-size weighted the *d* values within each facet to obtain mean *d* for each facet. Then, we equally weighted the mean facet-level *d* values to obtain a facet-based estimate of *d* for the Big-Five variable. We then calculated mean *d* for the Big-Five global measures by sample-size weighting each of the global Big-Five measure *d* values. We then sample-size weighted the mean facet-based estimate of *d* for the Big-Five variable and the mean global-based estimate of *d* for the Big-Five variable to obtain an overall estimate of *d* for the Big-Five variable. Other types of decisions and methods could have been pursued; we feel that ours is among the reasonable options.

Ethnic/cultural differences. Several researchers (Goldberg, Sweeney, Merenda and Hughes 1998; Hough 1998; Ones, Hough, Viswesvaran 1998; Ones and Viswesvaran 1998a) have examined personality differences between ethnic/cultural groups. We summarized mean score differences on personality scales for Blacks and Whites, Hispanic and Whites, American Indians and Whites, and Asian Americans and Whites. Table 2 summarizes our findings.

Minimal differences exist between ethnic/cultural groups at the Big Five level. The largest mean difference is between Blacks and Whites on Openness to Experience, where Blacks score about $d = .2$ lower than Whites. Sample sizes for American Indians and Asian Americans are small, so findings of minimal mean differences here are suggestive but far from conclusive.

The facet level of Conscientiousness and Extraversion reveals consistent differences by ethnicity. On Achievement and Dependability, two facets of Conscientiousness, Hispanics and Asian Americans tend to score slightly higher than Whites on Achievement but lower than Whites on Dependability. Blacks and American Indians have means similar to Whites on Achievement, with greater differences between their scores and Whites on Dependability. Similarly, on Affiliation and Surgency, two facets of Extraversion, Blacks score higher than Whites on Surgency but lower than Whites on Affiliation.

Differences between ethnic/cultural groups on the compound variables are small to nonexistent. Differences between different ethnic/cultural groups on Integrity scales range from -.04 to .14, Hispanics scoring an average of .14 standard deviations higher than Whites on Integrity scales. On Managerial Potential scales, Blacks score an average of .30 standard deviations lower than Whites, although because within-group sample sizes were not reported for this study finding, the stability of that difference is difficult to judge.

The largest mean-score differences between different ethnic/cultural groups and Whites are on Social Desirability scales. Hispanics ($N = 2285$) score about .56 and Asian Americans ($N = 91$) about .40 standard deviations higher than Whites on such scales. The Asian American sample size is, however, too small to provide a stable estimate.

Gender differences. Noteworthy differences exist between men and women on personality variables. For Agreeableness, the mean for women is about .40 standard deviations higher than for men. Differences appear at the facet level of Conscientiousness, where women tend to score higher than men on Dependability but about the same as men on Achievement. Differences also appear at the facet level of Extraversion, where women score higher on average than men on Affiliation but lower than men on Surgency (i.e., potency or dominance). Not surprisingly, women tend to score much lower than men on Rugged Individualism ($d = 1.75$ approximately), which is essentially Masculinity.

Age differences. Minimal differences exist between older and younger working adults. Only at the facet level of Conscientiousness do age differences of any consequence occur: Compared to younger working adults, older working adults have a mean Dependability score almost .50 standard deviations higher and for Achievement about .25 standard deviations lower, although these may be relatively sample-dependent, with only three and two samples, respectively. Table 3 summarizes our findings.

Table 2: Summary of mean-score differences on personality constructs for Ethnic/Cultural Groups

	N		Mean	*N*		Mean	*N*		Mean	*N*		Mean
Personality construct	Black	White	*d*[1]	Hispanic	White	*d*[1]	Amer. Indian	White	*d*[1]	Asian Amer.	White	*d*[1]
EXTRAVERSION[2]	***21,220***	***45,918***	***−.10***	***2,538***	***34,706***	***.01***	***155***	***24,513***	***−.10***	***283***	***11,818***	***−.15***
Affiliation	1,034	5,606	−.31									
Surgency (Potency)	19,735	37,104	.12	2,308	31,498	.01	132	21,364	−.11	105	8,669	−.01
CONSCIENTIOUSNESS[2]			***−.06***			***−.04***			***−.14***			***−.08***
Achievement	18,701	31,498	−.01	2,308	31,498	.04	132	21,364	−.09	105	8,669	.13
Dependability	35,873	86,504	−.11	8,097	80,898	−.11	633	67,916	−.19	279	11,503	−.29
ADJUSTMENT	***20,123***	***39,938***	***.04***	***1,515***	***34,335***	***.01***	***154***	***24,198***	***.03***	***279***	***11,503***	***−.08***
AGREEABLENESS	***20,488***	***42,786***	***−.02***	***2,764***	***37,180***	***−.06***	***154***	***24,198***	***−.13***	***279***	***11,503***	***−.01***
OPENNESS to Experience	***1,422***	***8,440***	***−.21***	***207***	***2,834***	***−.10***	***22***	***2,834***	***.00***	***174***	***2,834***	***−.18***
Integrity Scales[3]	108,871	481,523	−.04	59,790	481,523	.14	2,601	481,523	.08	13,594	481,523	.04
Managerial Potential Scales[4]	Total =	25,698	−.30				Total =	10,249	−.07			
Social Desirability Scales	18,638	31,124	−.05	2,285	10,265	.56	131	21,049	.03	91	8,354	.40

Notes: [1] Negative value means the group listed first scores lower.

[2] Mean *d* for the Big 5 construct was computed by equally weighting *d* values for the two facets to obtain a *d* for the Big 5 construct, then sample-size weighting that value and the *d* obtained on independent sample(s) (if any) that completed a global measure of the Big 5 construct.

[3] Integrity scales consist primarily of three Big Five factors: Conscientiousness, Agreeableness, and Adjustment (Ones and Viswesvaran 1998b).

[4] Managerial Potential Scales are very heterogeneous (Viswesvaran, Ones, and Hough 1998).

Mean *d* based on the following studies: Dohm and Hough (1993); Feingold (1994); Goldberg, Sweeney, Merenda, and Hughes (1998); Hough (1998); Hough, Carter, Dohm, Nelson, and Dunnette (1993); Houston and Schneider (1996); Ones and Viswesvaran (1998a); Ones, Hough, and Viswesvaran (1998); Smith and Reise (1998).

Table 3: Summary of mean-score differences on personality constructs for gender and age groups

	Gender[1]				Age[2,3]		
	N		Mean		*N*		Mean
Personality Construct	Women	Men	*d*[4]	*k*	≥40	< 40	*d*[4]
EXTRAVERSION	***177,103***	***181,382***	***−.09***[5]	***3***	***Total = 5,266***		***.01***
Affiliation	42,374	46,999	.12	0			
Surgency (Potency)	81,313	80,813	−.24	2	564	736	−.13
CONSCIENTIOUSNESS			***.08***[5]	***3***			***.12***
Achievement	28,231	23,150	−.06	2	564	736	−.24
Dependability	83,013	82,963	.22	3	Total = 4,929		.49
ADJUSTMENT	***83,636***	***82,608***	***-.24***	***3***	***Total = 4,929***		***.12***
AGREEABLENESS	***75,858***	***75,248***	***.39***	***3***	***Total = 4,929***		***.21***
OPENNESS to Experience	***23,527***	***28,023***	***-.07***	***1***	***Total = 3,629***		***-.02***
Rugged Individualism	1,498	1,481	−1.74	0			
Integrity Scales[6]	316,359	364,316	.16	2	9,743	68,477	.08
Managerial Potential Scales[7]	Total = 69,255		−.03	11	Total = 24,994		.07
Social Desirability Scales	28,129	22,785	.00	1	418	545	.09

Notes:

[1] Mean *d* based on the following studies: Dohm and Hough (1993); Feingold (1994); Goldberg, Sweeney, Merenda, and Hughes (1998); Hough (1998); Hough, Carter, Dohm, Nelson, and Dunnette (1993); Houston and Schneider (1996); Ones and Viswesvaran (1998a); Ones, Hough, and Viswesvaran (1998); Smith and Reise (1998).

[2] Mean *d* based on the following studies: Hough, Carter, Dohm, Nelson, and Dunnette (1993); Houston and Schneider (1996); Ones and Viswesvaran (1998b); Ones, Hough, and Viswesvaran (1998).

[3] Some studies reported correlations between age and the variable of interest. These correlations were converted to *d*.

[4] Negative *d* values means the group listed first scores lower.

[5] Mean *d* for the Big 5 construct was computed by equally weighting *d* values for the two facets to obtain a *d* for the Big 5 construct, then sample-size weighting that value and the *d* obtained on independent sample(s) that completed a global measure of the Big 5 construct.

[6] Integrity scales consist primarily of three Big Five factors: Conscientiousness, Agreeableness, and Adjustment (Ones and Viswesvaran 1998b).

[7] Managerial Potential scales are very heterogeneous (Viswesvaran, Ones and Hough 1998).

Lessons Learned

At the Big Five level, few subgroup differences exist between ethnic/cultural groups, the only stable finding of any consequence is the difference between Blacks and Whites on Openness to Experience, and it is likely that this difference is a function of facet-level differences. Specifically, intellectance is likely to show larger differences favoring Whites than are facets such as traditional values or need for experience or variety. Gender differences exist at both the Big Five level and facet level: women score higher on Agreeableness and lower on Adjustment than men and about the same on Conscientious and Extraversion, although at the facet level of these two factors, differences exist which are in the opposite direction. Findings from these studies and meta-analysis are based mostly on US data. In a large UK sample, Ones and Anderson (unpublished manuscript) reported mean scores across personality scales on three personality inventories in popular use. Findings were consistent with the US data. Gender differences were generally small to nonexistent. Ethnic subgroup differences were based on small sample sizes (mean $N < 50$), but results suggested small mean differences here as well.

As with cognitive abilities, one of the important conclusions from this research is that the specific construct (e.g., subscales of the Five Factor Model) moderates the magnitude of the effect size of mean differences between groups, often by amounts that are practically meaningful for personnel selection. Employers considering measures of personality in employment settings need to assess accurately the appropriate job requirements and their performance determinants. Ignoring job analysis information is clearly not appropriate.

Physical Abilities

Employment testing for physically demanding work typically includes tests that are based on either task sampling (work samples or job simulations) or tests measuring physical ability constructs such as muscular endurance. Tests involving work samples are often defended on the basis of content validity. Tests measuring physical ability constructs are often defended on the basis of criterion-related validity. Our summary findings below rely heavily on J. Hogan (1991a, 1991b) and on Sackett and Wilk (1994).

J. Hogan (1991b) proposed a hierarchy of physical ability constructs sufficient for describing physical requirements in work settings. Three broad physical abilities are at the top level: muscular strength, cardiovascular endurance, and movement quality. More specific abilities reside within two of the three broad physical abilities. We use J. Hogan's structure of physical abilities to summarize group mean-score differences.

Sackett and Wilk (1994) computed effect sizes of the differences between men and women for the studies J. Hogan (1991a) cited, as well as effect sizes between men and women in the Cooper Institute for Aerobic Research (n.d.) data set. The Cooper data set includes test scores for more than 30000 adults tested in the 1980s.

Sackett and Wilk also examined mean-score differences between men and women for different age groups in norms developed by the Cooper Institute for Aerobic Research (n.d.). The authors concluded that mean-score differences between men and women did not vary across age groups; they therefore combined data across age groups. This does not mean that age differences do not exist; nonetheless, it is the male–female mean-score differences that has been of greater interest to industrial/organizational psychologists because of the serious adverse impact against women when physical ability tests are used for personnel decision-making in the workplace.

Muscular Strength

Muscular strength is the ability 'to apply or resist force through muscular contraction' (J. Hogan 1991b, pp. 503). In J. Hogan's (1991b) taxonomic structure, muscular strength consists of three specific physical ability constructs: muscular tension, muscular power and muscular endurance.

Muscular tension. Muscular tension includes both isometric (static strength) and isotonic (dynamic strength) muscular contraction. According to the Sackett and Wilk (1994) computations of d between men and women for the J. Hogan (1991a) studies, $d = 2.28$, and for the Cooper (n.d.) data, $d = 1.79$, men scoring higher on average.

Muscular power. Muscular power, the second specific ability in Hogan's (1991b) muscular strength construct is defined as the rapid use of dynamic force. It is muscular tension plus speed. According to Sackett and Wilk (1994), the J. Hogan (1991a) studies report $d = 2.10$; men score on average more than two standard deviations higher than women on tests of muscular power.

Muscular endurance. The third specific ability in Hogan's (1991b) muscular strength construct is muscular endurance, defined as the 'capacity of muscle groups to sustain contraction over time while performing tasks' (p. 503). Sackett and Wilk (1994) found $d = 1.52$ in Hogan's (1991a) studies cited and $d = .99$ in the Cooper (n.d.) data, both effect sizes favoring men.

Cardiovascular Endurance

Cardiovascular endurance is the 'capacity to sustain gross (as contrasted with localized) muscular activity over prolonged periods' (Hogan 1991b, p. 504). It involves the large muscles of the body and requires both aerobic capacity and general systemic fitness. With strenuous muscular activity over time, muscles require oxygen and removal of the byproducts of metabolism; hence the need for cardiovascular endurance. According to Sackett and Wilk (1994), mean $d = 1.27$ between men and women in Hogan's (1991a) studies cited and $d = 1.06$ in the Cooper (n.d.) data, with the mean differences favoring men.

Movement Quality

This physical ability construct consists of three more specific abilities: flexibility, balance and neuromuscular integration. All three specific constructs have in common characteristics that contribute to skilled performance.

Flexibility. Flexibility refers to the range of motion through which joints allow limbs to rotate easily regardless of body position, be it an awkward, contorted, or extended position (Hogan 1991b). According to Sackett and Wilk (1994), $d = .05$ in Hogan's (1991a) studies cited, but $d = .69$ for flexibility in the Cooper (n.d.) data, with women scoring higher than men – the only physical ability in which women score higher than men.

Balance. Hogan (1991b) defines balance as the 'capacity to remain stable while the body's base of support is reduced or changed' (p. 504). Sackett and Wilk (1994) found $d = .53$ in the Hogan (1991a) studies, men scoring higher than women.

Neuromuscular integration. Neuromuscular integration, more commonly known as coordination, is the 'capacity to organize movements in sequence within acceptable temporal and spatial constraints in response to either internal or external stimuli' (Hogan 1991b, p. 504). According to Sackett and Wilk (1994), $d = .70$ in the studies Hogan (1991a) cited.

Lessons Learned

Table 4 summarizes data on mean-score differences between men and women on physical abilities. On average, men score much higher than women do on muscular strength ($d = 1.66$), higher on cardiovascular endurance (i.e., aerobic power, $d = 1.09$), and less high on movement quality, which consists of flexibility, balance and coordination ($d = .20$). Test users in employment settings need to assess accurately the physical requirements of work to determine the specific physical abilities that are important. For example, an appropriate physical ability measure for some jobs may not be balance but flexibility, two physical abilities that show gender differences in the *opposite* direction. If selection occurs solely on certain physical ability measures such as muscular power, then adverse impact against women is likely to occur.

Table 4: Summary of standardized mean-score differences on physical ability constructs for men and women

Construct[1]	k	Approximate Total N[2]	Approximate mean d[3,4]
Muscular Strength			−1.66[5]
Muscular tension	31	34,949	−1.86
Muscular power	6	2,252	−2.10
Muscular endurance	15	31,857	−1.02
Cardiovascular Endurance (Aerobic Power)	13	34,090	−1.09
Movement Quality			−.20[5]
Flexibility	11	31,985	.64
Balance	6	887	−.53
Neuromuscular integration (coordination)	4	1,246	−.70

Notes: [1] J. Hogan's (1991b) physical abilities taxonomic structure for physical activity in work settings.
[2] The Cooper (n.d.) data set includes over 30,000 adult men and women. We used 30,000 as the sample size for that data set. The remaining people are from the studies J. Hogan (1991a) cited.
[3] Negative d values mean women score lower than men.
[4] mean d calculated by sample-size weighting d values using Sackett and Wilk's (1994) calculations of d for Hogan's (1991a) cited studies and Cooper's (n.d.) data.
[5] d calculated by equally weighting specific ability d values within the construct.

Possible Causes and Correlates of Subgroup Differences

Several possible causes of subgroup differences have been hypothesized, and many strategies for reducing them have been attempted. This section summarizes causes and strategies in light of the following four broad categories: method of measurement, culture, test coaching, test-taker perceptions, others' perceptions (e.g., stereotype threat), and criterion issues.

Method of Measurement

This section summarizes empirical work and hypotheses addressing subgroup mean-score differences due primarily to measurement method (i.e., assessment centers, biodata inventories, interviews and job/work samples) in the context of the constructs they tend to assess. Much of our summary in this section deals with our own and other researchers' hypotheses more than empirical work, because most researchers researching measurement methods do not provide the descriptive statistics to compute d values between ethnic/cultural/gender/age groups.

Assessment Centers

The assessment center (AC) is a process of evaluating and combining information gathered from multiple assessment tools and methods to form ratings on multiple constructs. Almost always, ACs include at least one individual or group simulation exercise. Some of the different assessment tools that are often included in ACs are personality inventories, cognitive ability tests, role-playing exercises (type of job sample) and in-baskets (another type of job sample). Biodata inventories and interviews are also often included in ACs; we deal with these two assessment methods in a separate section.

The overall AC rating thus reflects performance on a variety of assessment methods as well as a variety of constructs. Assessment tools such as the role-playing exercises and in-basket often measure interpersonal competence, which research by R. Schneider, Ackerman and Kanfer (1996) indicates is composed of mostly personality constructs. Some of these, and others as well, also measure cognitive ability. We thus deal with mean score differences separately for the overall AC rating and different assessment tools such as role-playing exercises, which are more construct-specific.

Ethnic/cultural differences. Hoffman and Thornton (1997) summarized mean score differences between Blacks and Whites on overall assessment ratings concluding that the results are fairly evenly split between studies showing no significant differences in average performance and those showing Blacks scoring lower. Studies that do find differences usually find the difference is less than one standard deviation, the usual Black–White difference in general intelligence tests. They also noted that those comparisons showing Blacks scoring lower than Whites were probably not based on samples of Blacks and Whites that were comparable. The samples in these comparisons were probably not equally representative of their respective populations: broader samples of Blacks were selected for AC attendance in an effort to identify a larger pool of qualified Black candidates. Another possible explanation for overall AC ratings that show no differences between Blacks and Whites is that these ACs primarily measure interpersonal skills. Bobrow and Leonards (1997), for example, developed an AC for a first-line supervisory job in a customer contact/service provider division, a job with significant interpersonal skill requirements, which yielded essentially no differences between Whites and minorities.

Goldstein, Riley and Yusko (1999) examined Black–White mean score differences at the exercise level and found that group mean differences (d ranged from .03 to .40 with Blacks scoring lower) varied by type of exercise. Group differences appeared to be a function of the cognitive component of the exercise. They hypothesized and found that exercise components requiring interpersonal skills resulted in less adverse impact against Blacks ($d = .28$) than exercise components requiring cognitive abilities ($d = .60$).

Age differences. Clapham and Fulford (1997) examined age differences in overall AC ratings and found that in a high-level executive AC, partial correlations (controlled for level, years of service, verbal and math ability and gender) between construct ratings and age ranged from .00 to −.39, median $r = -.23$ (or $d = -.46$). Results were similar for a middle management AC. Partial correlations between construct ratings and age ranged from −.06 to −.32, median $r = -.21$ (or $d = -.42$).

Summary. The Black–White difference in mean overall AC ratings is smaller than is typically found for cognitive ability tests. When mean score differences are examined according to the type of AC exercise, the pattern of Black–White differences appears to mirror the pattern on cognitive and personality constructs – essentially no difference on interpersonal skill (i.e., personality characteristics) and significant differences on cognitive abilities. Older working adults are rated significantly lower than younger adults on overall AC performance – a surprising result if experience, which presumably comes with age, is important for developing managerial skills.

Biodata Inventories

Biographical (biodata) inventories ask respondents to report their behavior or experience in situations that occurred earlier in their lives. It is a method of gathering life-history information that provides multiple 'samples' of real-life behavior, enabling the investigator to infer, if desired, the respondent's relative standing on underlying constructs. Biodata questions can be developed to measure such constructs as cognitive abilities, interests, physical abilities and personality characteristics. Researchers have often developed biodata inventories without regard to construct measured. In these circumstances, the researchers have typically relied on content or criterion-related validity arguments to support their use in personnel decision-making.

Multiple-choice/true-false inventories. This is the typical biodata inventory. It consists of a list of life-history questions that is presented to the respondent who answers yes or no to each question or chooses from a set of options. Although aware of the confounding effects of summarizing across different constructs that are measured in biodata inventories, N. Schmitt, Rogers, Chan, Sheppard and Jennings (1997) summarized mean score differences between Whites and minorities, concluding minorities score lower than whites ($d = .20$). Gandy, Dye and MacLane (1994) describe the development and analysis of the federal government's 'Individual Achievement Record', a carefully developed biodata inventory. Their analysis of the factor structure of the inventory indicates it measures high school achievement, college achievement, work competency and leadership skills. They compared mean score differences on ethnic/cultural, gender and age groups on total, or overall biodata scores rather than on the separate factors. Nonetheless, the comparisons can be useful. Blacks scored lower than Whites ($d = .27$); Hispanics scored essentially the same as Whites ($d = .08$); men scored slightly lower than women ($d = .15$), and adults over 32 years of age scored the same as younger working adults ($d = .05$). These small differences are noteworthy given that two of the four factors focused on achievement in high school and college, factors that tap cognitive abilities as well as noncognitive variables.

Grades. Roth and Bobko (2000) examined the validity of college grade point average (GPA) as a predictor of job performance and mean score differences between Blacks and Whites. They found $d = .78$ on cumulative GPA in the senior year of college, with Blacks on average obtaining lower GPA than Whites. Recruiters consider GPA to reflect verbal ability, math ability, as well as motivation because GPA is an index of performance over time (Brown and Campion 1994). Given the d values for each of these constructs separately, a measure that combines all three is likely to produce an effect size of three-quarters of a standard deviation, just as Roth and Bobko found.

Accomplishment record. The accomplishment record is a biodata form that asks respondents to provide descriptions of accomplishment in highly relevant, behavioral job dimensions (Hough 1984). Raters then evaluate the information using rating guidelines for each accomplishment record dimension as well as anchored-rating scales that provide both examples of high-, medium-, and low-rated accomplishments for each accomplishment record dimension. The accomplishment record focuses directly on past job or work performance constructs, which of course relate to various individual-differences constructs (e.g., declarative and procedural knowledge, and motivation). Hough (1984) and Hough, Keyes and Dunnette (1984) compared ethnic/cultural and gender mean score differences on the overall accomplishment record scores for professional jobs. Sample-sized weighted $d = .24$ for the White-Minority comparison, with minorities scoring lower on average. There were essentially no mean differences between men and women, $d = .09$, women scoring slightly lower.

Interview

Huffcutt and Roth (1998) meta-analyzed mean score differences between ethnic/cultural groups. They found that Black and Hispanic job applicants scored on average only about one-quarter standard deviation below Whites. In addition, structured interviews produced smaller mean score differences than did unstructured interviews, a finding in line with Huffcutt, Roth and McDaniel (1996) who concluded that structured interviews focused less on ability. N. Schmitt *et al.* (1996) summarized six Black–White comparisons of interviews that measured mostly motivational factors; $d = .15$, with Blacks scoring lower; Hispanic–White comparisons were similar, $d = .19$, with Hispanics scoring lower.

Work/Job Samples

Many types of work/job samples exist – simulations, role-play exercises, in-baskets and portfolio assessments to name a few. Some authors (e.g., N. Schmitt *et al.* 1996) also include situational judgment in this category although Motowidlo, Dunnette and Carter (1990) refer to it as a 'low fidelity simulation'. Comparisons between groups that are based on summaries across all these different types of assessments confound not only construct measured with type of assessment method, but also stimulus mode (such as video versus paper-and-pencil) and response mode (such as hands-on-performance versus written). When the same construct is measured, however, it may be useful to compare types of assessment methods. When this was the case, we summarized d values.

Ethnic/cultural differences. Schmidt, Greenthal, Hunter, Berner and Seaton (1977), for example, compared mean score differences between Whites and minorities on a content-valid job sample test of metal trades skills and a well-constructed content-valid written achievement test of the same technical skills. They found $d = .81$ (comparing true scores) on the job sample whereas $d = 1.44$ (comparing true scores) for the written trades test, minorities scoring lower on both on average. They concluded that relative to written achievement tests, job samples of the same skill produced much less difference (significantly and practically) between minorities and Whites.

Pulakos, Schmitt and Chan (1996) compared Black–White mean score differences on three types of work samples with a written test of cognitive ability. They found that Blacks on average scored lower than Whites on all the measures. However, the work samples resulted in much smaller mean score differences. On the written test of cognitive ability $d = 1.25$ whereas on the video work sample $d = .83$, on the role-play work sample $d = .58$, and on the situational judgment work sample $d = .35$. Chan and Schmitt (1997) compared video-based versus paper-and-pencil situational judgment tests and found that video-based tests produced smaller mean score differences that paper-and-pencil situational judgment tests, attributing the reduced mean score differences to less need for reading comprehension in the video-based assessment. Similarly, N. Schmitt *et al.* (1996) meta-analyzed Black–White mean score differences on job sample tests. They found that Blacks lower scored on average than Whites, $d = .38$ (37 effect sizes summarized). In short, work samples have tended to produce smaller mean score differences between Blacks and Whites when compared with cognitive ability measures. The work sample often measures more than individual differences on cognitive ability, which may contribute to reduced mean differences. In many cases, work samples may be more legally defensible and job relevant in cases where the job applicant has prior work experience or when the job is not very complex and the work sample can be easily learned.

Several studies have focused on situational judgment tests paper-and-pencil and video based. We sample-size weighted the d values for the Black–White comparisons. Results are shown in Table 5. As shown there, Blacks score lower than Whites, $d = .61$ on paper-and-pencil situational judgment tests and $d = .43$ on video-based situational judgment tests. Although the sample sizes are approximately 500 for Blacks, the data do suggest that the mean score difference between Blacks and Whites on video-based situational judgment tests are smaller than on paper-and-pencil situational judgment tests.

A meta-analysis summarizing 20 Hispanic–White d-values on job sample tests found that Hispanics scored on average the same as Whites (N. Schmitt *et al.* 1996). Pulakos *et al.* (1996) compared Hispanic–White mean score differences on three types of work samples with a written test of cognitive ability. They found that Hispanics on average scored lower than Whites on all the measures. However, the work samples resulted in much smaller mean score differences. On the written test of cognitive ability $d = .92$ whereas on the video work sample $d = .59$, on the role-play work sample $d = .37$, and on the situational judgment work sample $d = .05$.

A few studies examined mean score differences between Hispanics and Whites on paper-and-pencil versus video-based situational

Table 5: Summary of mean score differences on situational judgment inventories according to stimulus mode

Stimulus mode	k	N	N	d
		Ethnic/Cultural comparisons		
		Black	White	
Paper-and-Pencil	7	493	3,064	−.61[1]
Video	4	468	1,489	−.43[4]
		Hispanic	White	
Paper-and-Pencil	3	340	2,672	−.26[2]
Video	1	44	930	−.39[5]
		Gender comparison		
		Female	Male	
Paper-and-Pencil	7	3,120	2,089	.26[3]
Video	2	109	310	.19[6]

Notes: A negative value means the group listed first scores lower.

[1] Mean d based on the following studies: Chan and Schmitt (1997); Motowidlo, Dunnette, and Carter (1990); Motowidlo and Tippins (1993); Pulakos and Schmitt (1996); Weekley and Jones (1999).

[2] Mean d based on the following studies: Weekley and Jones (1999); Pulakos and Schmitt (1996).

[3] Mean d based on the following studies: Chan and Schmitt, (1997); Lievens, Coetsier, and Decaesteker (in press); Motowidlo, Dunnette, and Carter (1990); Pulakos, Schmitt, and Chan (1996); Weekley and Jones (1999).

[4] Mean d based on the following studies: Chan and Schmitt (1997); Smiderle, Perry, and Cronshaw (1994); Weekley and Jones (1997).

[5] Mean d based on the following study: Weekley and Jones (1997).

[6] Mean d based on the following studies: Chan and Schmitt (1997); Smiderle, Perry, and Cronshaw (1994).

judgment tests. We sample-size weighted the *d* values for the Hispanic–White comparisons. Results are shown in Table 5. As shown there, Hispanics score lower than Whites, $d = .26$ on paper-and-pencil situational judgment tests and $d = .39$ on video-based situational judgment tests. The sample size for Hispanics on the video-based situational judgment test is, however, very small ($N = 44$).

Gender differences. N. Schmitt *et al*. (1996) meta-analyzed male and female mean score differences on work/job samples. On average women scored higher than men did ($d = .38$). Pulakos *et al*. (1996) found women scored higher than men on all three types of work samples they examined; $d = .42$ on the video work sample, $d = .22$ on the situational-judgment work sample, and $d = .13$ on the role-play work sample. Lievens, Coetsier and Decaesteker (in press) compared male–female mean score differences on situational tests, finding women scored on average higher than men on the situational tests ($d = .18$).

Several studies focused on paper-and-pencil and video-based situational judgment tests. We sample-size weighted the *d* values for the female-male comparisons. Results are shown in Table 5. As shown there, women score higher than men, $d = .26$ on paper-and-pencil situational judgment tests and $d = .19$ on video-based situational judgment tests.

Summary. Work samples appear to produce smaller mean score differences between minorities and Whites. The smaller difference is probably due, at least in part, to work samples measuring a combination of cognitive and motivational (non-cognitive) variables. Mode of stimulus presentation (paper-and-pencil versus video) appears to moderate mean score differences between groups, with video versions producing less adverse impact than paper-and-pencil versions.

Culture

The influence of culture on item responses has long been proposed as a reason for why certain demographic groups (e.g., White and Black) perform differently on standardized tests. *Culture* is not, however, a unitary concept (Kroeber and Kluckhohn 1952; Lytle, Brett, Barsness, Tinsley and Janssens 1995). Instead, culture consists of several dimensions (e.g., individualism/collectivism, masculinity/femininity; Hofstede 1980; Nyfield and Baron, 2000) that may differ by country, ethnicity/race, religion, gender, or any combination of these. The nature of culture and cultural differences thus needs to be examined carefully and specifically. Most theories and research tend to focus on White-Black differences; thus we do too.

Effect on Cognitive Ability

The impact of culture on cognitive ability scores has long been an area of question and debate (e.g., Bond 1993; Boykin 1983, 1994; Garth 1925; Helms 1992; Jensen 1969, 1998; Neisser *et al*. 1996; Sackett and Wilk 1994; Scarr 1981; Schmidt 1988; Valencia and Lopez 1992). However, identifying cultural factors that influence test responses has been particularly difficult (e.g., Bond 1993) and empirical data from most research on this topic show that cultural differences do not seem to account for racial differences on cognitive ability tests (see Jensen 1998, for an extensive review). For the General Aptitude Test Battery (GATB) on a large Dutch sample, te Nijenhuis and van der Flier (1997) find high factor-structure similarity between first-generation immigrants (Surinamese, Antillians, North Africans and Turks) and majority group members. Model fit to both Dutch native and Dutch immigrant data was high, although mean differences on subtests were found, especially for Vocabulary, Arithmetic Reasoning and Three-Dimensional Space.

The effect of culture on test scores and construct equivalence is often assessed using statistical indices of differential item functioning (DIF). DIF occurs when the probability of choosing a particular response on an item differs as a function of non-construct influences (Camilli and Shepard 1994). If any non-construct influences are related to culture, then cultural differences may be considered a partial cause of test score differences. For example, if Black and White respondents of equal ability choose different options for a particular item, that item shows DIF and may have different meanings for Whites and Blacks.

O'Neill and McPeek (1993) review research on standardized academic aptitude and achievement tests developed in the USA – the GRE, GMAT, ACT and LSAT – showing that DIF is more likely to be found when the item content is stereotypic of one group more than another. For example, an item that is couched in a sports context may be more difficult for women to answer than men. Across all the datasets reviewed, however, the effect is rather weak when it exists. Scheuneman (1987; Scheuneman and Gerritz 1990) shows DIF increases with the use of superlatives (e.g., most, best), with the number of words in the item, and with items that ask one to identify a single incorrect response. Despite these general research findings, studies have yet to find consistent cultural item characteristics that systematically and strongly influence item responses. Freedle and Kostin (1997) relate Black–White DIF to substantive item characteristics, but the amount of DIF is not reported, and substantive interpretations may be confounded by the nature of the DIF statistic that was applied to the data. Discouraging DIF findings have led many to conclude that this line of research is not likely to result in an understanding and successful attempts at reduction of subgroup test-score differences on cognitive ability tests. Typically when DIF for Blacks vs. Whites is found, it is usually (a) negligible in size; (b) found in either direction – favoring one group or another; (c) not necessarily found for items related to a cultural hypothesis or theory; and (d) not necessarily in the right direction given an item that seems culturally relevant (e.g., Guion 1998; Raju, van der Linden and Fleer 1995; Waller, Thompson and Wenk, 2000).

Although item-level DIF and consideration of item content for cultural bias are very important when developing new measures, explicit consideration of culture in assessment, either by formally incorporating or excluding dimensions of culture, has not reduced mean test-score differences. For example, DeShon, Smith, Chan and Schmitt (1998) presented cognitive ability test items in a social context (presumably more reflective of Black culture; see Boykin 1983) and found that the White-Black mean difference did not decrease, but actually increased slightly. Jensen and McGurk (1987) found that attempts to write items reflective of White and Black culture did not change the mean test score difference. Other attempts to develop 'culture-free' assessments, such as reaction time measures or nonverbal measures (e.g., Raven's Progressive Matrices), indicate that although Black–White differences are somewhat smaller, they still persist (Vernon and Jensen 1984; see Jensen 1998, for an extensive review). Table 1 illustrates these differences quite clearly. Our meta-analyses of mean score differences at the construct level do suggest smaller differences in measures of fluid intelligence (supposedly more culture free) compared with measures of crystallized intelligence.

Empirical results generally do not support the hypothesis that cognitive ability tests are culturally biased, although one might speculate that this occurs because most examinations of DIF are post-hoc (e.g., A. Schmitt, Holland and Dorans 1993). However, attempts to examine DIF in an *a priori* hypothesis-testing framework also have not been supportive (e.g., DeShon *et al.* 1998).

Helms (1992) asserts that research examining race and cognitive ability does not allow conclusions about equivalence because culture is inadequately considered. She suggests that studies using race as a proxy for culture cannot warrant claims about cultural equivalence because no substantive theory of culture is being tested (i.e., race is an imperfect indicator of culture). Because White and Black cultures stress different values and actions (e.g., individualism vs. group harmony), and most cognitive ability tests are developed in White culture, Helms suggests that existing studies of cultural equivalence actually assess Black acculturation rather than Black intelligence. Therefore, she calls for research to incorporate more formally White and Black cultural components into ability testing (see Boykin 1983; 1994, for descriptions of these components). Others in the testing community have made similar suggestions (e.g., Bond 1993; A. Schmitt *et al.* 1993).

The DeShon *et al.* (1998) and Jensen and McGurk (1987) studies adapted Black culture into traditional cognitive ability assessment and found essentially the same Black–White mean differences as in past research. The Helms hypothesis would predict that mean differences should have been less, because their college student samples are presumably more acculturated into White culture and thus more similar to Whites than the population as a whole. However, these study findings are qualified in the sense that the studies only considered one aspect of Black culture (e.g., social components) and did not incorporate other aspects.

Berry (1996) suggests a variety of potential cultural influences on test takers that influence the test taker's interpretation of the test. These variables are rarely considered in psychometric examinations of ability tests. Similarly, Greenfield (1997) argues that ability tests developed in one culture and transported to other cultures require cultural agreement in terms of (a) the values and meaning of the items and responses; (b) the focus on individual-level knowledge (not culturally-shared knowledge); and (c) the familiarity of item context and content (e.g., impersonal, one correct response). Similar to Helms, Berry (1996) and Greenfield (1997) argue that ability tests are not universal instruments that can be applied to different cultures without first identifying whether the test has the same cultural meaning within each culture.

In conclusion, empirical research shows that culture has at most a weak influence on cognitive ability test performance, and it does not seem to change the psychometric properties of cognitive ability tests. However, conclusions are weakened by the *post-hoc* nature of this research (e.g., Bond 1993) and almost exclusive use of race (usually Blacks and Whites) as an indicator of culture, rather than measuring level of acculturation directly (Helms 1992). To show more definitively that culture does or does not affect subgroup differences on ability test performance, research needs to adopt a more theoretical hypothesis-testing approach, whereby hypotheses about cultural dimensions relevant to test performance are specified and measured (see Greenfield 1997). Expanded conceptualizations and theories of culture (i.e., models that consider and incorporate culture-relevant individual differences, social structures and situations) will be necessary because broad racial group membership is likely an imperfect and simplistic measure of culture (Helms 1992). However, many have noted that the complexity of culture does not make this a straightforward task (e.g., Bond 1993; A. Schmitt *et al.* 1983), and recent empirical evidence supports this concern (e.g., DeShon *et al.* 1998).

Effect on Personality and Related Constructs

The latent structure of personality appears to be very similar across cultures (cf. McCrae and Costa 1997), but culture may still affect the expression of personality, or in other words, how personality traits are behaviorally manifested (Church and Katigbak 1988). For example, basic personality research suggests personality expression is in part determined by values (e.g., McCrae and Costa 1996). Given that values differ substantially across cultures (e.g., Hofstede 1980; Schwartz 1994), it is likely that the expression of personality traits differs across cultures. Markus and Kitayama (1991) provide support for this; they conducted several studies demonstrating that cultural dimensions and values influence the way the self is construed (e.g., self-efficacy, self-esteem). Specifically, people raised in collectivistic cultures tend to construe the self relative to social relations and contexts, whereas people raised in individualistic cultures tend to construe the self as being unique and independent. In support of this view, Ghorpade, Hattrup and Lackritz (1999) found that measurement equivalence was not established when comparing females' self-esteem and higher-order need strength from the United States and India (although locus of control was equivalent).

Schmit, Kihm and Robie (2000) used experts from numerous countries and cultures to develop a personality inventory consisting of items that were interpreted the same linguistically and meant the same behaviorally across many international cultures. By using within-culture experts to help with item development, they increased the chances that the instrument was sensitive to cultural differences yet produced measurement equivalence across cultures. Their results support the notion that personality constructs may be upheld cross-culturally even though the particular expression of personality traits may be culturally specific.

Some research has used differential item functioning (DIF) to examine the impact of culture on biographical data (biodata). In a multiracial sample of US government employees, N. Schmitt and Pulakos (1998) conducted log-linear analyses of item responses and revealed some tentative but substantively interpretable patterns for the differential prediction found when predicting overall job performance ratings. For example, Whites tended to report that they were more likely to initiate projects than were African Americans or Hispanics; African Americans appeared to be more interpersonally skilled and planned more in their work. Both African American and Hispanic groups, compared with Whites, reported more effort in maintaining their physical and professional appearance on the job as opposed to 'being themselves'. Whitney and Schmitt (1997) followed this study up by using items intended to reflect White or Black culture, testing whether Black–White cultural differences in values explained some of the DIF in biodata items. Over one quarter of the items exhibited DIF, but the culture items explained only a modest amount of the DIF at best. The only clear cultural factor that influenced responses was one's views about basic human nature, with Blacks more likely to express negative views.

Summary. Research linking culture to test scores and subgroup differences has long been an area of both hope and frustration. One major problem with past research has been the reliance on *post-hoc* examinations of DIF (Bond 1993). A second major problem has been an over-reliance on race as a proxy for culture (Helms 1992). Recent research has attempted to address both of these limitations by using more sophisticated theories of culture in a hypothesis-testing manner (e.g., DeShon *et al.* 1998; Whitney and Schmitt 1997). However, to date, the end result of this research has not changed the previous literature's conclusions: culture, at least given researchers best attempts to define it, does not appear to account for ethnic/cultural differences in cognitive ability items and test scores. In the few instances where culture does seem to be related to test scores and item responses, the effects have been rather small and hard to interpret (e.g., O'Neill and McPeek 1993).

However, this research may not yet have fully incorporated culture into the assessment process. Some evidence of this is offered by research linking culture to personality (e.g., Church and Katigbak 1988; Markus and Kitayama 1991), which tends to show that the basic structure of personality remains the same cross-culturally even if the expression of personality constructs does not. This is supported by Schmit *et al.*

(2000), who demonstrate the importance of developing test items within cultures initially and then refining and retaining those items that show equivalence cross-culturally. Whether or not generating items within-culture and then retaining items that hold up across cultures would benefit other predictor constructs remains uncertain. Such attempts have been recommended for cognitive ability tests (e.g., Greenfield 1997; Helms 1992) even though past research does not indicate this is very helpful (e.g., Jensen and McGurk 1987). More conceptual, theory-driven research linking specific individual-difference constructs to specific aspects of culture may be required to understand the impact of culture on the development and expression of individual differences.

Test Coaching

Test coaching programs (also called test orientation or test preparation programs) are increasingly used in large-scale selection procedures, particularly with public sector assessment. Most often test-coaching programs are designed to familiarize applicants with the content and nature of the testing process and thereby reduce error variance in test scores, as well as to foster more favorable perceptions about the testing process, which may lead to greater motivation and reduced test anxiety (Ryan, Ployhart, Greguras and Schmit 1998). Such programs may seek to improve the accuracy of one's standing on the constructs being measured (i.e., one's true score); however, the short-term nature of employment test-orientation programs does not typically allow for such long-term training effects.

Conclusions about the effectiveness of coaching programs on cognitive ability tests tend to be drawn from the educational literature because few studies on this subject have been conducted in employment settings. In a review of the educational literature, Sackett, Burris and Ryan (1989) show that meta-analyses indicate a small but consistent test score improvement for cognitive-ability coaching programs (e.g., Kulik, Bangert-Drowns and Kulik 1984; Kulik, Kulik and Bangert 1984; Powers 1986). However, test score improvement appears to occur mostly for people already scoring high on the ability, which would unfortunately lead to greater mean score differences between Blacks and Whites.

In contrast to the educational literature, unpublished data from recent employment testing programs indicate that subgroup cognitive ability test score differences are reduced by coaching, although only by $d = .10$ or less (Sackett, Schmitt, Ellingson and Kabin 2001). A review of other studies that examined the effects of coaching on Black–White mean differences on cognitive ability reveals that results are inconsistent. Examples are: Whites improve more than Blacks (Holden 1996); no difference in improvement rates (Maurer *et al.* 1998); no rise in Black test score means (Ryan *et al.* 1998); mixed results (Ryer, Schmidt and Schmitt 1999), or Blacks improve more than Whites (Schmit 1994). Moreover, self-selection effects and lack of an adequate control group make interpreting these results difficult (Sackett *et al.* 1989). Because attendance at coaching programs is often not mandatory, there is no easy way to determine whether the characteristics of those who self-selected into the program differ from those who did not. Powers (1993) has noted that research on coaching programs that do not consider self-selection show effect sizes nearly five times larger than those that consider these effects or include a control group.

Applicant perceptions are related to test performance and subgroup test differences (e.g., Arvey, Strickland, Drauden and Martin 1990); thus coaching programs may indirectly improve test scores by improving test-taking motivation of minority applicants. However, current data suggest small, if any, coaching effects on applicant perceptions beyond pre-existing perceptions (Ployhart, Ryan, Conley and West 1999; Ryan *et al.* 1998). This conclusion may be true primarily for types of tests with which applicants have had considerable exposure and experience (e.g., ability tests) and less true for more novel types of tests (e.g., video-based situational interviews).

Studies examining differential attendance rates of minorities, whites, men and women come to conflicting conclusions. Ryan *et al.* (1998) found that Blacks, women, and those with more test anxiety were more likely to attend a coaching program. In contrast, Schmit (1994) did not find any relationship between attendance and race.

Overall, it appears that test-coaching programs may exhibit a small, positive effect on cognitive ability test scores. Although in educational settings this effect seems to occur most for those individuals who need it the least (i.e., those who already score high on pre-training measures), in employment settings recent data suggest Black–White differences may be reduced, albeit only modestly.

Test-Taker Perceptions

Recent research suggests that applicants' perceptions and reactions toward selection procedures (e.g., face validity) may have consequences for both the applicant and organization (see the review by Ryan and Ployhart, 2000). Applicant test perceptions may differ across subgroups; they may relate to test performance as well as applicant withdrawal. To the extent that subgroup differences in test perceptions are related to test performance and withdrawal, adverse impact could be partially related to subgroup differences in test perceptions.

Applicant Perceptions and Test Performance

In one of the earliest studies on this issue, Lounsbury, Bobrow and Jensen (1989) assessed working adults' attitudes and beliefs about employment testing. They found few consistent differences between racial, gender, and age subgroups on most of their items. Of the few significant subgroup differences observed, minority job applicants reported having more favorable attitudes toward employment tests, whereas older individuals held more negative attitudes. Subsequent research has been devoted toward understanding what types of test perceptions are most related to test scores and how these may differ across subgroups.

Test-taking motivation. Arvey, Strickland, Drauden and Martin (1990) developed a measure of applicant test attitudes (the Test Attitude Survey, or TAS) which included scales for test-taking motivation, belief in tests, need for achievement, and comparative anxiety. The measure indicated that Whites tended to be more motivated to perform well on the tests than Blacks, and Whites were more likely to believe the tests would have an influence on their future relations with the organization. Gender and age differences were also found such that women and older individuals were more anxious about the exams and more likely to believe the tests were difficult. Finally, many of these test attitudes were related to ability and work sample test performance, and when the TAS motivational scores were partialled from the demographic group-test performance correlation, correlations significantly decreased. Thus, this study demonstrated that subgroup differences exist in test attitudes, and test attitudes contribute to subgroup test score differences. Further evidence supporting the impact of test attitudes on cognitive ability test performance was found in a lab study by Schmit and Ryan (1992), where test-taking motivation moderated validity.

Perceptions of validity. A program of laboratory research by Chan and colleagues has linked Black–White differences in test perceptions to test performance. Chan (1997) reported that Blacks and Whites had similar perceptions about the criterion-related validity of personality tests, but Blacks had more negative perceptions about the criterion-related validity of cognitive ability tests than Whites. Further, validity perceptions for the cognitive ability test were positively correlated with test performance. Chan and Schmitt (1997) found that Blacks perceived a video version of a situational judgment tests as more valid than a paper-and-pencil version of the same situational judgment test and the Black–White mean score difference was .5 standard deviations smaller in the video version than in the paper-and-pencil version. Finally, Chan, Schmitt, DeShon, Clause and Delbridge (1997) had groups of White and Black students complete two parallel forms of a cognitive ability test. Between test administrations, participants completed measures of face validity and test-taking motivation. The first test performance influenced the second test performance both directly and indirectly through face validity and test-taking motivation perceptions. Thus, this study suggests that a portion of the Black–White cognitive ability test score difference may be attributable to differences in subgroup test perceptions (most notably, face validity and test-taking motivation) beyond the amount of difference explained by previous test performance.

The body of research by Chan and colleagues is compelling, but this laboratory work should be carried out in applied settings. Recently, N. Schmitt and Mills (in press) conducted an important extension to this research in a sample of service representatives, comparing Black–White differences on both paper and pencil tests and high-fidelity simulations. Essentially the same constructs were assessed by the two sets of measures, but the high-fidelity simulation produced standardized White-Black differences (favoring Whites) nearly half of a standard deviation smaller than the paper and pencil measures. Furthermore, for all practical purposes the criterion-related validity of the measures was nearly identical ($r = .42$ for paper and pencil; $r = .40$ for simulation; both estimates corrected for range restriction). No data were available on test perceptions, however.

The extent to which reducing subgroup test perception differences may reduce adverse impact is not clear. Sackett *et al.* (2001) note that the effect of reducing adverse impact via changes in test perceptions is not likely to be large. Ployhart and Ehrhart (2000) conducted a simulation that examined this issue by varying the amount of Black–White difference in test-taking motivation and the strength of relationship between motivation and cognitive ability test performance to assess how much of a reduction in adverse impact would be present across various selection ratios. Their results found that by itself, reducing subgroup differences in test-taking motivation shows only a small reduction in adverse impact, although under certain conditions (e.g., high relationship between motivation and test performance) the reduction was practically meaningful. Thus, while the impact of test-taking perceptions on adverse impact may be practically important, by itself it does not appear to eliminate adverse impact.

Test Perceptions and Applicant Withdrawal

Most of the research on applicants' perceptions has focused on linking these perceptions to test performance. However, if applicant perceptions are related to withdrawal from the application process, they may indirectly affect adverse impact. Specifically, if different demographic groups have different rates of withdrawal and those with higher withdrawal rates are minorities, adverse impact ratios will be affected (e.g., Morris and Lobsenz, 2000).

In public service jobs, minorities do drop out at a higher rate than Whites (e.g., Ryan and McFarland 1997; Ryan, McFarland, Sacco and Kriska, 2000; Ryan, Ployhart, Greguras and Schmit 1997). Unfortunately very little research has examined whether or not applicant perceptions affect applicant withdrawal. Existing data do suggest applicant perceptions may be related to withdrawal, though only weakly. For example, Schmit and Ryan (1997) found Black police applicants had more negative test-taking perceptions (e.g., less motivation and more anxiety) and withdrew from the selection process at a much greater rate than White applicants. Test perceptions, however, were only weakly related to withdrawal; they did not help explain why Blacks withdrew more than Whites. Similarly, Ryan *et al.* (2000) found minority applicants were more likely to self-select out of the process and perceive the testing less favorably, but these negative perceptions were only weakly related to withdrawal from the process.

Research in this area is far from complete, however, and many fundamental questions are yet to be examined. For example, nearly all of this research has focused on Black–White differences; we know very little about test-perception differences between other ethnic/cultural groups in the USA (e.g., Asian, Hispanic), male–female differences, and international differences.

Stereotypic Perceptions of the Applicant

Social psychologists, primarily, have conducted research in this area which has received recent attention from the press. In particular, Claude Steele's research on stereotype threat (e.g., Steele 1992, 1997) has received a great deal of research attention and coverage in the US popular media (e.g., Begley, 2000). In general, *stereotype threat* refers to situations where one runs the risk of supporting a negative stereotype of a group with which one is identified. As it applies to testing situations, stereotype threat may exist when negative test performance stereotypes exist about one's demographic group. The threat of confirming this negative stereotype may activate a variety of psychological feelings, thoughts, and processes within an individual (e.g., anxiety, disengagement) that may interfere with test performance. The notion of stereotype threat has been forwarded as a partial explanation for lower Black test scores on cognitive ability tests, although some of the supporting statistical analyses by Steele and colleagues should be carefully revisited.

The mere presence of racial information (e.g., simply asking participants to record their race during the experiment) may produce stereotype-threat effects (Steele and Aronson 1995). Similar effects have been found in the area of gender stereotypes (Shih, Pittinsky and Ambady 1999; Spencer, Steel and Quinn 1999), age stereotypes (Levy 1996) and athletic performance (e.g., Stone, Lynch, Sjomeling and Darley 1999). However, it is important to note that the effects of stereotyping need not have a negative outcome. Some studies have found a Pygmalion effect, where test performance is enhanced if the stereotype held about one's group is actually positive, such as with Asian ability test performance (Shih *et al.* 1999) and Black athletic ability (Stone *et al.* 1999). On the other hand, when public expectations of success are salient (Baumeister, Hamilton and Tice 1985), fear of failing to confirm a positive stereotype may cause people to 'choke', thereby undermining performance (Cheryan and Bodenhausen, 2000).

Thus, stereotype threat may have implications for subgroup differences in test performance with a variety of tests and constructs. However, as Sackett *et al.* (2001) discuss, stereotype threat research has not so far been done in real-world testing contexts and results may not generalize to such settings. Although some field research is consistent with laboratory findings (MacLane, Sharpe and Nickels 2000), research testing the effects of stereotype threat needs to be conducted with applicants in actual employment settings.

Differences in Criterion Composition

The vast majority of research on adverse impact has been directed toward understanding and refining tests and *predictors*, but we cannot overstate the fact that understanding *criteria* as a prerequisite is essential, and we offer three reasons why. First, predictors derive their importance from their ability to predict criteria of interest to organizations (Wallace 1965). Job performance and job satisfaction are types of employee-controlled behaviors and attitudes that organizations value, and they (the criteria) should determine the appropriate content of the predictor battery (cf. Binning and Barrett 1989). Second, an unbiased criterion measure is a basic assumption in the most accepted model for assessing predictor bias (i.e., the Cleary model, Cleary 1968), and using predictors without a good understanding of the criteria being predicted can lead to bias. Third, how overall performance is defined and how its components are weighted impact the criterion-related validity of different predictors (e.g., Bobko, Roth and Potosky 1999; Hattrup, Rock and Scalia 1997; Murphy and Shiarella 1997; N. Schmitt, Rogers, Chan, Sheppard and Jennings 1997). Weighting specific performance dimensions leads to a particular set of rational or regression weights for the predictor battery, which in turn affects (increases or decreases) adverse impact effects in the predictor battery.

Recent conceptualizations and models of job performance help us understand why predictors relate to different parts of the performance criterion domain. A model of performance developed by Campbell and colleagues (Campbell 1990; Campbell, Gasser and Oswald 1996; Campbell, McCloy, Oppler and Sager 1993) defines performance across jobs in terms of eight latent factors. These factors are broad because they capture most jobs and most types of performance in the world of work. The eight-factor framework seems to be an informative starting point for studying job performance, but recent research has focused on two dimensions that are even broader: task performance and contextual performance. Customer service and teamwork are also types of performance receiving recent research attention; they may be viewed as additional broad performance constructs or as profiles of Campbell *et al.*'s eight latent performance factors.

Task Performance

Task performance represents behaviors that enhance the organization's technical core, transforming raw materials into products and services the organization provides to customers (Borman and Motowidlo 1993). Task performance is thus what has traditionally been considered overall job performance, although traditional overall job performance is more multidimensional than even the broad label 'task performance' implies (Campbell, McCloy, Oppler, Sager 1993; Murphy 1996). Because task performance is primarily composed of the technical aspects of one's job, cognitive ability measures tend to be more predictive of task performance than non-cognitive predictors (McHenry, Hough, Toquam, Hanson and Ashworth 1990; Motowidlo and Van Scotter 1994). Therefore, when overall performance is composed primarily of task performance, cognitive ability is likely to be most highly correlated with the performance criterion, and adverse impact against, for example, Blacks is likely, given a $d = 1.0$ mean difference favoring Whites on general cognitive ability measures. Hattrup *et al.* (1997) show statistically that when task performance is a heavily weighted component of overall performance, minority selection ratios decrease and adverse impact increases dramatically. Thus, the more weight given to task performance criteria, predictors dealing with cognitive ability will be weighted more heavily in a regression equation, and more adverse impact is likely to be present in the predictor battery.

Contextual Performance

Contextual performance behaviors help enact task performance and contribute to the organization's climate (Borman and Motowidlo 1993). Supporting others, volunteering to take on more work, exhibiting extra effort on the job, and defending the organization's image and reputation are all individual behaviors relating to contextual performance (Campbell *et al.* 1996; Borman and Motowidlo 1993, 1997). Where cognitive ability predicted task performance better, personality constructs appear to be more strongly related to contextual performance (Hough 1992; McHenry *et al.* 1990; Motowidlo and Van Scotter 1994; Organ and K. Ryan 1995). As discussed above, personality measures generally show smaller subgroup differences and less adverse impact than cognitive ability. Therefore a test battery designed to predict overall performance will tend to show smaller levels of adverse impact to the extent that overall performance is composed of contextual performance variance (Hattrup *et al.* 1997), and to the extent the cognitive and personality predictors are uncorrelated (Sackett and Ellingson 1997).

Knowledge- and Service-Based Performance Constructs

In addition to task and contextual performance distinctions, other dimensions of performance are pertinent to the growing number of knowledge- and service-based organizations found internationally. In these organizations, work is often characterized as highly interactive and social in nature, e.g., dealing with customers and working as a team to carry out tasks. Good prediction of these interpersonal criteria primarily occurs with personality variables or personality composites.

Customer service. Organizations increasingly are competing in a global marketplace, and therefore customer service has become an increasingly important factor in promoting and maintaining good business relationships. Customer service does have bottom-line effects for the organization: measures of organizational climate for customer service have been found to relate empirically to customer perceptions of quality service (Johnson 1996). Providing customers with good service is very different from the traditional definition of task performance. First, the *product* created is often intangible (i.e., the product is a customer's perception or feeling of satisfaction), and thus requires participation by the customer. Second, service production and consumption are closely interdependent (B. Schneider and Bowen 1992), and thus requires a great deal of social interaction and sensitivity (George and Jones 1991). Common dimensions of customer service performance include factors such as demonstrating courtesy, tact, and empathy for customers, providing appropriate communication, and being responsive to customers' needs (Hogan, Hogan and Bush 1984; Parasuraman, Berry and Zeithaml 1991).

Given these interpersonal behaviors are relatively distinct from more technical KSAs (George and Jones 1991), customer service seems related more to personality and temperament constructs than to cognitive ability constructs. Several researchers (Hogan *et al*. 1984; Mount, Barrick and Stewart 1998; Rosse, Miller and Barnes 1991) have found that 'service orientation', a compound measure (see Hough and Schneider 1996) consisting of basic personality traits, predicted customer service performance. The Frei and McDaniel (1998) meta-analysis indicates that these service orientation measures are primarily composed of emotional stability, agreeableness, and conscientiousness, are relatively independent of cognitive ability, and are related to customer service performance (corrected average r = .50). Because customer service performance is primarily determined by personality-based constructs, less adverse impact should be found when customer service performance is the criterion than when traditional task performance is the criterion of interest. However, it is important to note that when situational judgment tests (SJTs) are used to predict customer service performance, small racial and gender mean differences may be found, at least to the extent that cognitive ability is related to situational judgment test performance (Weekley and Jones 1997, 1999).

Teamwork. Over the past few decades, organizations have moved toward carrying out many types of work with groups or teams with interdependent tasks, both out of necessity and out of the advantages that teamwork provides. Introducing team-based structures into organizations creates many challenges for traditional selection system designs because predictors and criteria exist at both the individual and team levels (Jones, Stevens and Fischer, 2000; Klimoski and Jones 1995). Furthermore, team effectiveness may be defined as a function of team process and output criteria (e.g., Brannick and Prince 1997; Hackman 1987; Salas, Dickinson, Converse and Tannenbaum 1992). Because our focus is on individual differences, we are interested in measures of individual teamwork, that is, the extent that an individual supports and facilitates coworkers' duties, helps others with their tasks, and contributes to the effective and efficient functioning of the team (Campbell *et al*. 1993). Notice that teamwork is not the same as contextual performance but is subsumed by it. Teamwork involves interpersonal interactions with individuals in a team context, whereas contextual performance involves supporting task performance with behaviors that may or may not be interpersonal in nature (e.g., volunteering, demonstrating extra effort). Teamwork can further be contrasted with taskwork, which reflects the more technical, task performance behaviors noted above (Jones *et al*., 2000).

Given that teamwork reflects interpersonal behaviors, determinants of teamwork should reflect personality and motivational traits. When teamwork is the focal criterion, test batteries should exhibit relatively lower levels of adverse impact. Some researchers have focused on the process aspects of teamwork. Stevens and Campion (1994) report that the KSAs required for teamwork involve conflict resolution, collaborative problem solving, effective communication, goal setting/performance management and planning/task coordination. Fleishman and Zaccaro (1992) report similar dimensions, suggesting that personality variables are likely to be most predictive of teamwork behaviors. Others have found meaningful empirical relations between teamwork and individual differences on various combinations of personality constructs, including agreeableness, dependability, achievement and emotional stability (Hough 1992); emotional stability, agreeableness and conscientiousness (Mount, Barrick and Stewart 1998), and agreeableness and conscientiousness (vs. cognitive ability and job-specific skills; Neuman and Wright 1999). MacLane (1996) determined that the degree of interpersonal skills required for a job moderated the relationship between a personality-based biodata measure and performance.

Lessons Learned

We have summarized research that relates to possible causes of subgroup differences on predictor measures: measurement method, culture, test coaching, applicant perceptions such as test anxiety and perceptions of the validity of the test, others' stereotype perceptions and criterion composition. Data suggest each one of these factors can contribute to subgroup differences, some appear to contribute significantly to subgroup differences on cognitive ability tests, where Black–White mean differences are most pronounced. Much more research needs to be completed with actual job applicants before we can estimate the magnitude that these factors contribute individually and jointly to subgroup mean score differences.

Evaluation of Statistical Methods for Detecting Adverse Impact and Differential Prediction and for Reducing Adverse Impact

Statistical methods help researchers and practitioners estimate and understand important subgroup differences on psychological measures and constructs systematically. This section addresses statistical approaches for detecting adverse impact (which are based on subgroup mean differences on one variable) and differential prediction (which are based on subgroup mean and/or covariance differences on two variables). Also addressed are statistical methods for detecting test fairness/unfairness and construct equivalence as well as mathematical/statistical methods for reducing adverse impact.

Researchers and practitioners alike routinely draw conclusions about adverse impact and differential prediction that lead to encouraging (or discouraging) critical personnel selection related activities. Such activities may include refining and updating selection tests, adding measures with less adverse impact to a selection test battery, and aggressively recruiting talented job applicants that reflect the types of diversity the organization wants. Whether or not the organization should spend the time, money, and labor on these activities is an important decision. The decision may be proactively motivated, and it may be partially informed by statistical analysis and interpretation of existing adverse impact and differential prediction in the organization. Obviously, statistical findings leading to conclusions about adverse impact or differential prediction (or a lack thereof) should be solidly based on appropriate statistical methods. Furthermore, adequate statistical power ensures accurate judgments about whether the amount of adverse impact or differential prediction is *practically* relevant (given various stakeholders' definitions of *practical*). The US judicial system may give credence to findings based on statistical significance tests or the 4/5ths rule (protected group's selection ratio is at least 4/5ths of the majority group's), but the results of a statistical test do not necessarily inform practical and substantive concerns about adverse impact or differential prediction. For example, the difference between minority-majority selection ratios of .04 and .05 may arguably be less important than the difference between selection ratios of .72 and .90 selection ratios, although both threaten the 4/5ths rule equally.

Differential Prediction

Organizations need to be concerned about whether or not differential prediction exists across subgroups. Using the same overall regression line to predict a performance criterion across subgroups versus using regression lines tailored to maximize prediction within each subgroup reflects 'fairness' in a different way. Only with the Cleary model (Cleary 1968) do these two types of fair prediction agree. Subgroup regression lines can differ due to a variety of subgroup differences on both predictors and criteria: reliability differences, range restriction differences (likely in concurrent validity studies), mean differences, differences in the meaning of the latent construct measured, or any combination thereof. As will be discussed, sample sizes in differential prediction research have tended to be less than is required to detect meaningful subgroup regression line differences when such differences exist.

Moderated Multiple Regression (MMR)

For personnel selection purposes, moderated multiple regression (MMR) tends to be the preferred method for detecting slope and intercept differences between subgroup regression lines (e.g., AERA, APA, NCME 1999, Standard 7.6). Statistically testing for the homogeneity of subgroup criterion-related validity coefficients is the less desirable and less informative alternative (Stone-Romero and Anderson 1994; Stone and Hollenbeck 1989). MMR is preferred because (a) MMR is influenced by subgroup predictor and criterion mean criterion differences as well as subgroup predictor and criterion covariance differences, whereas validity coefficients deal only with standardized covariance differences; and (b) comparing subgroup validity coefficients is equivalent to comparing predictor and criterion scores that are standardized within each subgroup, which is inappropriate in personnel selection.

The basic formula in moderated multiple regression is:

$$Y^{\exists} = a + b_1X + b_2G + b_3XG$$

where X is the predictor score, $Y^{\exists}$ is the predicted criterion score, and G represents group membership. Without loss in generality, assume group membership in G is either 0 or 1. Then it is useful to rewrite this equation as:

$$Y^{\exists} = (a + b_2G) + (b_1 + b_3G)X$$

so it is apparent that:

$$Y^{\exists} = a + b_1X \text{ when } G = 0, \text{and}$$

$$Y^{\exists} = (a + b_2) + (b_1 + b_3)X \text{ when } G = 1$$

This makes it clear that when b_2 and b_3 are not statistically significant, one regression line, $Y^{\exists} = a + b_1X$, can be said to apply across both groups. Regression lines are said to be different but parallel when Group 0 and Group 1 intercepts are different but not the slopes (b_2 is statistically significant but b_3 is not); regression lines are different and not parallel when Group 0 and Group 1 slopes are different (i.e., b_3 is statistically significant). In the latter case, slope differences reflect interaction effects between the predictor and group membership, which precludes interpretation of main effects that would be due solely to intercept differences.

It has been recommended that when differential prediction is an *a priori* hypothesis, the slope-difference coefficient b_3 can be interpreted when it reaches statistical significance, even if the R^2 for the overall equation is not statistically significant (Bedeian and Mossholder 1994; Lautenschlager and Mendoza 1986). Of course, one should then plot the regression lines by subgroup to determine whether the type and extent of statistically significant differences in subgroup regression lines make a practical difference (see Mossholder, Kemery and Bedeian 1990). One can then use the Johnson-Neyman technique to compare statistical differences in subgroup regression lines for exact points along the range of the predictor, for example at points where one might set for predictor cutoff scores (see Pedhazur 1997). In addition to plots, the regression coefficients can be given some substantive interpretation in terms of the change in predicted value of Y given a unit change in X and the individual's subgroup membership, or when the slopes are the same, in terms of systematic over- or underprediction compared with the overall regression line across subgroups.

Statistical power of MMR. Statistical power in organizational research has traditionally been found to be extremely low for detecting non-zero relationships (Mone, Mueller and Mauland 1996). Perhaps it is now a recognized fact that the required statistical power of MMR studies should be much higher, because one wants to investigate whether slope and intercepts differ *from each other*, not just from the 'nil hypothesis' of zero effect. As a result of low sample sizes and the low statistical power that results, is has been possible to overlook even large amounts of differential prediction when it exists.

The statistical power of MMR increases when the slope differences to be detected are high, when sample sizes are high, and when subgroup proportions approach equality (Stone-Romero, Alliger and Aguinis 1994). Even given these favorable conditions, sample sizes often need to be extremely high, higher than what is typically found in organizational research. Aguinis, Beaty, Boik and Pierce (2000) reviewed the statistical power of moderated multiple regression for detecting differential prediction in 30 years' worth (1969–1998) of differential prediction literature covered in three major business and psychology journals: *Journal of Applied Psychology*, *Personnel Psychology* and *Academy of Management Journal*. They focused on 261 differential prediction studies and varied the reliabilities and overall selection ratios when such information was not reported in a study. The median level of power across conditions was disappointingly low, .18. The variability about the median did not offer much more promise: across conditions power typically hovered around .20, with an average of only .43 in the best-case scenario (recall that .80 is a well-known rule-of-thumb for adequate statistical power). Generally, more contemporary MMR studies showed only slightly better statistical power; much more power is often needed.

Sample sizes generally need to be extremely high to detect slope differences in MMR. For a total sample size of 100, Aguinis and Pierce (1998b) found that power was adequate only when sample sizes were nearly equal (minority group no less then 40% of the majority group), and slope differences were extreme (correlations of .2 and .8). Aguinis and Stone-Romero (1997) report similar trends and include the power-reducing effects of range restriction: power was generally acceptable ($\geq$.70) for total $N = 300$ when slope differences were high (correlation differences of .4 or greater) and direct range restriction on the predictor was low (20% or less).

Johnson, Carter, Davison and Oliver (in press) suggest a statistical method that leverages the statistical power from the total sample size across job families to better estimate differential prediction for the smaller sample sizes within job families. Their method, called synthetic differential prediction analysis, or SDPA, tests for differential prediction by using correlations between performance scores on job-family-relevant job components, test scores, subgroup membership, and test-subgroup cross-products across job families. Such correlations need to be available from the data at hand, each job family must have some components that overlap with components in other job families, and the measures of job performance components must be equivalent across job families. The advantage of SDPA is that it capitalizes on performance component overlap across job families, so estimates about differential prediction can be conducted for job families with small *N* – or even for job families lacking data for a particular subgroup.

Homogeneity of error variance assumptions in MMR. Homogeneity of errors in prediction is required to satisfy the F-test in moderated multiple regression. Computer simulations have shown that error variance ratios at or exceeding a 1.5:1 ratio – especially with unequal sample sizes and when the larger sample size is paired with the smaller error variance – tends to create misleading results in the F-test for homogeneous regression slopes (i.e., a greater tendency to assume mistakenly that slopes are homogeneous or heterogeneous; see Alexander and DeShon 1994; DeShon and Alexander 1996). Fortunately for the body of past MMR studies, Oswald, Saad and Sackett (2000) found that, at least in a large sample of ability and personality data predicting job performance ratings, variance heterogeneity was generally not great enough to distort conclusions from the traditional F-test. Violations did occur enough (about 10% of the time) and in unpredictable ways such that a routine statistical test for the homogeneity of error variance is warranted when the F-test is applied to an MMR analysis. A few alternative statistical tests do not require satisfying the variance-homogeneity assumption, and those tests should be considered. Note that low overall sample sizes, disproportionate subgroup sample sizes, statistical artifacts (predictor and criterion unreliability, range restriction) all serve to impair the ability to detect differential prediction regardless of the statistical test (see the thorough MMR review by Aguinis and Pierce 1998a).

Reliability and Latent Variable Relationships

Although predictors in test-score banding (a strategy used to reduce adverse impact that is described below) are treated in a manner that explicitly recognizes error of measurement, prediction in regression assumes the predictor is without error (fixed-effects). Low reliability attenuates or eliminates observed interaction effects; reliability differences between subgroups would distort interaction effects. Fuller (1987) provides methodology estimating regression weights and associated standard errors for linear regression equations based on a correlation matrix corrected for attenuation due to measurement unreliability, although this method has not been fleshed out for interaction terms.

Differential reliabilities could make regression lines appear similar when they are not, or not similar when they are. Ghiselli (1963) noted that both the standard error of measurement and the standard error of the estimate might not be the same for all subgroups; reliabilities and validities may differ for important subgroups within a dataset, although he was mindful of the effects of sampling error on estimating these differences. Errors-in-variables regression (EIVR) corrects for predictor unreliability; Anderson, Stone-Romero and Tisak (1996) found that EIVR increases the chance for moderator detection when reliabilities are high (at least .65) and sample sizes are high – no method works very well when sample sizes are low. High reliability is a must in order to determine the nature and extent of differential prediction via MMR or any other method.

Latent variable models require constructs that have adequate indicators and measurement models associated with them. For example, both overall and dimensional job performance ratings have been known to have low interrater reliabilities (around .50 on average, according to Viswesvaran, Ones and Schmidt 1996), which can muddy any understanding of construct differences and differences in latent relationships between subgroups. Similar to MMR, latent structural model comparisons between subgroups requires very large samples for good estimation; they also require making difficult decisions about appropriate model constraints and dealing with non-normality with the product or interaction terms in the model (Jöreskog and Yang 1996). Single-indicator models seem to fare better in terms of model convergence and model fit than multiple-indicator models. This does not imply that such models are better, just that they are more tractable (Li, Harmer, Duncan, Duncan, Acock and Boles 1998). In fact single-indicator models challenge the very purpose of structural equation models with latent variables that are derived for multiple indicators. Latent variable models with multiple indicators are useful for model-fitting and accurate parameter estimation, but they do not appear to increase the power to detect an interaction effect (e.g. subgroup-test interaction) when one is present (DeShon, Horvath and Sacco, 2000), although this is somewhat at variance with the EIVR findings above.

Factorial and Predictive Invariance

Before one focuses on differential prediction in predicting a latent criterion, it is important to determine whether the criterion is psychometrically equivalent across subgroups, so that we can assume the same construct is being measured. A set of interesting papers (Millsap and Meredith 1992; Millsap 1995, 1997; Terris 1997) demonstrates that data essentially can never show both construct equivalence and a lack of differential prediction between two groups at the same time: if slopes and intercepts are the same between the observed variables x and y, the latent slopes and intercepts must differ, unless a particular mathematical constraint is satisfied that no data would be expected to fit. Research needs to determine the extent to which *practical* construct equivalence and a *practical* lack of differential prediction can both hold.

Test Fairness

The construct equivalence between protected and non-protected subgroups has not been closely linked to the test fairness literature to date. Traditionally, the type of personnel selection model adopted and the resulting selection outcomes, independently or together, define a *fair* selection test. Petersen and Novick (1976) reviewed several models of test fairness. Although it is a relatively old paper, it covers most of the test fairness models discussed today. Each test fairness model sets different equalities between two or more subgroups:

(a) Equal regression lines (Cleary 1968, as clarified by Norborg 1984).
(b) An equal ratio of (1) the proportion within a subgroup the test selects to (2) the proportion within a subgroup who would be successful absent a selection test (Thorndike 1971).
(c) Equal proportions of successful performers are selected by the test (Cole 1973).
(d) Equal base rates of success out of those selected (Petersen and Novick 1976).

Darlington's cultural-fairness models essentially parallel models (b), (c), and (d) above (Darlington 1971). Petersen and Novick (1976) go on to suggest three equally viable alternatives focusing on fairness for rejected applicants from different subgroups:

(e) An equal ratio of (1) the proportion within a subgroup the test rejects to (2) the proportion within a subgroup who would be unsuccessful absent a selection test.
(f) Equal rejection rates for unsuccessful performers.
(g) Equal failure rates in those not selected.

The authors also cite Einhorn and Bass (1971), who suggested a model with

(h) Equal base rates of success for each subgroup *at* their respective cutoff scores.

In summary, the choices about how to define and operationalize *test fairness* will in part determine whether fairness is supported or challenged psychometrically or substantively. Psychometrically, given mean subgroup differences combined with top-down selection (or the same test score cutoff in fixed selection), all models except the Cleary model are unlikely to be satisfied without some form of score adjustment within subgroups (Hunter and Schmidt 1976), which the 1991 Civil Rights Act makes illegal in practice (Sackett and Wilk 1994). Further, with the exception of the Cleary model and the Einhorn and Bass model, it is virtually impossible to have a personnel selection decision that is simultaneously 'fair' to both selected applicants and rejected applicants (i.e., making the fairness definition apply to all false positives and false negatives). The Cleary model of equal regression lines has turned out to be the preferred model of fairness (cf. Maxwell and Arvey 1993; Campbell 1996). Perhaps that is because of (a) this problem; (b) MMR historically has had low statistical power to detect actual slope and intercept differences; (c) a single regression line across all individuals can be considered 'race neutral'; or (d) a single regression line is easier to apply and interpret.

Substantively, the research we have summarized goes beyond psychometric models of test fairness, and future research should continue to do so. Although most definitions imply some form of psychometric comparability (e.g., similar group-mean differences, similar levels of criterion-related validity, similar implications for test score use), psychometric comparability alone does not imply test fairness. Test fairness involves psychometrics, but it also involves a number of important judgment calls about test design, development, and administration; the meaning of test scores; how test scores are used; and the implications, risks, and rewards for test score use within a complex organizational, legal, and social context (Sireci and Geisinger 1999; Willingham 1999).

Evaluating the utility of selection decisions should reflect the above considerations because useful measures of utility incorporate the statistical outcomes of a selection process as well as some process allowing organizational stakeholders to state and incorporate the relative *value* of various selection decisions explicitly (Austin, Klimoski and Hunt 1996). The utility of a fair test and a fair overall selection system may be found to constrain the organization's expected marginal utility gains from correct prediction (true positives, true negatives) and losses from errors in prediction (false positives, false negatives). Utility for all stakeholders may not be entirely satisfied, but perhaps a maximally satisficing strategy can be adopted (e.g., the minimax strategy). Even though most measures of utility are imperfect and limited, attempts to establish utility are invaluable for the fact that it leads to an explicit participative process for evaluating the components a selection system versus buying into the selection system automatically.

Construct Equivalence

Construct equivalence between subgroups is an important prerequisite before one can give any clear and substantive interpretation to adverse impact and differential prediction findings. Even though the statistical methods have been around for a while, 'measurement invariance is rarely tested in organizational research' (Vandenberg and Lance 2000, p. 4). Researchers and practitioners are just beginning to use and appreciate various tools and standards for determining construct equivalence. DIF has become a popular method for assessing construct equivalence (Holland and Wainer 1993) which we have already referred to in the context of assessing potential cultural influences on construct equivalence for cognitive ability tests. Vandenberg and Lance (2000) wrote a thorough review of confirmatory factor analysis approaches to assessing construct equivalence.

Differential item functioning. Statistical tests for determining whether items function differently across different subgroups (e.g., for Blacks and Whites, whether ability items are more or less difficult, or whether items are more or less accurate at measuring the underlying ability level, etc.). DIF is traditionally based on unidimensional item-response-theory models, and unidimensionality is rarely tenable in a strict sense, especially when the construct is broad (e.g., general cognitive ability). DIF may still, however, provide useful information about whether a measure operates differently between subgroups of individuals (e.g., Black/White, male/female).

We mentioned that ability test items have not been shown to have systematic differential item functioning between minority subgroups. This does not mean that subgroup differences cannot exist or should not be researched further; it means that ability and aptitude data collected and analyzed thus far tend to support interpreting mean test score differences between subgroups as differences in levels of the same construct, not different constructs. In a high-school test of academic skills (reading, writing, mathematics), Fan, Willson and Kapes (1996) found little evidence that varying the proportion of ethnic representation (White, African American and Hispanic) in test-development groups affects bias against those groups, even when the test-development group was 100% of one ethnicity.

As with other statistical methods discussed, we encourage exploratory data analysis to supplement statistical tests for DIF. Range restriction and mean differences on the latent construct can affect DIF statistics, and empirical item characteristic curves can explore those effects. Empirical curves can demonstrate the magnitude of item-difficulty differences across the range of the construct being measured (Dorans and Kulick 1986), and whether those differences are practical differences (e.g., differences in the mid-range or lower-range of ability may be more important than differences at the high extreme).

Weighting of Criteria and Predictors: Implications for Adverse Impact

We have mentioned that differential weights to both predictors and criteria can lead to different implications for adverse impact and its effects on criterion-related validity. A group of recent studies shows that varying predictor and criterion weights affects both adverse impact and resulting validity coefficients (Bobko, Roth and Potosky 1999; Hattrup, Rock and Scalia 1997; N. Schmitt, Rogers, Chan, Sheppard and Jennings 1997). It is a complex multivariate problem to weight predictors and criteria using statistical weights (e.g., least-squares regression weights), rational weights (e.g., weights derived from a job analysis), or both. Statistical and rational weights both reflect the value of a predictor in the context of the other predictors and criteria, but *value* is defined differently. Rational weights reflect the values of the organization and decision-makers; the weights may account for intercorrelation between predictors and criteria implicitly. Statistical (least-squares) weights account for intercorrelations between variables explicitly, such that the weighted composite predicts in an optimal (least-squares) sense.

It is difficult, perhaps, to compare or combine rational and statistical weights applied to predictors and criteria, although in general it is sensible and appropriate for statistical approaches to inform rational approaches and vice-versa. The problem of combining rational and statistical weights may be a little more straightforward when dealing with domains with low correlations (e.g., ability and personality predictors, task and contextual performance) where multicollinearity is less of a problem or confound. Schmidt and Kaplan (1971) recommend the use of criterion composites for statistical prediction, and the study of predictor battery – single criterion relationships for psychological understanding. Psychological variables are intercorrelated in almost every domain (Meehl 1997), which may affect interpretability of both rational and statistical weights. To deal with this for regression weights, a recent statistical method transforms regression weights into weights that reflect the predictors' importance, where *importance* is defined as the proportionate contribution to the criterion in a regression model after accounting for predictor intercorrelations (Johnson, 2000). This method is a related but computationally efficient alternative to dominance analysis (Budescu 1993), which computes regression analysis for all (2^n-1) subsets of n predictor variables.

In addition to these weighty weighting issues, consider the fact that regression weights cannot simultaneously minimize adverse impact for all protected groups (e.g., race and gender) and maximize criterion-related validity (Hoffman and Thornton 1997; Sager, Peterson, Oppler, Rosse and Walker 1997). It is a necessary compromise to balance factors the organization values, such as criterion-related validity, low adverse-impact, job-relatedness and legal defensibility. It is best to be explicit about trying to achieve such a balance, because such compromising is done whether the organization recognizes it or not.

Particular personnel selection situations may justify including and weighting personality-based (and other less-cognitively-based) predictors and criteria more heavily than one might typically or traditionally. As we have mentioned, two general and international trends describing the nature of work suggest a greater need for personality variables in conducting work-related activities successfully: (1) the increase in service-oriented jobs (employees deal with customers or clients) and (2) the increase in the use of teams and teamwork in organizations (employees deal with and depend on coworkers). Job analysis data have also suggested the importance of personality within jobs. For example, job analysts have rated interpersonal skills and physical abilities as work requirements that are at least as important as cognitive ability for jobs such as firefighter (Bownas 1988; Hornick and Axton 1998). Including personality measures in the predictor battery in combination with cognitive ability measures has been found to increase the criterion-related validity for performance ratings, while also serving to reduce – though not eliminate – adverse impact.

Cognitive predictors often must receive some weight in a regression equation for personnel selection because most job performance criteria are to some extent cognitive ability laden. It is important to note, then, that adverse impact as defined by the 4/5ths rule in the US legal system is often extremely difficult to overcome because predictor batteries often contain cognitive ability tests. Take, for example, a typical cognitive ability test with a one-standard-deviation standardized mean difference between Blacks and Whites. One may seek to reduce this one-standard deviation difference by adding non-cognitive predictors with little or no adverse impact. However, without high selection ratios (select > 50% of the applicants), adding a non-cognitive predictor will tend to have little impact on satisfying the 4/5ths rule (i.e., making selection ratios between subgroups more similar). Both fundamental psychometric principles (Sackett and Ellingson 1997) and applied research (Ryan, Ployhart and Friedel 1998; N. Schmitt *et al.* 1997) bear out this unfortunate fact. Equating the means of the cognitive ability or predictor composite scores within each subgroup is clearly not a legal option (Sackett and Wilk 1994), but fortunately there exists a host of legal alternatives with potential (see the review by Sackett *et al.* 2001).

Test-score Banding

To reduce adverse impact, one personnel selection practice involves test score banding, i.e., establishing fixed or sliding score bands based on the standard error of measurement, treating members within a band as 'equivalent' and turning to other methods for selecting individuals from within the test score band. The US judicial system has ruled that fixed test-score banding, with selection based on race within bands, is an appropriate remedy for employment discrimination, as long as that practice is removed once the lingering effects of past discrimination are remedied (*Chicago Fire Fighters Union Local No. 2 v. City of Chicago* 1999).

In a move away from traditional banding methods, Aguinis, Cortina and Goldberg (1998) offered a new method for test score banding, adjusting test score bands for criterion reliability and the criterion-related validity coefficient. This new method reflects the philosophy of the standard error of the estimate: lower validities and unreliable criteria should translate into wider predictor bands, because test scores are less able to predict accurately. This philosophy has already been challenged (Hanges, Grojean and Smith 2000) and defended (Aguinis, Cortina and Goldberg 2000).

This standard error of the estimate approach is a compelling alternative to more traditional banding procedures. Consider a general point made through an admittedly unrealistic situation: Assume a perfectly reliable test ($r_{xx} = 1$), but a validity of zero ($r_{xy} = 0$). Traditional test-score bands would be of zero width, because any observed test score difference reflects a true score difference, and top-down selection would be recommended. However, because in this particular case validity equals zero, *random* selection on the test (or any type of selection) would be as useful as top-down selection. To the extent that criterion-related validity exists, then top-down selection has added value; the main point here is whether test-score bands should reflect criterion-related validity. Extending Aguinis *et al.* (2000), one could create bandwidths that incorporate *differential* predictor and criterion reliability as well as *differential* validities, but this clearly opens the door to a tremendous amount of complexity for test score banding practices. Fancy banding procedures incorporating these additional complexities may be unnecessary when one keeps in mind the bottom-line purpose of test-score banding: balancing the criterion-related validity of a selection test with diversity in hiring outcomes for an organization.

In summary, the goals of banding for promoting diversity within an organization are often worthwhile, and banding procedures may be useful to this end, but the psychometric underpinnings of test-score banding do deserve a closer look. Test-score bands may define statistical differences in test scores that may not translate into practical differences in predicted performance scores, or conversely, test-score bands may not recognize statistical differences in test scores that would actually translate into practical differences in predicted performance if predictors were more reliable across subgroups. Furthermore, regardless of the test-score banding method used, organizations should attempt to assess explicitly – at least through discussion if not calculation – the expected gains and losses from banding. Practical gains in diversity may or may not lead to practical amounts of marginal loss in performance utility.

New Directions in Adverse Impact and Differential Prediction Research

We have summarized the mean subgroup differences across several predictor domains (cognitive ability, personality and physical ability). We reviewed empirical and conceptual work on the possible causes and remedies for adverse impact (measurement method *vis-à-vis* the constructs measured, culture, test coaching, test-taker perceptions, stereotype threat and criterion conceptualizations). We then concluded with a review and evaluation of statistical methods of detecting adverse impact, differential prediction, test fairness/unfairness, and construct equivalence as well as a review of mathematical/statistical strategies for reducing adverse impact. Each major section in this article has been summarized above. In our brief parting thoughts, we offer new directions and ideas in research and practice dealing with adverse impact and differential prediction in employment testing.

Correlates, Processes and Outcomes of Group Differences

We think that future research and practice could better recognize distinctions and relationships between the *outcomes* resulting from decisions related to group differences and the *correlates* and *processes* that contribute to group differences. Organizations are interested in group-related outcomes when it comes to reducing adverse impact and meeting legal standards for avoiding employment discrimination (e.g., the 4/5ths rule in the USA). For example, in test-score banding, applicants' group membership within a band is a salient characteristic used for diversity-promoting purposes. In this case, focusing on the outcome of group membership in test-score banding is often a practical and necessary expedient for reducing adverse impact. It is no surprise, then, that adverse impact research has focused on group-mean differences on predictor variables and the differential subgroup selection ratios they might produce in hiring.

However, in addition to an outcome focus, our review suggested different correlates and processes that might be responsible for some of these group-mean differences. A real science of group-level differences on predictor measures will require an understanding of more specifics about the nature of group membership, the substantive causes for group differences, when to expect to find group differences, and to what extent group differences can change.

In the particular case of race, we suppose that those performance differences over which individuals have some control are those due to those individual differences that may be *related* to race (e.g., culture, stereotype threat, test-taking strategies, exposure to educational and vocational opportunities), not to race itself. Race is a distal indicator of what is really causing the difference. Only when we attempt to understand the *meaning* behind race differences will we find more proximal and powerful explanations of various group mean differences, covariance differences, differential item functioning, or differential test functioning (Sue 1999). To address important adverse impact and discrimination issues in employment testing adequately, researchers should not only examine group mean differences on variables, but should start investigating the processes/mechanisms that lead toward proactively understanding and reducing unfairness in employment testing. I/O psychology is beginning this sort of research with the studies on culture, test-taking motivation and strategy, and test perceptions reviewed above. Wolfle (1985) represents a good example of this sort of research in educational psychology. In his study, background variables (parents' education and occupation), situational characteristics (high school curriculum), and personal characteristics (ability, high-school grades) were found for Blacks and Whites in the USA to be equally predictive of US educational attainment after one adjusts for measurement error between groups. Educational attainment appears to be a matter of ability, motivation and being exposed to equal opportunity regardless of race, and presumably this would be true for on-the-job performance that education contributes to and predicts.

Psychometric Requirements and Practical Judgments in Differential Prediction and Test-Score Banding

The research imperative for detecting differential prediction can be simply stated. Statistical power required for testing differences in subgroup regression lines has been sorely lacking. Future differential prediction research needs to parallel Cohen's (1988) directive to pay greater attention to statistical power rather than focusing on the null hypothesis. For differential prediction, one should: (1) determine the minimal amount of differential prediction (slope differences, intercept differences) that makes a practical difference; (2) determine the sample size required overall and between groups to detect such a difference; and (3) when statistical differences in prediction are not found, one can say with some confidence that practical differences in prediction do not exist.

A similar line of argument can be made for future test-score banding work, in that enough test-score reliability is needed to estimate with reasonable accuracy whether score bands used for reducing adverse impact leads to meaningful amounts of performance or utility loss. Employment tests of all types should strive toward higher reliability coefficients (given an appropriate definition of *reliability*), even though greater reliability leads to smaller test-score bands and higher d values. To the extent that tests in a predictor composite are unreliable, this adds random 'noise' to the composite, reducing d values and adverse impact. Using low-reliability tests as predictors is clearly *not* an appropriate way to reduce adverse impact in employment settings.

Consideration of Alternative Measurement Methods and Constructs for Reducing Adverse Impact

In documenting group mean differences on predictors, we have reflected on past research and indicated what the future might predict for measures of similar constructs for similar groups. Given high d values, Sackett and Ellingson (1997) convincingly illustrate a psychometric truism: it is extremely difficult to add tests to a selection battery to ameliorate a large subgroup difference on a single test, such as the $d = 1.0$ US Black–White difference in means on general cognitive ability tests. By no means do we imply, however, that reducing adverse impact is a lost cause.

Personnel selection research has begun to explore new predictor constructs that are important in today's world of work, which is becoming more service-oriented and international in scope. Measures of predictor constructs related to teamwork and customer service show promise for reducing adverse impact, as do measures that retain their essence while reducing content that is not job-related (e.g., a video format versus a paper-and-pencil format, as the latter may be pitched at an unnecessarily high reading level). Profitable research on alternative predictor constructs and methods remains to be conducted. Developing job-relevant or job-related employment tests with adequate criterion-related validity and reduced adverse impact is a highly relevant pursuit.

Finally, as organizations become increasingly international in scope, it is essential that employment test development and measurement methods adapt and accommodate diverse cultural, cross-cultural, and legal contexts. Furthermore, internationalization has resulted in new types of jobs that require different or expanded batteries of personnel selection tests. Take the job of international manager, a job that is shown to require adaptability, resilience, political and cultural awareness and sensitivity, and family support (Nyfield and Baron 2000). Measures of these types of constructs may be increasingly used in personnel selection for jobs that require their employees to work with different coworkers and clients from different countries and cultures. The tasks and challenges of personnel psychologists, human resources professionals, and organizational consultants will only increase as technology, travel, and trade bring the world of work – and the work of worlds – closer together.

Note

1. Carroll's (1993) model has eight factors at the level of fluid intelligence and crystallized

intelligence: memory and learning, visual perception, auditory perception, cognitive speediness, retrieval ability, and decision speed.

References

Ackerman, P.L. (1996) A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence,* **22**, 227–257.

Ackerman, P.L. and Heggestad, E.D. (1997) Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin,* **121**, 219–245.

Ackerman, P.L. and Rolfhus, E.L. (1999) The locus of adult intelligence: Knowledge, abilities, and nonability traits. *Psychology and Aging,* **14**, 314–330.

Aguinis, H., Beaty, J.C., Jr., Boik, R.J. and Pierce, C.A. (2000) Statistical power of differential predication analysis: A 30-year review. Symposium presented at the 15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Aguinis, H., Cortina, J.M. and Goldberg, E. (1998) A new procedure for computing equivalence bands in personnel selection. *Human Performance,* **11**, 351–365.

Aguinis, H., Cortina, J.M. and Goldberg, E. (2000) A clarifying note on difference between the W.F. Cascio, J. Outtz, S. Zedeck and I.L. Goldstein (1991) and H. Aguinis, J.M. Cortina and E. Goldberg (1998) banding procedures. *Human Performance,* **13**, 199–204.

Aguinis, H. and Pierce, C.A. (1998a) Heterogeneity of error variance and the assessment of moderating effects of categorical variables: A conceptual review. *Organizational Research Methods,* **1**, 296–314.

Aguinis, H. and Pierce, C.A. (1998b) Statistical power computations for detecting dichotomous moderator variables with moderated multiple regression. *Educational and Psychological Measurement,* **58**, 668–676.

Aguinis, H. and Stone-Romero, E.F. (1997) Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology,* **82**, 192–206.

Alexander, R.A. and DeShon, R.P. (1994) Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin,* **115**, 308–314.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999) *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Anastasi, A. (1958) *Differential Psychology.* New York: Macmillan.

Anderson, L.E., Stone-Romero, E.F. and Tisak, J. (1996) A comparison of bias and mean squared error in parameter estimates of interaction effects: Moderated multiple regression versus errors-in-variables regression. *Multivariate Behavioral Research,* **31**, 69–94.

Arvey, R.D. *et al.* (1994) Mainstream science on intelligence. *Wall Street Journal,* editorial, December 13, p. 356.

Arvey, R.D., Strickland, W., Drauden, G. and Martin, C. (1990) Motivational components of test-taking. *Personnel Psychology,* **43**, 695–716.

Austin, J.T., Klimoski, R.J. and Hunt, S.T. (1996) Dilemmatics in public sector assessment: A framework for developing and evaluating selection systems. *Human Performance,* **9**, 177–198.

Bahrick, H.P., Bahrick, P.O. and Wittlinger, R.P. (1974) Long-term memory: Those unforgettable high-school days. *Psychology Today,* **8**, 50–56.

Baltes, P.B. and Schaie, K.W. (1976) On the plasticity of intelligence in adulthood and old age: Where Horn and Donaldson fail. *American Psychologist,* **31**, 720–725.

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

Baumeister, R.F., Hamilton, J.C. and Tice, D.M. (1985) Public versus private expectancy of success: Confidence booster or performance pressure? *Journal of Personality and Social Psychology,* **48**, 1447–1457.

Becker, B.J. (1989) Gender and science achievement: A reanalysis of studies from two meta-analyses. *Journal of Research in Science Teaching,* **26**, 141–169.

Bedeian, A.G. and Mossholder, K.W. (1994) Simple question, not so simple answer: Interpreting interaction terms in moderated multiple regression. *Journal of Management,* **20**, 159–165.

Begley, S. (2000) The stereotype trap. *Newsweek,* November 6, pp. 66–68.

Benbow, C.P. (1988) Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects and possible causes. *Behavioral and Brain Sciences,* **11**, 169–183.

Berry, J.W. (1996) A cultural ecology of cognition. In I. Dennis and P. Tapsfield (eds.), *Human Abilities: Their Nature and Measurement*. Mahwah, NJ: Lawrence-Erlbaum, Inc.

Binning, J.F. and Barrett, G.V. (1989) Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology,* **74**, 478–494.

Block, J. (1995) A contrarian view of the five-factor approach to personality description. *Psychological Bulletin,* **117**, 187–215.

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

Bobrow, W. and Leonards, J.S. (1997) Development and validation of an assessment center during organizational change. *Journal of Social Behavior and Personality,* **12**, 217–236.

Bond, L. (1993) Comments on the O'Neill and McPeek paper. In P.W. Holland and H. Wainer (eds.), *Differential Item Functioning.* New Jersey: Erlbaum.

Borman, W.C. and Motowidlo, S.J. (1993) Expanding the criterion domain to include elements of contextual performance. In N. Schmitt, W.C.

Borman, and Associates, *Personnel Selection in Organizations.* San Francisco: Jossey-Bass.

Borman, W.C. and Motowidlo, S.J. (1997) Task performance and contextual performance: The meaning for personnel selection research. *Human Performance,* **10**, 99–109.

Botwinick, J. (1967) *Cognitive Processes in Maturity and Old Age*. New York: Springer.

Botwinick, J. (1977) Intellectual abilities. In J.E. Birren and K.W. Schaie (eds.), *Handbook of the Psychology of Aging*. New York: Van Nostrand Reinhold.

Bownas, D.A. (1988) Firefighter. In G. Gael (ed.), *The Job Analysis Handbook for Business, Industry, and Government*, vol. 2. New York: John Wiley and Sons.

Boykin, A.W. (1983) The academic performance of Afro-American children. In J.T. Spence (ed.), *Achievement and Achievement Motives*. San Francisco: Freeman.

Boykin, A.W. (1994) Harvesting talent and culture: African-American children and educational reform. In R. Rossi (ed.), *Schools and Students at Risk*. New York: Teachers College Press.

Brannick, M.T. and Prince, C. (1997) An overview of team performance measurement. In M.T. Brannick, E. Salas and C. Prince (eds.), *Team Performance Assessment and Measurement: Theory, Methods, and Applications*. New Jersey: Erlbaum.

Brown, B. and Campion, M.A. (1994) Bidoata phenomenology: Recruiters' perceptions and use of biographical information in resume screening. *Journal of Applied Psychology,* **79**, 897–908.

Budescu, D.V. (1993) Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin,* **114**, 542–551.

Camilli, G. and Shepard, L.A. (1994) *Methods for Identifying Biased Test Items.* Thousand Oaks, CA: Sage.

Campbell, J.P. (1990) Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette and L.M. Hough (eds.), *Handbook of Industrial and Organizational Psychology* (2nd edn.; Vol. 1). Palo Alto, CA: Consulting Psychologists Press.

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

Campbell, J.P., Gasser, M.B. and Oswald, F.L. (1996) The substantive nature of job performance variability. In K.R. Murphy (ed.), *Individual Differences and Behavior in Organizations*. San Francisco, CA: Jossey-Bass.

Campbell, J.P., McCloy, R.A., Oppler, S.H. and Sager, C.E. (1993) A theory of performance. In N. Schmitt and W.C. Borman (eds.), *Personnel Selection in Organizations*. San Francisco: Jossey-Bass.

Carroll, J.B. (1993*) Human Cognitive Abilities: A Survey of Factor-Analytic Studies.* New York: Cambridge University Press.

Cattell, R.B. (1987) *Intelligence: Its Structure, Growth, and Action*. (revised and reprinted from *Abilities: Their Structure, Growth and Action* 1971, Boston: Houghton-Mifflin). Amsterdam: North Holland.

Chan, D. (1997) Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology,* **82**, 311–320.

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

Chan, D., Schmitt, N., DeShon, R.P., Clause, C.S. and Delbridge, K. (1997) Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions and test-taking motivation. *Journal of Applied Psychology,* **82**, 300–310.

Cheryan, S. and Bodenhausen, G.V. (2000) When positive stereotypes threaten intellectual performance: The psychological hazards of 'model minority' status. *Psychological Science,* **11**, 399–402.

Chicago Fire Fighters Union Local No. 2 vs. City of Chicago, 1999 U.S Dist. LEXIS 20310 (Dec. 30, 1999).

Church, A.T. and Katigbak, M.S. (1988) The emic strategy in the identification and assessment of personality dimensions in a nonwestern culture: Rationale, steps, and a Philippine illustration. *Journal of Cross-Cultural Psychology,* **19**, 140–163.

Clapham, M.M. and Fulford, M.D. (1997) Age bias in assessment center ratings. *Journal of Managerial Issues,* **9**, 373–387.

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

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd edn.). Hillsdale, NJ: Erlbaum.

Cole, N.S. (1973) Bias in selection. *Journal of Educational Measurement,* **10**, 237–255.

Cooper Institute for Aerobic Research (n.d.). *Physical Fitness Assessment*. Dallas, TX: Authors.

Craik, F.I.M. (1977) Age differences in human memory. In J.E. Birren and K.W. Schaie (eds.), *Handbook of the Psychology of Aging*. New York: Van Nostrand Reinhold.

Darlington, R.B. (1971) Another look at 'culture fairness'. *Journal of Educational Measurement,* **8**, 71–82.

DeShon, R.P. and Alexander, R.A. (1996) Alternative procedures for testing regression slope homogeneity when group error variances are unequal. *Psychological Methods,* **1**, 261–277.

DeShon, R.P., Horvath, M. and Sacco, J. (2000) Using SEM to overcome limitations of the regression test for differential prediction. Symposium at the 15th Annual Society of Industrial and Organizational Psychology, New Orleans, LA.

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

Desmarais, L.B. and Sackett, P.R. (1993) Investigating a cognitive complexity hierarchy of jobs. *Journal of Vocational Behavior,* **43**, 279–297.

Dohm, T.E. and Hough, L.M. (1993) *'Universal Test Battery' Cut Score Analyses and Revisions* (Institute Report No. 242). Minneapolis: Personnel Decisions Research Institutes, Inc.

Dorans, N.J. and Kulick, E. (1986) Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, **23**, 355–368.

Einhorn, H.J. and Bass, A.R. (1971) Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, **75**, 261–269.

Fan, X., Willson, V.T. and Kapes, J.T. (1996) Ethnic group representation in test construction samples and test bias: The standardization fallacy revisited. *Educational and Psychological Measurement*, **56**, 365–381.

Feingold, A. (1992) Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*, **62**, 61–84.

Feingold, A. (1994) Gender differences in personality: A meta-analysis. *Psychological Bulletin*, **116**, 429–456.

Fleishman, E.A. and Quaintance, M.K. (1984) *Taxonomies of Human Performance: The Description of Human Tasks*. Orlando, FL: Academic Press, Inc.

Fleishman, E.A. and Zaccaro, S.J. (1992) Toward a taxonomy of team performance functions. In R.W. Swezey and E. Salas (eds.), *Teams: Their Training and Performance*. Norwood, NJ: Ablex.

Fleming, M.L. and Malone, M.R. (1983) The relationship of student characteristics and student performance in science as viewed by meta-analysis research. *Journal of Research in Science Teaching*, **20**, 481–495.

Flynn, J.R. (1984) The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, **95**, 29–51.

Flynn, J.R. (1987) Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, **101**, 171–191.

Flynn, J.R. (1998) IQ gains over time: Toward finding the causes. In U. Neisser (ed.), *The Rising Curve: Long-term Gains in IQ and Related Measures*. Washington, DC: American Psychological Association.

Freedle, R. and Kostin, I. (1997) Predicting Black and White differential item functioning in verbal analogy performance. *Intelligence*, **24**, 417–444.

Frei, R.L. and McDaniel, M.A. (1998) Validity of customer service measures in personnel selection: A review of criterion and construct evidence. *Human Performance*, **11**, 1–27.

Fuller, W.A. (1987) *Measurement Error Models*. New York: Wiley.

Gandy, J.A., Dye, D.A. and MacLane, C.N. (1994) Federal government selection: The individual achievement record. In G.S. Stokes, M.D. Mumford and W.A. Owens (eds.), *Biodata Handbook: Theory, Research, and Use of Biographical Information in Selection and Performance Prediction*. Palo Alto, CA: Consulting Psychologist Press, Inc.

Garai, J.E. and Scheinfeld, A. (1968) Sex differences in mental and behavioral traits. *Genetic Psychology Monographs*, **77**, 169–299.

Gardner, H. (1993) *Multiple Intelligences: The Theory in Practice*. New York: Basic Books.

Garth, T.R. (1925) A review of racial psychology. *Psychological Bulletin*, **22**, 343–364.

George, J.M. and Jones, G.R. (1991) Towards an understanding of customer service quality. *Journal of Managerial Issues*, **111**, 220–238.

Ghiselli, E.E. (1963) Moderating effects and differential reliability and validity. *Journal of Applied Psychology*, **47**, 81–86.

Ghorpade, J., Hattrup, K. and Lackritz, J.R. (1999) The use of personality measures in cross-cultural research: A test of three personality scales across two countries. *Journal of Applied Psychology*, **84**, 670–689.

Goldberg, L.R., Sweeney, D., Merenda, P.F. and Hughes, J.E., Jr. (1998) Demographic variables and personality: The effects of gender, age, education, and ethnic/racial status on self-descriptions of personality attributes. *Journal of Individual Differences*, **24**, 393–403.

Goldstein, H., Riley, Y. and Yusko, K.P. (1999) Exploration of black-white subgroup differences on interpersonal constructs. In B. Smith (Chair), Subgroup Differences in Employment Testing. Symposium conducted at the 14th annual conference of the Society for Industrial and Organizational Psychology, Atlanta.

Goleman, D. (1995) *Emotional Intelligence*. New York: Bantam Books.

Gottfredson, L.S. (1998) The general intelligence factor. *Scientific American Presents*, **9**, 24–29.

Greenfield, P.M. (1997) You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist*, **52**, 1115–1124.

Guilford, J.P. (1967) *The Nature of Human Intelligence*. New York: McGraw-Hill.

Guion, R.M. (1998) *Assessment, Measurement, and Prediction for Personnel Decisions*. Mahwah, NJ: Elrbaum.

Gustafsson, J. (1999) Measuring and understanding g: Experimental and correlational approaches. In P.L. Ackerman, P.C. Kyllonen and R.D. Roberts (eds.), *Learning and Individual Differences*. Washington, DC: American Psychological Association.

Hackman, J.R. (1987) The design of work teams. In J. Lorsch (ed.), *Handbook of Organizational Behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Hanges, P.J., Grojean, M.W. and Smith, D.B. (2000) Bounding the concept of test banding: Reaffirming the traditional approach. *Human Performance*, **13**, 181–198.

Hartigan, J.A. and Wigdor, A.K. (eds.) (1989) *Fairness in Employment Testing: Validity Generalization, Minority Issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Hattrup, K., Rock, J. and Scalia, C. (1997) The effects of varying conceptualizations of job performance on adverse impact, minority hiring, and predicted performance. *Journal of Applied Psychology*, **82**, 656–664.

Hauser, R.M. (1998) Trends in black-white test-score differentials: I. Uses and misuses of NAEP/SAT data. In U. Neisser (ed.), *The Rising Curve: Long-term Gains in IQ and Related Measures*. Washington, DC: American Psychological Association.

Hedges, L.V. and Nowell, A. (1995) Sex differences in mental test scores, variability, and number of high-scoring individuals. *Science*, **269**, 41–45.

Helms, J.E. (1992) Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist*, **47**, 1083–1101.

Herrnstein, R.J. and Murray, C. (1994) *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Hoffman, C.C. and Thornton, G.C., III. (1997) Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology*, **50**, 455–470.

Hofstede, G. (1980) *Culture's Consequences: International Differences in Work-Related Values*. Beverly Hills, CA: Sage.

Hogan, J.C. (1991a) Physical abilities. In M.D. Dunnette and L.M. Hough (eds.), *Handbook of Industrial and Organizational Psychology* (2nd edn., Vol. 2). Palo Alto, CA: Consulting Psychologists Press.

Hogan, J.C. (1991b) Structure of physical performance in occupational tasks. *Journal of Applied Psychology*, **76**, 495–507.

Hogan, J., Hogan, R. and Bush, C.M. (1984) How to measure service orientation. *Journal of Applied Psychology*, **69**, 167–173.

Holden, L.M. (1996) The effectiveness of training in improving test performance. Paper presented at the 11th annual conference for the Society of Industrial and Organizational Psychology, San Diego, CA.

Holland, P.W. and Wainer, H. (eds.) (1993*) Differential Item Functioning: Theory and Practice*. Hillsdale, NJ: Erlbaum.

Horn, J.L. (1968) Organization of abilities and the development of intelligence. *Psychological Review*, **75**, 242–259.

Horn, J.L. (1970) Organization of data on life-span development of human abilities. In L.R. Goulet and P.B. Baltes (eds.), *Life-Span Developmental Psychology*. New York: Academic.

Horn, J.L. (1975) Psychometric studies of aging and intelligence. In S. Gershon and A. Raskin (eds.), *Genesis and Treatment of Psychological Disorders in the Elderly* (Vol. 2). New York: Raven.

Horn, J.L. (1976) Human abilities: A review of research and theory in the early 1970s. *Annual Review of Psychology*, **27**, 437–485.

Horn, J.L. (1982) The theory of fluid and crystallized intelligence in relation to concepts of cognitive psychology and aging in adulthood. In F.I.M. Craik and S. Trehub (eds.), *Aging and Cognitive Processes*. New York: Plenum Press.

Horn, J.L. (1989) Models of intelligence. In R.L. Linn (ed.), *Intelligence: Measurement Theory and Public Policy*. Urbana: University of Illinois Press.

Horn, J.L. and Cattell, R.B. (1967) Age differences in fluid and crystallized intelligence. *Acta Psychologica*, **26**, 107–129.

Horn, J.L. and Donaldson, G. (1976) On the myth of intellectual decline in adulthood. *American Psychologist*, **31**, 701–719.

Horn, J.L. and Hofer, S.M. (1992) Major abilities and development in the adult period. In R.J. Sternberg and C.A. Berg (eds.), *Intellectual Development*. New York: Cambridge University Press.

Hornick, C.W. and Axton, T.R. (1998) Weighing issues: Balancing low adverse impact and high validity. Paper presented at the 106th annual convention of the American Psychological Association, San Francisco, CA.

Hough, L.M. (1984) Development and evaluation of the 'accomplishment record' method of selecting and promoting professionals. *Journal of Applied Psychology*, **69**, 135–146.

Hough, L.M. (1992) The 'big five' personality variables – construct confusion: Description versus prediction. *Human Performance*, **5**, 139–155.

Hough, L.M. (1997) The millennium for personality psychology: New horizons or good old daze. *Applied Psychology: An International Review*, **47**, 233–261.

Hough, L.M. (1998) Personality at work: Issues and evidence. In M.D. Hakel (ed.), *Beyond Multiple Choice: Evaluating Alternatives to Traditional Testing for Selection*. Hillsdale, NJ: Erlbaum.

Hough, L.M., Carter, G.W., Dohm, T.E., Nelson, L.C. and Dunnette, M. (1993) *Development and Validation of the 'Universal Test Battery': A Computerized Selection System for Non-Management Employees* (Institute Report No. 219). Minneapolis: Personnel Decisions Research Institutes, Inc.

Hough, L.M., Keyes, M.A. and Dunnette, M.D. (1984) *Development and Validation of Personnel Selection Systems for Eight Library of Congress Professional Series* (Institute Report No. 91). Minneapolis, MN: Personnel Decisions Research Institute.

Hough, L.M. and Schneider, R.J. (1996) Personality traits, taxonomies, and applications in organizations. In K.R. Murphy (ed.), *Individual Differences and Behavior in Organizations*. San Francisco, CA: Jossey-Bass.

Houston, J.S. and Schneider, R.J. (1996) *Development and Validation of a Selection System for Insurance Agents* (TR#292). Minneapolis, MN: Personnel Decisions Research Institutes, Inc.

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

Huffcutt, A.I., Roth, P.L. and McDaniel, M.A. (1996) A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology*, **81**, 459–473.

Hunter, J.E. and Hunter, R.F. (1984) Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, **96**, 72–98.

Hunter, J.E. and Schmidt, F.L. (1976) Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin*, **83**, 1053–1071.

Hyde, J.S. (1981) How large are cognitive gender differences? *American Psychologist*, **36**, 892–901.

Hyde, J.S., Fennema, E. and Lamon, S.J. (1990) Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, **107**, 139–155.

Hyde, J.S. and Linn, M.C. (1988) Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, **104**, 53–69.

Hyde, J.S. and McKinley, N.M. (1997) Gender differences in cognition: Results from meta-analyses. In P.J. Caplan, M. Crawford, J.S. Hyde and J.T.E. Richardson (eds.), *Gender Differences in Human Cognition*. New York: Oxford University Press.

Jensen, A.R. (1969) How much can we boost IQ and educational achievement? *Harvard Educational Review*, **39**, 1–123.

Jensen, A.R. (1972) *Genetics and Education.* New York: Harper and Row.

Jensen, A.R. (1980) *Bias in Mental Testing.* New York: Free Press.

Jensen, A.R. (1986) *g*: Artifact or reality. *Journal of Vocational Behavior*, **29**, 301–331.

Jensen, A.R. (1998) *The g Factor*. Westport, CN: Praeger.

Jensen, A.R. and McGurk, F.C.J. (1987) Black-white bias in 'cultural' and 'noncultural' test items. *Personality and Individual Differences*, **8**, 295–301.

Johnson, J.W. (1996) Linking employee perceptions of service climate to customer satisfaction. *Personnel Psychology*, **49**, 831–851.

Johnson, J.W. (2000) A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, **35**, 1–19.

Johnson, J.W., Carter, G.W., Davison, H.K. and Oliver, D. (in press) A synthetic validity approach to testing differential prediction hypotheses. *Journal of Applied Psychology*.

Jones, R.G., Stevens, M.J. and Fischer, D.L. (2000) Selection in team contexts. In J.F. Kehoe (ed.), *Managing Selection in Changing Organizations*. San Francisco: Jossey-Bass.

Jöreskog, K.G. and Yang, F. (1996) Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G.A. Marcoulides and R.E. Schumacker (eds.), *Advanced Structural Equation Modeling: Issues and Techniques*. Mahwah, NJ: Erlbaum.

Klimoski, R. and Jones, R.G. (1995) Staffing for effective group decision making: Key issues in matching people to teams. In R. Guzzo, E. Salas and Associates (eds.), *Team Effectiveness and Decision Making in Organizations.* San Francisco: Jossey-Bass.

Kroeber, A.L. and Kluckhohn, F. (1952) *Culture: A Critical Review of Concepts and Definitions.* Cambridge, MA: Harvard University Press.

Kulik, J.A., Bangert-Drowns, R.L. and Kulik, C.C. (1984) Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, **95**, 179–188.

Kulik, J.A., Kulik, C.C. and Bangert, R.L. (1984) Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, **21**, 435–447.

Lautenschlager, G.J. and Mendoza, J.L. (1986) A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement*, **10**, 133–139.

Law, D.J., Pellegrino, J.W. and Hunt, E.B. (1993) Comparing the tortoise and the hare: Gender differences and experience in dynamic spatial reasoning tasks. *Psychological Science*, **4**, 35–40.

Levy, B. (1996) Improving memory in old age through implicit self-stereotyping. *Journal of Personality and Social Psychology*, **71**, 1092–1107.

Li, F., Harmer, P., Duncan, T.E., Duncan, S.C., Acock, A. and Boles, S. (1998) Approaches to testing interaction effects using structural equation modeling methodology. *Multivariate Behavioral Research*, **33**, 1–39.

Lievens, F., Coestsier, P. and Decaesteker, C. (in press) Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment.*

Linn, M.C. and Petersen, A.C. (1985) Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, **56**, 1479–1498.

Loehlin, J.C., Lindzey, G. and Spuhler, J.N. (1975) *Race Differences in Intelligence*. New York: Freeman.

Loevinger, J. (1994) In search of grand theory. *Psychological Inquiry*, **5**, 142–144.

Lounsbury, J.W., Bobrow, W. and Jensen, J.B. (1989) Attitudes toward employment testing: Scale development, correlates, and 'known group' validation. *Professional Psychology: Research and Practice*, **20**, 340–349.

Lubinski, D. (2000) Scientific and social significance of assessing individual differences: 'Sinking shafts at a few critical points'. *Annual Review of Psychology*, **51**, 405–444.

Lubinski, D. and Dawis, R.V. (1992) Attitudes, skills, and proficiencies. In M.D. Dunnette and L.M. Hough (eds.), *Handbook of Industrial and Organizational Psychology* (Vol. 3). Palo Alto, CA: Consulting Psychologists Press.

Lynn, R. (1987) The intelligence of the mongoloids: A psychometric, evolutionary and neurological theory. *Personality and Individual Differences*, **8**, 813–844.

Lynn, R. (1991) Race differences in intelligence: A global perspective. *Mankind Quarterly*, **31**, 254–296.

Lytle, A.L., Brett, J.M., Barsness, Z.I., Tinsley, C.H. and Janssens, M. (1995) A paradigm for confirmatory cross-cultural research in organizational behavior. *Research in Organizational Behavior*, **17**, 167–214.

Maccoby, E.E. and Jacklin, C.N. (1974) *The Psychology of Sex Differences.* Stanford, CA: Stanford University Press.

MacLane, C.N. (1996) Relationship of biodata validities and social demands of jobs. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

MacLane, C.N., Sharpe, J.P. and Nickels, B.J. (2000) An examination of stereotype threat theory in an applied setting. Paper presented at the 15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Markus, H.R. and Kitayama, S. (1991) Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, **98**, 224–253.

Martin, M. (1978) Memory span as a measure of individual differences in memory capacity. *Memory and Cognition*, **6**, 194–198.

Masters, M.S. and Sanders, B. (1993) Is the gender difference in mental rotation disappearing? *Behavior Genetics*, **23**, 337–341.

Maurer, T., Solamon, J. and Troxtel, D. (1998) Relationship of coaching with performance in situational interviews. *Journal of Applied Psychology*, **83**, 128–136.

Mayer, J.D. and Geher, G. (1996) Emotional intelligence and the identification of emotion. *Intelligence*, **22**, 89–113.

Maxwell, S.E. and Arvey, R.D. (1993) The search for predictors with high validity and low adverse impact: Compatible or incompatible goals? *Journal of Applied Psychology*, **78**, 433–437.

Mayer, J.D. and Geher, G. (1996) Emotional intelligence and the identification of emotion. *Intelligence*, **22**, 89–113.

McAdams, D.P. (1992) The five-factor model in personality: A critical appraisal. *Journal of Personality*, **60**, 329–361.

McCrae, R.R. and Costa, P.T, Jr. (1996) Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J.S. Wiggins (ed.), *The Five-Factor Model of Personality: Theoretical Perspectives*. New York: The Guilford Press.

McCrae, R.R. and Costa, P.T, Jr. (1997) Personality trait structure as a human universal. *American Psychologist*, **52**, 509–516.

McHenry, J.J., Hough, L.M., Toquam, J.L., Hanson, M.A. and Ashworth, S. (1990) Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology*, **43**, 335–354.

Meehl, P.E. (1997) The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L.L. Harlow, S.A. Mulaik and J.H. Steiger (eds.), *What If There Were No Significance Tests?* Mahwah, NJ: Erlbaum.

Millsap, R.E. (1995) Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, **30**, 577–605.

Millsap, R.E. (1997) Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, **2**, 248–260.

Millsap, R.E. and Meredith, W. (1992) Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement*, **16**, 389–402.

Minton, H.L. and Schneider, F.W. (1980) *Differential Psychology*. Monterey, CA: Brooks/Cole Publishing Co.

Mone, M.A., Mueller, G.C. and Mauland, W. (1996) The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology*, **49**, 103–120.

Morris, S.B. and Lobsenz, R.E. (2000) Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology*, **53**, 89–111.

Mossholder, K.W., Kemery, E.R. and Bedeian, A.G. (1990) On using regression coefficients to interpret moderator effects. *Educational and Psychological Measurement*, **50**, 255–263.

Motowidlo, S.J., Dunnette, M.D. and Carter, G.W. (1990) An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, **75**, 640–647.

Motowidlo, S.J. and Tippins, N. (1993) Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology*, **66**, 337–344.

Motowidlo, S.J. and Van Scotter, J.R. (1994) Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology*, **79**, 475–480.

Mount, M.K., Barrick, M.R. and Stewart, G.L. (1998) Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance*, **11**, 145–165.

Mullis, I.V., Dossey, J.A., Foertsch, M.A., Jones, L.R. and Gentile, C.A. (1991) *Trends in Academic Progress: Achievement of U.S. Students in Science, 1969–70 to 1900; Mathematics, 1973 to 1990; Reading, 1971 to 1990; and Writing, 1984 to 1990* (National Center for Education Statistics, Office of Educational Research and Improvement, US Department of Education). Washington, DC: US Government Printing Office.

Murphy, K.R. (1996) Individual differences and behavior in organizations: Much more than *g*. In K.R. Murphy (ed.), *Individual Differences and Behavior in Organizations*. San Francisco, CA: Jossey-Bass.

Murphy, K.R. and Shiarella, A.H. (1997) Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology*, **50**, 823–854.

Neisser, U. (ed.) (1998) *The Rising Curve: Long-Term Gains in IQ and Related Measures*. Washington, DC: American Psychological Association.

Neisser, U., Boodoo, G., Bouchard, T.J., Jr., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J. and Urbina, S. (1996) Intelligence: Knowns and unknowns. *American Psychologist*, **51**, 77–101.

Neuman, G.A. and Wright, J. (1999) Team effectiveness: Beyond skills and cognitive ability. *Journal of Applied Psychology*, **84**, 376–389.

Norborg, J.M. (1984) A warning regarding the simplified approach to the evaluation of test fairness in employee selection procedures. *Personnel Psychology*, **37**, 483–486.

Nyfield, G. and Baron, H. (2000) Cultural context in adapting selection practices across borders. In J.F. Kehoe (ed.), *Managing Selection in Changing Organizations*. San Francisco: Jossey-Bass.

O'Neill, K.A. and McPeek, W.M. (1993) Item and test characteristics that are associated with differential item functioning. In P.W. Holland and H. Wainer (eds.), *Differential Item Functioning*. New Jersey: Erlbaum.

Ones, D.S., Hough, L.M. and Viswesvaran, C. (1998) Personality correlates of managerial performance constructs. In R.C. Page (Chair), Personality Determinants of Managerial Potential, Performance, Progression and Ascendancy. Symposium at 13th annual meeting of Society for Industrial and Organizational Psychology, Dallas.

Ones, D.S. and Viswesvaran, C. (1998a) Gender, age, and race differences on overt integrity tests: Results across four large-scale job applicant data sets. *Journal of Applied Psychology*, **83**, 35–42.

Ones, D.S. and Viswesvaran, C. (1998b) Integrity testing in organizations. In R.W. Griffin, A.O'Leary-Kelly and J.M. Collins (eds.), *Dysfunctional Behavior in Organizations: Vol. 2, Nonviolent Behaviors in Organizations*. Greenwich, CT: JAI.

Organ, D.W. and Ryan, K. (1995) A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel Psychology*, **48**, 775–802.

Oswald, F.L., Saad, S.A. and Sackett, P.R. (2000) The homogeneity of error variances assumption in differential prediction analysis: Does it really matter? *Journal of Applied Psychology*, **85**, 536–541.

Parasuraman, A., Berry, L.L. and Zeithaml, V.A. (1991) Refinement and reassessment of the SERVQUAL scale. *Journal of Retailing*, **69**, 140–147.

Pedhazur, E.J. (1997) *Multiple Regression in Behavioral Research: Explanation and Prediction* (3rd edn.). Fort Worth, TX: Harcourt Brace College Publishers.

Pervin, L.A. (1994) A critical analysis of current trait theory. *Psychological Inquiry*, **5**, 103–113.

Petersen, N.S. and Novick, M.R. (1976) An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, **13**, 3–29.

Ployhart, R.E. and Ehrhart, M. (2000) Modeling the Practical Effects of Applicant Reactions. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Ployhart, R.E., Ryan, A.M., Conley, P.R. and West, B.J. (1999) Effects of Test Preparation Programs on Applicants' Perceived Predictive Validity and Self-Efficacy. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Powers, D.E. (1986) Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin*, **100**, 67–77.

Powers, D.E. (1993) Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice*, **12**, 24–30.

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

Pulakos, E.D., Schmitt, N. and Chan, D. (1996) Models of job performance ratings: An examination of ratee race, ratee gender, and rater level effects. *Human Performance*, **9**, 103–119.

Raju, N.S, van der Linden, W.J. and Fleer, P.F. (1995) IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, **19**, 353–368.

Ree, M.J. and Carretta, T.R. (1994) Factor analysis of the ASVAB: Confirming a Vernon-like structure. *Educational and Psychological Measurement*, **54**, 459–463.

Ree, M.J. and Earles, J.A. (1991) Predicting training success: Not much more than *g*. *Personnel Psychology*, **44**, 321–332.

Reese, H.W. (1976) The development of memory: Life-span perspectives. In H.W. Reese (ed.), *Advances in Child Development and Behavior* (Vol. 2). New York: Academic Press.

Reynolds, C.R., Chastain, R.L., Kaufman, A.S. and McLean, J.E. (1987) Demographic characteristics and IQ among adults: Analysis of the WAIS-R standardization sample as a function of the stratification variables. *Journal of School Psychology*, **25**, 323–342.

Roberts, R.D., Pallier, G. and Goff, G.N. (1999) Sensory processes within the structure of human cognitive abilities. In P.L. Ackerman, P.C. Kyllonen and R.D. Roberts (eds.), *Learning and Individual Differences: Process, Trait, and Content Determinants*. Washington, DC: American Psychological Association.

Rosse, J.G., Miller, H.E. and Barnes, L.K. (1991) Combining personality and cognitive ability predictors for hiring service-oriented employees. *Journal of Business and Psychology*, **5**, 431–445.

Roth, P.L., BeVier, C.A., Bobko, P., Switzer, F.S., III, Tyler, P. (in press) Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis.

Roth, P.L. and Bobko, P. (2000) College grade point average as a personnel selection devise: Ethnic group differences and potential adverse impact. *Journal of Applied Psychology*, **85**, 399–406.

Roznowski, M., Dickter, D.N., Hong, S., Sawin, L.L. and Shute, V.J. (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? *Journal of Applied Psychology*, **85**, 940–955.

Ruch, F.L. and Ruch, W.W. (1963) *Employee Aptitude Survey Technical Report.* Los Angeles, Psychological Services, Inc.

Ryan, A.M. and McFarland, L.A. (1997) Organizational influences on applicant withdrawal from selection processes. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Ryan, A.M., McFarland, L.A., Sacco, J.M. and Kriska, S.D. (2000) Applicant self-selection: Correlates of withdrawal from a multiple hurdle process. *Journal of Applied Psychology*, **85**, 163–179.

Ryan, A.M. and Ployhart, R.E. (2000) Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management*, **26**, 565–606.

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

Ryan, A.M., Ployhart, R.E., Greguras, G.J. and Schmit, M.J. (1997) Predicting applicant withdrawal from applicant attitudes. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Ryan, A.M., Ployhart, R.E., Greguras, G.J. and Schmit, M.J. (1998) Test preparation programs in selection contexts: Self-selection and program effectiveness. *Personnel Psychology*, **51**, 599–621.

Ryer, J.A., Schmidt, D.B. and Schmitt, N. (1999) Candidate orientation programs: effects on test scores and adverse impact. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Sackett, P.R., Burris, L.R. and Ryan, A.M. (1989) Coaching and practice effects in personnel selection. In C.L. Cooper and I.T. Robertson (eds.), *International Review of Industrial and Organizational Psychology.* Chichester: John Wiley and Sons.

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

Sackett, P.R., Schmitt, N., Ellingson, J.E. and Kabin, M.B. (2001) High-stakes testing in employment, credentialing, and higher education: Prospects in a post affirmative-action world. *American Psychologist*, **56**, 302–318.

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

Sager, C.E., Peterson, N.G., Oppler, S.H., Rosse, R.L. and Walker, C.B. (1997) An examination of five indexes of test battery performance: Analysis of the ECAT battery. *Military Psychology*, **9**, 97–120.

Salas, E., Dickinson, T.L., Converse, S.A. and Tannenbaum, S.I. (1992) Toward an understanding of team performance and training. In R.W. Swezey and E. Salas (eds.), *Teams: Their Training and Performance*. Norwood, NJ: Ablex Publishing Company.

Salthouse, T.A. (1991) Mediation of adult age differences in cognition by reduction in working memory and speed of processing. *Psychological Science*, **2**, 179–183.

Saucier, G and Goldberg, L.R. (1998) What is beyond the Big Five? *Journal of Personality*, **66**, 495–524.

Scarr, S. (1981) *Race, Social Class, and Individual Differences in I.Q.* Hillsdale, NJ: Erlbaum.

Schaie, K.W. and Strother, C.R. (1968) A cross-sequential study of age changes in cognitive behavior. *Psychological Bulletin*, **68**, 671–680.

Scheuneman, J. (1987) An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement*, **24**, 97–118.

Scheuneman, J. and Gerritz, K. (1990) Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, **27**, 109–131.

Schmidt, F.L. (1988) The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior*, **33**, 272–292.

Schmidt, F.L., Greenthal, A.L., Hunter, J.E., Berner, J.G. and Seaton, F.W. (1977) Job sample vs. paper-and-pencil trades and technical tests: Adverse impact and examinee attitudes. *Personnel Psychology*, **30**, 187–197.

Schmidt, F.L. and Hunter, J.E. (1984) A within setting empirical test of the situational specificity hypothesis in personnel selection. *Personnel Psychology*, **37**, 317–326.

Schmidt, F.L. and Hunter, J.E. (1998) The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, **124**, 262–274.

Schmidt, F.L. and Kaplan, L.B. (1971) Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology*, **24**, 419–434.

Schmit, M.J. (1994) Pre-employment processes and outcomes, applicant belief systems, and minority-majority group differences. Unpublished doctoral dissertation, Bowling Green, OH: Bowling Green State University.

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

Schmit, M.J. and Ryan, A.M. (1992) Test-taking dispositions: A missing link? *Journal of Applied Psychology*, **77**, 629–637.

Schmit, M.J. and Ryan, A.M. (1997) Applicant withdrawal: The role of test-taking attitudes and racial differences. *Personnel Psychology*, **50**, 855–876.

Schmitt, A.P. and Dorans, N.J. (1990) Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, **27**, 67–81.

Schmitt, A.P., Holland, P.W. and Dorans, N.J. (1993) Evaluating hypotheses about differential item functioning. In P.W. Holland and H. Wainer (eds.), *Differential Item Functioning*. Mahwah, NJ: Erlbaum.

Schmitt, N., Clause, C.S. and Pulakos, E.D. (1996) Subgroup differences associated with different measures of some common job-relevant constructs. *International Review of Industrial and Organizational Psychology*, **11**, 115–139.

Schmitt, N. and Mills, A.E. (in press) Traditional tests and job simulations: Minority and majority performance and test validities *Journal of Applied Psychology.*

Schmitt, N. and Pulakos, E.D. (1998) Biodata and differential prediction: Some reservations. In M.D. Hakel (ed.), *Beyond Multiple Choice: Evaluating Alternatives to Traditional Testing for Selection*. Mahwah, NJ: Erlbaum.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L. and Jennings, D. (1997) Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology*, **82**, 719–730.

Schneider, B. and Bowen, D.E. (1992) Personnel/Human resources management in the service sector. In G.R. Ferris and K.M. Rowland (eds.), *Research in Personnel and Human Resources Management*, **10**, 1–30.

Schneider, R.J., Ackerman, P.L. and Kanfer, R. (1996) To 'act wisely in human relations': Exploring the dimensions of social competence. *Personality and Individual Differences*, **21**, 469–481.

Schneider, R.J. and Hough, L.M. (1995) Personality and industrial/organizational psychology. In C.L. Cooper and I.T. Robertson (eds.), *International Review of Industrial and Organizational Psychology*. Chichester: Wiley.

Schwartz, S.H. (1994) Beyond individualism and collectivism: New cultural dimensions of values. In U. Kim, H.C. Triandis, C. Kagitcibasi, S.C. Choi and G. Yoon (eds.), *Individualism and Collectivism: Theory, Method, and Applications.* Thousand Oaks, CA: Sage.

Shih, M., Pittinsky, T.L. and Ambady, N. (1999) Stereotype susceptibility: identity salience and shifts in quantitative performance. *Psychological Science*, **10**, 80–83.

Sireci, S.G. and Geisinger, K.F. (1999) Equity issues in employment testing, *Test Interpretation and Diversity: Achieving Equity in Assessment.* Washington, DC: American Psychological Association.

Smiderle, D., Perry, B.A. and Cronshaw, S.F. (1994) Evaluation of video-based assessment in transit operator selection. *Journal of Business and Psychology*, **9**, 3–22.

Smith, L.L. and Reise, S.P. (1998) Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology*, **75**, 1350–1362.

Spearman, C. (1927) *The Abilities of Man*. New York: Macmillan.

Spencer, S.J., Steele, C.M. and Quinn, D.M. (1999) Stereotype threat and women's math performance. *Journal of Experimental and Social Psychology*, **35**, 4–28.

Steele, C.M. (1992) Race and the schooling of Black Americans. *The Atlantic Monthly*, April.

Steele, C.M. (1997) A threat in the air: How stereotypes shape intellectual ability and performance. *American Psychologist*, **52**, 613–629.

Steele, C.M. and Aronson, J. (1995) Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology*, **69**, 797–811.

Sternberg, R.J. (1985) *Beyond IQ: A Triarchic Theory of Human Intelligence.* Cambridge: Cambridge University Press.

Sternberg, R.J. and Detterman, D.K. (eds.) (1986) *What is Intelligence? Contemporary Viewpoints on its Nature and Definition*. Norwood, NJ: Ablex.

Stevens, M.J. and Campion, M.A. (1994) The knowledge, skill, and ability requirements for teamwork: Implications for human resource management. *Journal of Management*, **20**, 503–530.

Stone, E.F. and Hollenbeck, J.R. (1989) Clarifying some controversial issues surrounding statistical procedures for detecting moderator variables: Empirical evidence and related matters. *Journal of Applied Psychology*, **74**, 3–10.

Stone, J., Lynch, C.I., Sjomeling, M. and Darley, J.M. (1999) Stereotype threat effects on Black and White athletic performance. *Journal of Personality and Social Psychology*, **77**, 1213–1227.

Stone-Romero, E.F., Alliger, G.M. and Aguinis, H. (1994) Type II error problems in the use of moderated multiple regression for the detection of moderating effects of dichotomous variables. *Journal of Management*, **20**, 167–178.

Stone-Romero, E.F. and Anderson, L.E. (1994) Relative power of moderated multiple regression and the comparison of subgroup correlation coefficients for detecting moderating effects. *Journal of Applied Psychology*, **79**, 354–359.

Storfer, M.D. (1990) *Intelligence and Giftedness: The Contributions of Heredity and Early Environment*. San Francisco: Jossey-Bass.

Sue, S. (1999) Science, ethnicity and bias: Where have we gone wrong? *American Psychologist*, **12**, 1070–1077.

Teasdale, T.W. and Owen, D.R. (1989) Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence*, **13**, 255–262.

Tellegen, A. (1993) Folk concepts and psychological concepts of personality and personality disorder. *Psychological Inquiry*, **4**, 122–130.

te Nijenhuis, J. and van der Flier, H. (1997) Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal of Applied Psychology*, **82**, 675–687.

Terris, W. (1997) The traditional regression model for measuring test bias is incorrect and biased against minorities. *Journal of Business and Psychology*, **12**, 25–37.

Thorndike, E.L. (1920) Intelligence and its uses. *Harper's Monthly Magazine*, **140**, 227–235.

Thorndike, R.L. (1936) Factor analysis of social and abstract intelligence. *Journal of Educational Psychology*, **27**, 231–233.

Thorndike, R.L. (1971) Concepts of culture-fairness. *Journal of Educational Measurement*, **8**, 63–70.

Thurstone, L.L. (1938) *Primary Mental Abilities*. Chicago: University of Chicago Press.

Troll, L.E. (1975) *Early and Middle Adulthood*. Monterey, CA: Brooks/Cole Publishing Co.

Valencia, R.R. and Lopez, R. (1992) Assessment of racial and ethnic minority students: Problems and prospects. In M. Zeidner and R. Most (eds.), *Psychological Testing: An Inside View*. Palo Alto, CA: Consulting Psychologists Press.

Vandenberg, R.J. and Lance, C.E. (2000) A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, **3**, 4–69.

Verive, J.M. and McDaniel, M.A. (1996) Short-term memory tests in personnel selection: Low adverse impact and high validity. *Intelligence*, **23**, 15–32.

Vernon, P.A. and Jensen, A.R. (1984) Individual and group differences in intelligence and speed of information processing. *Personality and Individual Differences*, **5**, 411–423.

Vernon, P.E. (1950) *The Structure of Human Abilities*. New York: Wiley.

Viswesvaran, C., Ones, D.S. and Hough, L.M. (1998) Construct validity of managerial potential scales. In R. Page (Chair), Personality Determinants of Managerial Potential, Performance, Progression and Ascendancy. Symposium conducted at the 13th annual convention of the Society for Industrial and Organizational Psychology, Dallas, April.

Viswesvaran, C., Ones, D.S. and Schmidt, F.L. (1996) Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, **81**, 557–574.

Voyer, D., Voyer, S. and Bryden, M.P. (1995) Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, **117**, 250–270.

Wallace, S.R. (1965) Criteria for what? *American Psychologist*, **20**, 411–417.

Waller, N.G., Thompson, J.S. and Wenk, E. (2000) Using IRT to separate measurement bias from true group differences on homogeneous *and* heterogeneous scales: An illustration with the MMPI. *Psychological Methods*, **5**, 125–146.

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

Weekley, J.A. and Jones, C. (1999) Further studies in situational tests. *Personnel Psychology*, **52**, 679–700.

Welford, A.T. (1977) Motor performance. In J.E. Birren and K.W. Schaie (eds.), *Handbook of the*

Psychology of Aging. New York: Van Nostrand Reinhold.

Whitney, D.J. and Schmitt, N. (1997) Relationship between culture and responses to biodata employment items. *Journal of Applied Psychology*, **82**, 113–129.

Wiggins, J.S. and Trapnell, P.D. (1997) Personality structure: The return of the Big Five. In R. Hogan, J. Johnson and S. Briggs (eds.), *Handbook of Personality Psychology*. San Diego: Academic.

Wilk, S.L., Desmarais, L.B. and Sackett, P.R. (1995) Gravitation to jobs commensurate with ability: Longitudinal and cross-sectional tests. *Journal of Applied Psychology*, **80**, 79–85.

Williams, W.M. and Ceci, S.J., (1997) Are Americans becoming more or less alike? Trends in race, class and ability differences in intelligence. *American Psychologist*, **52**, 1226–1235.

Willingham, W.W. (1999) A systemic view of test fairness. In S.J. Messick (ed.), *Assessment in Higher Education: Issues of Access, Quality, Student Development, and Public Policy*. Mahwah, NJ: Erlbaum.

Wolfle, L.M. (1985) Postsecondary educational attainment among Whites and Blacks. *American Educational Research Journal*, **22**, 501–525.