

Differential Item Functioning by Sex and Race in the Hogan Personality Inventory

Richard Sheppard
Kyunghee Han
Stephen M. Colarelli
Guangdong Dai
Central Michigan University

Daniel W. King
National Center for Posttraumatic Stress Disorder and Boston University

The authors examined measurement bias in the Hogan Personality Inventory by investigating differential item functioning (DIF) across sex and two racial groups (Caucasian and Black). The sample consisted of 1,579 Caucasians (1,023 men, 556 women) and 523 Blacks (321 men, 202 women) who were applying for entry-level, unskilled jobs in factories. Although the group mean differences were trivial, more than a third of the items showed DIF by sex (38.4%) and by race (37.3%). A content analysis of potentially biased items indicated that the themes of items displaying DIF were slightly more cohesive for sex than for race. The authors discuss possible explanations for differing clustering tendencies of items displaying DIF and some practical and theoretical implications of DIF in the development and interpretation of personality inventories.

Keywords: differential item functioning; item bias; measurement bias; Hogan Personality Inventory; personality testing; employee selection

Research throughout the past decade has shown that some dimensions of normal personality structure are valid predictors of job outcomes (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991). Moreover, when used in concert with cognitive ability tests, several personality traits demonstrate incremental validity, providing additional predictive power beyond that of cognitive ability alone (Day & Silverman, 1989). Thus, personality testing in employee selection is likely to continue and expand. With the increased interest in and use of employment-oriented personality inventories, questions are emerging about bias. Yet, the possibility of test bias in employment-oriented personality inventories remains largely unexplored (Sackett & Wilk, 1994). This stands in sharp contrast to the extensive research on bias in ability tests (Jensen, 1980; Thissen, Steinberg, & Gerrard, 1986).

This contrast is probably due to the historical difference between the uses of ability and personality tests and mean differences between racial and ethnic subgroups on cognitive ability test scores (Hartigan & Wigdor, 1989). Ability tests have primarily served a gate-keeping function, regulating entry to higher education, professional careers, and some employment opportunities. Moreover, Blacks typically score about 1 standard deviation lower than Caucasians on cognitive ability tests. Because these differences often translate into differential college admissions and hiring rates, researchers and policy makers deem it important to examine whether test bias causes group differences. If group differences are due to test bias rather than true differences, differential hiring rates would be considered unfair. Test bias would provide a

We would like to thank Andrew Cox, Mark Deskovitz, Chris Joles, Nathan Weed, and Kevin Young for labeling the components. This research is based on the doctoral dissertation of Richard Sheppard. Correspondence concerning this article should be addressed to Kyunghee Han, Department of Psychology, Central Michigan University, Mt. Pleasant, MI 48859; e-mail: han1k@cmich.edu.

Assessment, Volume 13, No. 4, December 2006 442-453
DOI: 10.1177/1073191106289031
© 2006 Sage Publications

sound rationale for implementing score adjustment procedures in admissions and hiring.

The situation with personality inventories has been quite different. Until recently, personality inventories were used primarily in clinical settings and had little impact on personnel selection. They served a minimal gate-keeping function in schools, occupations, and business organizations. Moreover, ambiguity existed over the meaning of personality and little consensus existed on the appropriate taxonomy of normal personality traits. This theoretical ambiguity precluded rigorous mathematical analyses. Therefore, there were few compelling reasons to examine bias in personality inventories. However, the situation is changing. With the emergence of the Big Five and similar taxonomies, a consensus is emerging on the dimensions of the normal personality structure (Goldberg, 1990). The development of psychometrically sound personality inventories for normal populations allows the same level of psychometric rigor to be applied to personality tests as is typical of cognitive ability tests, and the use of personality tests has broadened to the point where they serve a gate-keeping function in many employment settings (Kaihla, 2003). Thus, it is now important and technically possible to assess potential bias in personality inventories.

Test bias reflects psychometric inequalities among subgroups and can take the form of relationship bias or measurement bias (Drasgow, 1984). Relationship bias is concerned with the association between a test score and an external criterion measure; measurement bias is concerned with the properties of test items. A test exhibits relationship bias when individuals from different subgroups have equal test scores but unequal probabilities of success on the criterion (Cleary, 1968). Comparing regression parameters (slope and intercept) is the typical method for assessing relationship bias (Bobko & Bartlett, 1978). Relationship bias can have implications for fairness in admissions and selection. For example, relationship bias can result in false negative decisions whereby those who could perform capably on the criterion measure have a lower probability of having a passing score on the predictor. With measurement bias, a test is considered biased if individuals from different subgroups who possess the same quantity of an underlying (latent) trait have unequal probabilities of obtaining the same test score. Measurement bias is typically assessed through procedures that estimate differential item functioning (DIF)—the extent to which people from different subgroups with equal amounts of a trait have unequal probabilities of responding correctly or, in the case of personality tests, responding in the keyed direction to an item (Drasgow & Hulin, 1990). Among its possible implications, measurement bias can affect relationship bias, mean differences between subgroups, and construct validity (Smith, 2002).

Two widely used techniques for assessing DIF are estimating item parameter differences and calculating item odds ratios (Lord, 1980; Mantel & Haenszel, 1959; Millsap & Everson, 1993).

The majority of studies of bias in personality measures have focused on inventories that assess psychopathology, such as the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & KcKinley, 1940; MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). Most of these studies examined relationship bias. They found that the predictive validity of MMPI-2 scales was not different across racial groups (Arbisi, Ben-Porath, & McNulty, 2002; McNulty, Graham, Ben-Porath, & Stein, 1997; Timbrook & Graham, 1994). Waller, Thompson, and Wenk (2000) examined measurement bias on MMPI items and scales. They found that 38% of the items on each scale were, on average, biased against Blacks or Caucasians but that item bias canceled out when aggregated at the scale score level. The authors concluded that the “presence of differential item functioning does not lead inexorably to differential *test* functioning” (p. 143, *italics added*). Even fewer studies have investigated bias in inventories for nonclinical populations. Using a military sample, Saad and Sackett (2002) examined relationship bias by sex in three personality domains—adjustment, achievement, and dependability. They found little evidence of sex differences in the slopes of the regression lines. However, they did find considerable evidence of intercept differences, and here, the personality measures, for the most part, overpredicted female performance when compared to a common regression line. We were unable to locate any studies examining measurement bias in inventories containing traits common in normal populations.

Therefore, the first purpose of this study is to examine the extent of measurement bias toward sex and race subgroups on an inventory that assesses dimensions of normal personality functioning. The second purpose of the current study is to investigate the themes of the biased items. It is important to go beyond identifying and eliminating DIF items and gain an understanding of differences in how people respond to items (Ellis, Becker, & Kimmel, 1993). Smith (2002) demonstrated that DIF analysis can be used effectively to explore personality constructs.

METHOD

Participants

A personnel selection program used by an automobile supplier in the Detroit, Michigan, metropolitan area provided the data for this study. This program involved

the selection of factory laborers for entrance into any of eight entry-level, unskilled factory jobs. Our sample consisted of 2,102 adults. These included 1,579 Caucasians and 523 Blacks; 1,344 were men and 758 were women. The mean ages of these subgroups were 32.8 ($SD = 8.6$), 29.8 ($SD = 7.4$), 29.7 ($SD = 7.5$), and 33.1 years ($SD = 9.0$), respectively.

Measures

We used the Hogan Personality Inventory (HPI; Hogan & Hogan, 1992). The HPI contains 206 items that are keyed true and false and 13 scales, 7 of which are primary. The primary scales are Adjustment, Ambition, Sociability, Likeability, Prudence, Intellectance, and School Success and consist of 182 items. There is no item overlap among the primary scales. Each primary scale consists of several homogeneous item composites; the HPI has a total of 44 homogeneous item composites. The HPI generally corresponds to the five-factor model of personality (Digman, 1990). The primary divergence from the five-factor model results from the bifurcation of the five-factor model's Extraversion scale into separate Sociability and Ambition scales and the bifurcation of the Openness to Experience scale into the Intellectance and School Success scales. In the HPI, Adjustment measures the degree to which a person appears calm and self-accepting. Ambition includes initiative, surgency, ambition, and impetuosity; Sociability means being sociable, exhibitionist, and expressive. Likeability measures the degree to which a person is perceptive, tactful, and socially sensitive. Prudence measures the degree to which a person seems conscientious, conforming, and dependable. Intellectance is the degree to which a person is perceived as bright, creative, and interested in intellectual and cultural matters; School Success measures the degree to which a person enjoys academic activities and values educational achievement for its own sake. The other six scales (termed Occupational Performance scales) are formed by combining and differentially weighting items from the primary scales (Hogan & Hogan, 1992).

Developed especially for use in personnel selection (Hogan, 1986), the HPI emphasizes constructs relevant to job, career, and occupational performance. The HPI scales correlate with various measures of job performance across a variety of jobs, ranging from $r = .15$ to $r = -.62$ (Hogan & Hogan, 1992).

Procedure

Applicants completed the HPI as part of a personnel selection procedure and voluntarily provided sex and race information at the time of testing by marking the appropriate

boxes on a scannable answer sheet. Following test administration, answer sheets were mailed to the test publisher, Hogan Assessment Systems, Inc., Tulsa, Oklahoma, which scored the tests and returned the results to researchers via mail. Scores for each applicant included seven primary scale scores, 44 homogeneous item composite scores, and individual item responses. At the time of the data collection, the original 307-item version (Hogan, 1986) of the HPI was used. However, at the time of data analysis, an updated, and psychometrically stronger, 206-item version (Hogan & Hogan, 1992) of the HPI replaced the earlier version. Therefore, this study uses data scored according to the latter version. Scores based on the earlier and longer version were converted to the 206-item version scores by the test publisher. Investigating the DIF on any test requires a scoring key as well as knowledge of which items constitute each scale. Unfortunately, the publisher of the HPI would not provide us with this information. Therefore, prior to conducting the analyses, it was necessary to develop a scoring key as well as to assign each item to the correct scale, which is explained in the appendix.

Analyses

There are several statistical methods for detecting measurement bias, such as the Mantel-Haenszel procedure, logistic regression models, and item response theory approaches (Millsap & Everson, 1993). We used the Mantel-Haenszel chi-square ($MH-\chi^2$) method to detect DIF (Mantel & Haenszel, 1959) because it has been widely used and a focus of research (e.g., Azocar, Arean, Miranda, & Munoz, 2001; Roznowski & Reith, 1999; Scheuneman & Gerritz, 1990; Schmitt, Hattrup, & Landis, 1993). This method asks: Given two groups of job applicants matched on the amount of an attribute, do the two groups differ significantly in the rate at which they endorse each item that measures that attribute? For the present study, we assumed equivalent amounts of attribute from identical scale scores. The MH tests the null hypothesis that the odds ratio is 1 for an item with no DIF. Significant deviations from 1 are typically associated with a significant $MH-\chi^2$ value. A significant $MH-\chi^2$ value, reflecting an association between the classification variable (e.g., sex or race) and rate of item endorsement, is taken as evidence of DIF for the studied item. We use a significance level of .01 for all statistical tests. Although we conducted the DIF analysis on the items that we could assign to scales (138 out of 182 items), we used the total scores of each scale as matching variables. The test publisher provided total scores, which were based on all possible items.

In each scale, we compared the reference group (men or Caucasians) with the focal group (women or Blacks) in two ways: (a) the percentage of items flagged with DIF for

TABLE 1
Means and Standard Deviations for Seven Primary Scales
by Sex and Race and Standardized Mean Differences

Scale	Sex				Race							
	Men		Women		Caucasians				Blacks			
	M	SD	M	SD	M	SD	t	d	M	SD	t	d
Intellectance	13.99	5.00	11.52	4.83	13.20	5.20	11.53***	.49	12.53	4.65	2.61**	.13
Sociability	12.16	4.65	10.57	4.47	11.63	4.81	8.02***	.34	11.27	4.22	1.53	.08
Prudence	21.74	4.26	23.12	3.75	22.08	4.25	-7.82***	-.33	22.64	3.79	-2.67**	-.13
Likeability	19.65	2.13	20.24	1.74	20.01	1.97	-6.82***	-.29	19.45	2.11	5.53***	.28
Ambition	23.03	4.82	21.90	4.99	22.19	5.14	5.38***	.23	24.01	4.02	-7.37***	-.37
Adjustment	27.51	5.64	27.75	5.95	27.65	5.91	-.97	-.04	27.44	5.45	.74	.04
School success	8.55	3.289	8.67	3.47	8.52	3.41	-.80	-.04	8.89	3.13	-2.24*	-.11

* $p < .05$. ** $p < .01$. *** $p < .001$.

reference and focal groups and (b) mean delta values. Delta values, which are equal to -2.35 times the natural logarithm of odds ratio (Holland & Thayer, 1988), were calculated to examine the degree of DIF. Although the odds ratio is an estimate of DIF effect size, it is asymmetrical. Its values range from 0 to $+\infty$, and this makes interpretation difficult. Therefore, we transformed the odds ratio to delta values that are symmetric around zero and range from $-\infty$ to $+\infty$ (Holland & Thayer, 1988). The delta scale is an inverse normal transformation of percentage correct to a linear scale with a mean of 13 and standard deviation of 4. Delta values that are negative indicate items that the reference group members endorsed more frequently on average than did focal group members, whereas positive delta values indicate the opposite. A value of 1.00 indicates that the focal group members were more likely to endorse the item by one delta point than reference group members, and a difference of one delta is roughly equal to a difference of 10 points in percentage endorsed between groups (Zieky, 1993). Delta values of the items of each scale were averaged for the focal and reference groups separately.

To identify meaningful clusters among potentially biased items, we submitted items flagged with DIF to a series of principal component analyses. One clinical psychology professor and four graduate students who were blind to the purpose of the study and to the sex and racial characteristics of the participants then independently reviewed the items and clusters and identified themes suggested by each cluster. We then reviewed those themes and determined the final labels for each cluster.

RESULTS

Table 1 presents the means, standard deviations, significance tests, and effect sizes (Cohen's d ; Cohen, 1988)

for the seven primary scales, separately for sex and race. The largest d ratio (.49) was found for the men ($M = 13.99$) and women ($M = 11.52$) comparison on the Intellectance scale. The Ambition scale showed the largest race difference ($d = -.37$), with Blacks scoring higher ($M = 24.01$) than Caucasians ($M = 22.19$). All of the d ratios, except for Intellectance, were less than .5, indicating that the sex or race differences were small, although most of those differences were statistically significant. The mean d ratios¹ (.14 for the sex and .01 for the race) indicate that sex differences are greater than race differences on the seven primary scales.

DIF by Sex

The results of the DIF analysis based on sex are presented in Table 2. The second column shows the number of items analyzed and the percentages of the total number of items identified in relation to the total number of items each scale has. Of a total of 182 seven primary scale items, we identified 138 items (75.8%). The School Success scale had the highest identification rate (13 out of 14 items, 92.9%), whereas the Prudence scale had the lowest identification rate (18 out of 31 items, 58.1%). Of a total of 138 identified items, 53 (38.4%) showed evidence of DIF. Twenty-eight items (20.3%) were potentially biased for men; 25 (18.1%) for women. The magnitude of the absolute mean delta values for men ($M = -1.37$) and women ($M = 1.26$) across all scales were considered marginally biased by the ETS convention.² Although the absolute mean delta value for men was slightly greater than that for women, it is equivalent to only 1.1 difference in endorsement percentages.

The Intellectance scale had the greatest number of items flagged with DIF. Of the 21 Intellectance items analyzed, 15 (71.4%) showed DIF. Ten of the items were

TABLE 2
Results of DIF Analyses for Sex and Race Comparisons

Scale	Number of Items Analyzed ^a	DIF Items ^b	Sex					Race			
			Number ^c of Items Endorsed More Frequently By		Mean Delta ^d		Number ^c of Items Endorsed More Frequently By		Mean Delta ^d		
			Men	Women	Men	Women	DIF Items ^b	Caucasian	Black	Caucasian	Black
Intellectance	21 (84%)	15 (71.4%)	10 (13)	5 (8)	-1.50 (-1.23)	1.43 (.98)	10 (47.6%)	6 (9)	4 (12)	-1.83 (-1.33)	1.37 (.62)
Ambition	22 (75.9%)	12 (54.5%)	6 (11)	6 (11)	-1.49 (-1.13)	1.38 (.91)	9 (40.9%)	5 (7)	4 (15)	-1.63 (-1.30)	1.39 (.73)
School success	13 (92.9%)	6 (46.2%)	2 (6)	4 (7)	-1.52 (-1.16)	1.31 (.97)	5 (38.5%)	3 (7)	2 (6)	-.97 (-.55)	1.50 (.81)
Adjustment	24 (64.9%)	9 (37.5%)	5 (13)	4 (11)	-1.23 (-.70)	.83 (.50)	9 (37.5%)	7 (12)	2 (12)	-1.31 (-1.02)	1.75 (.62)
Sociability	21 (87.5%)	7 (33.3%)	2 (10)	5 (11)	-1.01 (-.64)	1.28 (.75)	7 (33.3%)	5 (9)	2 (12)	-.93 (-.70)	1.07 (.60)
Prudence	18 (58.1%)	3 (16.7%)	3 (12)	0 (6)	-1.01 (-.63)	.84	9 (50.0%)	3 (11)	6 (7)	-1.61 (-1.06)	1.25 (1.12)
Likeability	19 (86.4%)	1 (5.3%)	0 (7)	1 (12)	(-.65)	1.28 (.58)	3 (15.8%)	1 (12)	2 (7)	-2.11 (-.95)	1.17 (.75)
Total	138 (75.8%)	53 (38.4%)	28 (72)	25 (66)	-1.37 (-.88)	1.26 (.77)	52 (37.7%)	30 (68)	22 (70)	-1.43 (-.99)	1.34 (.72)

a. Percentages are based on the total number of items identified in relation to the total number of items each scale has.

b. Percentages are based on total number of items identified in each scale.

c. Top values are the number of items endorsed more frequently by each sex or race, which reached a significance level (defined as DIF), whereas those inside the parentheses indicate the total number of items endorsed more frequently by each sex or race regardless of significance level.

d. Top values are the mean delta of items flagged with DIF, whereas those inside the parentheses are the mean delta of the total items. Positive values favor the focal group (women or Black), whereas negative values favor the reference group (men or Caucasian).

potentially biased for men and 5 for women, indicating that the items in the Intellectance scale were more likely to be endorsed by men when men and women are matched on the Intellectance scale score. The largest mean delta value (2.93) summed across sex also indicated the largest magnitude of potential bias for this scale. Four other scales, Ambition (54.5%), School Success (46.2%), Adjustment (37.5%), and Sociability (33.3%) also had numerous items flagged with DIF. The Prudence scale had relatively fewer potentially biased items; only 3 out of 18 items (16.7%) were flagged with DIF. Measurement equivalence across sex was found in the Likeability scale, which had only 1 (5.3%) potentially biased item. Potentially biased items of School Success and Intellectance scales for men reached the greatest degree of DIF, whereas potentially biased items of the Adjustment scale for women showed an ignorable degree of DIF (see Note 2).

DIF by Race

The MH analysis of DIF by race also was shown in Table 2. Of a total of 138 items contrasted by race, 52 (37.7%) showed DIF. The items of the HPI scales were potentially biased more for Caucasians than Blacks. A greater number showing DIF and larger absolute mean delta values occurred for Caucasians (30 items; $M = -1.43$) than for Blacks (22 items; $M = 1.34$), although the mean

delta difference is roughly equivalent to only 0.9% difference in endorsement rates.

The Prudence scale had the greatest proportion of potentially biased items (50%). Among 9 items flagged with DIF, 6 were in favor of Blacks and 3 were in favor of Caucasians. The Intellectance scale, however, showed the largest magnitude of potential bias, with a delta value of 3.20.³ Of 21 items, 10 (47.6%) were showing DIF; 6 were in favor of Caucasians, whereas 4 were in favor of Blacks. Ambition, School Success, Adjustment, and Sociability also had substantial numbers of items indicating potential bias, with 9 (40.9%), 5 (38.5%), 9 (37.5%), and 7 (33.3%) with DIF, respectively. Again, the Likeability scale had the smallest number of potentially biased items. Of the 19 Likeability items, only 3 (15.8%) were potentially biased. Two Likeability items were in favor of Blacks, whereas one was in favor of Caucasians. The proportion and magnitude of potentially biased items on some scales differed by race. In the Prudence scale, for example, more items were potentially biased in favor of Blacks (6 items) than Caucasians (3 items), but the absolute mean delta of those items was higher for Caucasians (-1.61) than for Blacks (1.25). The greatest magnitude of potential bias was found in Intellectance, Ambition, and Prudence scales for Caucasians and Adjustment scale for Blacks. Sociability had the weakest magnitude of potential bias for Caucasians.

TABLE 3
Themes Among Items Potentially Biased by Sex

Scale	Potentially Biased By	Themes	Number of Items	Mean Delta
Intellectance	Men	C1: Thrill-seeking/Risky hobbies or sports (scuba diving, skydiving, etc.)	4	-1.55
		C2: Scientific interests (using microscope, knowing stars twinkle)	6	-1.46
	Women	C3: Interests in reading or word games (solving puzzles, inventing stories)	3	1.60
		C4: Enjoyment of classical music (classical music)	2	1.18
Ambition	Men	C1: Leadership/Social dominance (being a leader, taking charge of group)	5	-1.10
		C2: Competitiveness (being a competitive person)	2	-2.47
	Women	C3: Life satisfaction and happiness with life (life being not meaningful [-])	5	1.38
School success	Men	C2: Mathematical ability (multiply quickly)	3	-.64
	Women	C1: Reading interests/enjoyment (reading books, liking reading)	3	1.37
Adjustment	Men	C1: Closeness to parents (being loved by parents)	2	-.98
		C2: Emotional stability/Invulnerability (being calm, physical complaints [-])	3	-1.40
	Women	C3: Self-criticism/Distressed alienation	4 ^a	.83
Sociability	Men	C1: Showmanship/Exhibitionism (showing off)	2	-1.01
	Women	C2: Extroversion/Social involvement (like parties)	5	1.28

NOTE: The dashes within the brackets indicate that a presented example item had a negative loading on its corresponding component.

a. Item 12 had a slightly higher loading on Emotional stability/Invulnerability (.47) than Self-Criticism/Distressed alienation (.44) but we decided to assign it to the latter based on the second criterion listed in the text.

Themes Among Items Flagged With DIF

To identify meaningful clusters among potentially biased items, we first independently submitted intercorrelation matrices of DIF items of each scale to principal components analyses (PCA) with varimax rotation.⁴ The scales containing small numbers of DIF, Likeability, and Prudence for sex comparisons and Likeability for race comparisons were excluded from PCA. A total of 11 PCAs were performed, using interpretability as a guide to determine the number of components to be extracted. We had to consider not only that each cluster of potentially biased items hang together in a meaningful way but also that the items in each cluster should be in the same direction in terms of potential bias. Therefore, we used the following decision rules to address these concerns. First, we allowed a component to be defined by only two items because we have a small number of DIF items in a certain scale (e.g., School Success) and wanted to avoid the item clusters that have the opposite direction of bias. Second, when an item loaded on more than one component (defined by less than .1 difference between the loadings), it was assigned to the component that contained the items whose bias direction is consistent with that particular item in the calculation of mean delta. Third, if an item loaded weakly (< .3) on all the components, it was eliminated in the naming of the components and in calculation of mean delta. Finally, if an item clustered with

the items that were biased in the opposite direction, we excluded it in calculating mean delta but kept it in the labeling of the components. Based on the rules listed above, we rearranged the component loading matrices and presented them to the raters who were asked to name the clusters. We then determined the final labels based on the consensus from the raters.

Table 3 displays themes among items potentially biased by sex for the following scales: Intellectance, Ambition, School Success, Adjustment, and Sociability. The component structures of the items flagged with DIF by sex were fairly stable in that the items differentially functioning in one direction tended to hang together in a meaningful way (i.e., high interpretability). Thrill-seeking and scientific interests were the dominant themes among the items potentially biased in favor of men on the Intellectance scale. The two themes for women were interests in reading and enjoyment of classical music. For the Ambition scale, the leadership and competitiveness were thematic among the items endorsed by men, whereas life satisfaction and happiness with life were reflected in the items endorsed by women. Mathematical ability was reflected in the items endorsed more by men in the School Success scale, whereas items reflecting reading interests predominated among women. For the Adjustment scale, men favored items about closeness to parents and emotional stability, whereas women responded more to items reflecting self-criticism and alienation. Unlike the

TABLE 4
Themes Among Items Potentially Biased by Race

<i>Scale</i>	<i>Potentially Biased By</i>	<i>Themes</i>	<i>Number of Items</i>	<i>Mean Delta</i>
Prudence ^a	Caucasians	C1: Concerns of opinions of others (caring what other people think of one's self)	2	-.97
	Blacks	C2: Impulse control (not doing things spur of moment)	3 ^b	1.07
		C3: Perfectionism/Conscientiousness (striving for perfection)	3	1.43
Intellectance	Caucasians	C1: Thrill-seeking/Risky hobbies or sports (scuba diving, mountain climbing, etc.)	5	-1.99
	Blacks	C2: Intellect/Intellectualism/Inquisitiveness (using a microscope, thinking of good ideas)	4 ^c	1.37
Ambition	Caucasians	C1: Life-direction/Positive attitude (life passing by [-], self-confidence)	5 ^c	-1.36
	Blacks	C2: Competitiveness/Social dominance/Leadership (taking orders than giving them [-], enjoy talking in front of people)	3	1.41
School success	Caucasians	C1: Mathematical abilities (math being easy)	2	-1.01
	Blacks	C2: School aptitude/Achievement (in the top of the class)	2 ^c	1.50
Adjustment	Caucasians	C1: Trust/Lacks of cynicism/Neuroticism (trust people, no regret)	4	-1.42
		C2: Lack of somatic complaints (no headaches, indigestion, changing body temperature)	3	-1.15
	Blacks	C3: Happy homelife/Good relations with one's parents (not resenting parents and not running away from home)	2	1.75
Sociability	Caucasians	C1: Attention-seeking/Showmanship (showing off)	2 ^c	-.97
		C2: Enjoyment of large groups/Extroversion (like to be in noisy crowds)	3 ^d	-.96

NOTE: The dashes within the brackets indicate that a presented example item had a negative loading on its corresponding component.

a. Item 204 was excluded in the naming and calculation of delta due to the fourth criterion listed in the text.

b. Although item 122 had a slightly higher loading on the Concerns of opinions of others component (.39) than the Impulse Control component (.33), we decided to assign it to the Impulse control component based on the second rule.

c. One item from each cluster (174 in Intellect; 94 in Life-Direction; 99 in School Aptitude; 98 in Attention-Seeking; 116 in Enjoyment of Large Group) was excluded in the calculation of delta due to the third criterion.

d. Item 14 was excluded in the naming and calculation of delta due to the fourth criterion.

other scales, themes of Adjustment scale were not well defined except the first theme (closeness to parents). Independent raters produced very diverse labels for the second and third components. For the Sociability scale, showing off was thematic among the items endorsed by men, whereas for women it was social involvement.

The two-item component, competitiveness, showed the greatest mean delta (-2.47), indicating that men endorse this dimension approximately 25% more frequently, on average, than did women when the Ambition total score was adjusted. The item with greatest DIF among all items was a competitiveness item (not enjoying games unless winning). It had an odds ratio of .27, indicating that men were likely to endorse this item 3.7 times (1/.27) more frequently than women. A moderate amount of potential bias was found in Intellectance components,

and potential bias was minimal in Mathematical Ability, Closeness to Parents, and Self-Criticism.

Table 4 presents themes among items potentially biased by race. The items flagged with DIF by race were relatively unstable. Several items clustered with the items where bias was in the opposite direction, whereas others had low loadings on all the components (see Table 4).

In response to the Prudence scale, Blacks described themselves as having impulse control and being perfectionists. Two DIF items in favor of Caucasians referred to concerns of opinion of others, and one DIF item was eliminated because it had very low loadings on all of the three components. For the Intellectance scale, thrill-seeking or risky hobbies/sports themes were potentially biased in favor of Caucasians. The intellectualism theme was potentially biased in favor of Blacks, although item 174

(curiosity about stars), which loaded on the intellectualism theme, was potentially biased in favor of Caucasians. For the Ambition scale, Caucasians endorsed more items measuring life direction or satisfaction, whereas Blacks more frequently endorsed items associated with competitiveness, social dominance, or leadership when total Ambition score was controlled. One life direction item was, however, in favor of Blacks. In the School Success scale, three items with DIF were related to school aptitude/achievement, two of which were in favor of Blacks and one of which was in favor of Caucasians. Two items with DIF for Caucasians related to math. In the Adjustment scale, two items that were associated with happy home-life/good relations with one's parents were potentially biased in favor of Blacks. Two common themes identified among DIF items in favor of Caucasians were lack of somatic complaints and trust. In the Sociability scale, attention-seeking/showmanship and extroversion themes were found, but the direction of potential bias was inconsistent in that two of the three items in each component were in favor of Caucasians and one was in favor of Blacks.

The thrill-seeking theme was found to have the greatest magnitude of potential bias. When the mean difference on Intellectance scale between racial groups was controlled, Caucasians were more likely to endorse (about 20%) the items associated with thrill-seeking or risky sports than were Blacks. Item 75 (mountain climbing) showed the greatest mean delta value (-3.01) of the 138 items. The next highest mean delta was found in the themes of happy homelife/good relations with parents (1.75), which was in favor of Blacks. The themes of concerns of opinion of others, attention-seeking, and enjoyment of large groups had very low magnitude of potential bias.

DISCUSSION

Although personality tests are increasingly used in personnel selection, few studies have examined bias in employment-oriented personality tests. Accordingly, the purposes of this study were to evaluate potential measurement bias by sex and race subgroups in an employment-oriented personality test and to investigate the themes of the potentially biased items. Eighty-one out of the 138 items that we tested in the HPI exhibited DIF. Thirty-eight percent of the total identified items appeared potentially biased by sex; of those, 53% were in favor of men. Thirty-eight percent of the items showed DIF by race; of these, 58% were in favor of Caucasians.⁵ With the exception of the Adjustment scale, DIF items for sex comparison produced much more conceptually coherent

and simpler component structures than those for race. Caution should be taken not to interpret these results as an overall rate of DIF items of the HPI because our analyses were based only on 76% of the total items that we could identify (138 out of 182 items).⁶

Results of other studies also suggest that about one third to one half of items in personality tests show potential DIF. Waller et al. (2000) found potential DIF by race (Black vs. Caucasian) on 38% of the items in the MMPI, and Dai, Han, Hu, and Colarelli (2006) found potential DIF by race/culture (Chinese vs. American) on 58% of the items in the conscientiousness scale of the Revised NEO Personality Inventory (NEO-PI-R). Assuming that these results are reasonably representative, they raise several questions for the development and use of personality inventories. With more than a third of the items exhibiting DIF, is measurement bias a serious problem for personality inventories? If so, what should be done about biased test items? Or is item bias, as some suggest, generally not a significant practical problem (Roznowski & Reith, 1999)? If so, what then is the value of DIF analyses?

One approach—and this is perhaps still the standard approach—is to view item bias as a serious problem and to recommend that personality inventories be analyzed for measurement bias and that biased items be replaced with unbiased items (Drasgow, 1984). From this perspective, measurement bias is considered a serious problem. It is a source of relationship bias, it can contribute to adverse impact by lowering mean test scores of minority groups (Sackett & Wilk, 1994), and it limits the generalizability of tests across subgroups (Schmit, Kihm, & Robie, 2000). Others suggest that moderately biased items are not a problem, especially when the direction of bias is unsystematic. Several studies show that item-level bias does not necessarily weaken measurement quality or the predictive validity of the overall test (Roznowski & Reith, 1999; Waller et al., 2000), and item bias does not necessarily lead to differential test functioning (Waller et al., 2000). Therefore, detecting and removing moderately biased items may have little practical value for improving test quality.⁷

We support this latter approach. Given that a fairly substantial percentage of items in both ability and personality test often exhibit DIF, that moderately based items tend not to affect test quality, and that we still do not have an adequate theory that predicts which items are likely to exhibit bias, there is little to no advantage, in most circumstances, to removing items that exhibit moderate bias.

DIF analysis is an important tool for test development. Tests are likely to be more valid and less biased when items that exhibit extreme levels of DIF are identified and eliminated. Our content analysis of the items showing

DIF illustrated this point well. When controlling for mean differences between sexes on the Intellectance scale, all items measuring thrill seeking and scientific interests were easier for men to endorse, whereas some of items measuring culture were easier for women to endorse (also see Smith, 2002). This result confirms sex differentiation in Intellectual pursuits but at the same time suggests a differential magnitude of DIF between the sexes. Men tend to engage more in thrill-seeking activities than women, but the HPI's operationalization of thrill-seeking as male-oriented activities (race-car driving, scuba diving, mountain climbing, etc.) may inflate the sex difference. This suggests that the thrill-seeking and culture components of the HPI could be improved by including more gender-neutral items or perhaps by developing separate thrill-seeking and culture subscales for men and women.

Another interesting finding in sex comparisons is that men express high Ambition levels through competitiveness and leadership/social dominance, whereas women express high Ambition levels through life satisfaction and happiness with life. Magnitudes of potential bias of leadership/social dominance and life satisfaction were not substantial. However, one particular competitiveness item, not enjoying games unless winning, showed the greatest magnitude (-3.07) of DIF between men and women out of all of the items we examined. Although men tend to be more competitive than women (Daly & Wilson, 1999; Eagly & Carli, 1981; Roy & Benenson, 2002), the HPI's operationalization of competitiveness in games may inflate the sex difference, suggesting that this item could be deleted or replaced with more gender-neutral items.

The themes of racial DIF items were somewhat unexpected and inconsistent. It was surprising to find that the theme of thrill-seeking/risky hobbies was again found to be potentially biased. Careful inspection of risk-taking items revealed that they consisted of activities (race-car driving, scuba diving, mountain climbing, etc.) in which people with social and economic advantages tend to engage. This again suggests that the thrill-seeking of the HPI could be improved by including more race-neutral as well as gender-neutral items.

Although we found DIF by race on six of the seven scales, the themes of the items showing DIF were generally inconsistent, making it difficult to draw inferences regarding the type of items that are biased against or for racial groups. One theme, which might be labeled cautiousness, however, was somewhat consistent. On several scales, items related to cautiousness were more likely to be potentially biased in favor of Blacks. For example, on the Intellectance scale, risk-taking was potentially biased in favor of Caucasians; on the Prudence scale, items related

to impulse control, perfectionism, and conscientiousness were potentially biased in favor of Blacks; and on the Sociability scale, attention-seeking and showing off were potentially biased in favor of Caucasians. If these results are generalizable, they suggest a hypothesis that may help explain the persistent economic disparity between the two groups: group differences in risk-taking. In a culture that values business enterprise, risk-taking is a useful trait for activities leading to the creation of wealth. An interesting research question would be to examine why the differences between Blacks and Caucasians on risk taking and prudence exist. The differences may, for example, be due to different historical opportunity structures, reward systems, or perhaps differences in the distribution of the sensation-seeking allele between groups that immigrated freely into a country compared to those that arrived involuntarily (Chen, Burton, Greenberger, & Dmitrieva, 1999; Whybrow, 2005).

The difference in the thematic consistency of the items showing DIF by sex versus by race suggests how personality inventories may be useful for understanding the validity of how we categorize people into groups and subgroups. Subgroup categories that have strong psychological or biological underpinnings may evidence stronger thematic consistency among items showing DIF, whereas subgroup categories with weaker biological or psychological foundations (e.g., categories based more on arbitrary social constructions) may show less thematic consistency among items with DIF. For example, we found little conceptual similarity among the HPI items that were potentially biased by race, a theoretically and genetically ambiguous category. Items potentially biased in favor of Caucasians had relatively little conceptual coherency (impulsivity, not autonomous, thrill seeking, and lack of somatic complaints). On the other hand, items potentially biased by sex—an unambiguous biological category—clustered together along conceptually similar lines. Items potentially biased in favor of men clustered around agentic content (competitiveness, leadership) and interests in science and math. These results suggest that there is more semantic coherency among personality trait descriptors within sex than racial subgroups. This is consistent with the literature on the evolution of sex differences in psychological traits (Geary, 1998) as well as with modern genetics. Most racial differences are genetically trivial and ambiguous, whereas sex differences are more clearly delineated (Cavalli-Sforza, 2000; Gorski, 1991). Moreover, biological sex is straightforward—an individual has either two X chromosomes or an X and a Y chromosome, making the person either a biological woman or man. Race, on the other hand, is more biologically ambiguous. For example, common classifications

of race (e.g., Negroid, Caucasian) do not have an isomorphic set of "racial" genes. At this point, the use of DIF in personality tests for developing hypotheses about the underpinnings of subgroup classifications is entirely speculative; however, our results are sufficiently intriguing to merit further research along these lines.

Limitations and Conclusion

Our study represents an important step in the effort to systematically evaluate DIF by sex and race in an employment-oriented personality inventory with a large sample. It was, however, limited in several respects. First, although the sample size in our study was large, the extent to which the results are generalizable is unknown. Applicants came from a circumscribed geographical area and were applying for unskilled, entry-level manufacturing positions. It may therefore be problematic to generalize these findings to the job applicants for other types of work, which presumably, could tend to attract different types of people. A second limitation is in regard to our HPI item assignments. There is a possibility that a few items were assigned to the wrong scales. However, the strong correlations between the scales derived by us with the publisher's scales indicate that we accurately assigned most items. The other issue has to do with low item identification rates for some scales. This may be due to ambiguity or nontransparency in content. For example, only 58% of Prudence scale items were identified. It is, therefore, unknown whether the percentages of bias among these unidentified items would be the same as the percentage of the assigned items and whether the magnitude of DIF is the same. The last limitation involves the assumption of unidimensionality for assessing DIF. Factor analysis of each of the seven primary scales showed that none of the seven primary scales were fully unidimensional, suggesting multidimensionality of the underlying attribute. Although some people believe that "violations of this assumption have not been found to adversely affect some DIF statistics" (Schmitt et al., 1993, p. 604), there is a possibility that lack of unidimensionality might distort the results.

In conclusion, this study suggests that there is a fairly substantial degree of measurement bias in an employment-oriented personality test, although the magnitude of bias was, for most items, relatively small. The content of the items displaying DIF by sex was much clearer and thematic than for items biased by race. Our approach to the identification of the themes among potentially biased items suggests a new approach for using DIF that could contribute in future studies to the development of theory in a confirmatory manner.

APPENDIX

Assignment of Items to Scales and Scoring the Hogan Personality Inventory (HPI)

We assigned test items to relevant Hogan Personality Inventory (HPI) primary scales using (a) information from the HPI manual, (b) intuitive judgment, (c) trained raters, and (d) factor analysis. The HPI manual (Hogan & Hogan, 1992) indicated the proper assignment of 41 items to their scales. Face validity of many items enabled assignment of items to their proper homogeneous item composites (HIC). Once HIC assignment was made, the scale assignment could be made because scale membership of HIC was provided in the manual. For example, the item "going to a party" was judged to belong to the homogeneous item composite Likes Parties, which belongs to the Sociability scale. Other items, which were not so transparent, were given to three to five raters—all advanced graduate students in psychology—who underwent a training session and then independently assigned the items to HIC. During training, raters were presented with the construct of each scale and the HIC belonging to each. Definitions of each HIC and examples of known items were provided. Then, each rater assigned each unassigned item to an HIC. Those items for which a high degree of agreement was reached by raters were assigned to the agreed-on scale.

We also factor analyzed the HPI items using the same extraction and rotational procedures (principal components analysis with varimax rotation) as did Hogan and Hogan (1992) in construction of the inventory. The component loadings of the items and the groups of items on components also were used as a guide to assign items to their scales. Where the results of our factor analysis supported the raters' assignment of items to scales, we were confident that correct item assignment was made.

Using the preliminary item assignments described above, we then determined the scoring direction (i.e., true or false) for each item based on our understanding of the constructs measured. A total of 138 out of 182 items (75.8%) were assigned to the seven primary scales and, thus, were available for item-level DIF analysis. The other 44 items were not analyzed due to uncertainty of item assignment to the correct scale. Items identified out of the total number (in parentheses) for the seven scales were 24 (37) for Adjustment, 22 (29) for Ambition, 21 (24) for Sociability, 19 (22) for Likeability, 18 (31) for Prudence, 21 (25) for Intellectance, and 13 (14) for School Success. To evaluate the accuracy of item assignment, we correlated two sets of seven primary scale scores—those that were newly derived by us and those

that were provided from the publisher—and compared internal consistency coefficients using our current sample ($N = 2,102$). The correlations between the two sets of scores were extremely high, with correlation coefficients of .96, .98, .96, .98, and .99 for Adjustment, Ambition, Sociability, Intellectance, and School Success, respectively. The correlations between two sets of scores were moderately high for the Prudence (.84) and Likeability (.80) scales. Despite the shortened length of the scales, internal consistency coefficients of the newly derived scales were fairly comparable with those of the original scales except for the Prudence and Likeability scales. The alpha coefficients of the shortened scales (.57 and .52) were noticeably smaller than that of the original scale (.78 and .71) (Hogan & Hogan, 1992). Plausible explanations include (a) not identifying critical items, (b) incorrect identification of items, and (c) lack of internal consistency among identified items for this sample.

NOTES

1. Signs of d ratios were ignored in the mean d calculations.
2. Educational Testing Service (ETS) researchers have adapted conventions by which items are classified as to the level of bias. We have utilized this same classification scheme to classify Hogan Personality Inventory (HPI) items. "A" items are those free from bias and are indicated by a delta value that is not significantly different from 0 ($p < .05$) or delta values possessing an absolute value less than 1.00. "B" items are considered marginally biased and are identified by (a) an absolute value greater than 1.00 but less than 1.50 or (b) at least 1.00 but not significantly greater than 1.00, either at $p < .05$. "C" items carry the greatest magnitude of bias. For these items, delta values are at least 1.5 and are significantly greater than 1.00 ($p < .05$; Schmitt, Hattrup, & Landis, 1993; Zwick & Erickson, 1989).
3. The largest delta value is actually found in the Likeability scale. But, because it contains the smallest number of biased items (3), the mean delta value of this scale is considered unreliable.
4. Although we chose to use principal components analyses (PCA) with varimax rotation because we wanted to reduce the intercorrelations into smaller subsets and did not expect that the components are correlated, we examined whether similar factor solutions also could emerge from other types of extraction and rotation methods. Although factor loadings are slightly lower with principal axis factoring, the pattern of item clusters is almost identical. The factor correlations that were generated from oblique rotation were very low, indicating orthogonal rotation was permissible.
5. On the other hand, group mean differences were, as expected, trivial. This is consistent with the findings from Ones and Anderson's (2002) study using British university students. On the seven HPI scales that we examined, they found the same average effect size as we did (mean $d = .14$), and the race differences they reported were almost identical to ours (mean $d = .02$ vs. $.01$, respectively).
6. Certainly, additional studies should be carried out to determine if similar levels of item bias are found on the HPI with different populations and on other employment-oriented personality inventories.

7. Another argument for keeping moderately biased items is the low probability of bias correspondence across group categories. A test that is unbiased against one group may still be biased against other groups. For example, items biased by sex are unlikely to be biased by race. We found little cross-group classification on differential item functioning (DIF) correspondence (these analyses are available from the second author). Only a small number of items (30%) showed both gender and race DIF. Therefore, eliminating items with DIF by gender does not guarantee that the test will be free of items biased by race, or vice versa. The multiplicity of possible groups and simultaneous membership in multiple groups are two related issues that suggest that it may be unwise to eliminate moderately biased items. People can be categorized in a multiplicity of subgroups, such as sex, race, ethnicity, culture, and language, and within each subgroup there are multiple classifications. For example, an individual could be simultaneously female, ethnically and linguistically Hispanic, and of Black descent. What is the proper classification of this person for bias analysis? These multiplicities of subgroup classifications imply ad infinitum possibilities of bias analyses. Finding a sufficient pool of unbiased items for so many types of subgroup analyses would be exceedingly difficult.

REFERENCES

- Arbisi, P. A., Ben-Porath, Y. S., & McNulty, J. (2002). A comparison of MMPI-2 validity in African American and Caucasian psychiatric inpatients. *Psychological Assessment, 14*, 3-15.
- Azocar, F., Arean, P., Miranda, J., & Munoz, R. F. (2001). Differential item functioning in a Spanish translation of the Beck Depression Inventory. *Journal of Clinical Psychology, 57*, 355-365.
- Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.
- Bobko, P., & Bartlett, C. (1978). Subgroup validities: Differential definitions and differential prediction. *Journal of Applied Psychology, 63*, 12-14.
- Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Manual for the restandardized Minnesota Multiphasic Personality Inventory: MMPI-2. An administrative and interpretive guide*. Minneapolis: University of Minnesota Press.
- Cavalli-Sforza, L. L. (2000). *Genes, peoples, and languages*. Berkeley: University of California Press.
- Chen, C., Burton, M., Greenberger, E., & Dmitrieva, J. (1999). Population migration and the variation of dopamine D4 receptor (DRD4) allele frequencies around the globe. *Evolution and Human Behavior, 20*, 309-324.
- 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*. Hillsdale, NJ: Lawrence Erlbaum.
- Dai, G., Han, K., Hui, H., & Colarelli, C. M. (2006, May). *Examining measurement invariance of the Chinese Version NEO-PI-R Conscientiousness Scale*. A poster presented at the 21st annual conference at the Society for Industrial and Organizational Psychology, Dallas, Texas.
- Daly, M., & Wilson, M. (1999, July 1). Darwinism and the roots of machoism. *Scientific American, 19*, 8-14.
- Day, D. V., & Silverman, S. B. (1989). Personality and job performance: Evidence of incremental validity. *Personnel Psychology, 42*, 25-36.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.
- Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95*, 134-135.

- Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 3, pp. 577-636). Palo Alto, CA: Consulting Psychologists Press.
- Eagly, A. H., & Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinants of sex differences in influential: A meta-analysis of social influence studies. *Psychological Bulletin*, *90*, 1-20.
- Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version the Trier Personality Inventory. *Journal of Cross-Cultural Psychology*, *24*, 133-148.
- Geary, D. C. (1998). *Male, female*. Washington, DC: American Psychological Association.
- Goldberg, L. R. (1990). An alternative "description of personality": The Big Five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216-1229.
- Gorski, R. A. (1991). Sexual differentiation of the endocrine brain and its control. In M. Motta (Ed.), *Brain endocrinology* (2nd ed.). New York: Raven.
- Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC: National Academy Press.
- Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology*, *10*, 249-254.
- Hogan, R. (1986). *Manual for the Hogan Personality Inventory*. Minneapolis, MN: National Computer Systems.
- Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Kaihla, P. (2003, November). Getting inside the boss's head. *Business 2.0*, pp. 49-51.
- Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.
- McNulty, J. L., Graham, J. R., Ben-Porath, Y. S., & Stein, L. A. R. (1997). Comparative validity of MMPI-2 scores of African American and Caucasian mental health center clients. *Psychological Assessment*, *9*, 464-470.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.
- Ones, D. S., & Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology*, *75*, 255-276.
- Roy, R., & Benenson, J. F. (2002). Sex and contextual effects on children's use of interference competition. *Developmental Psychology*, *38*, 306-312.
- Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement*, *59*, 248-269.
- Saad, S., & Sackett, P. R. (2002). Investigating differential prediction by sex in employment-oriented personality measures. *Journal of Applied Psychology*, *87*, 667-674.
- Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, *49*, 929-954.
- Scheuneman, J. D., & 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.
- Schmit, M. J., Kihm, A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology*, *53*, 153-193.
- Schmitt, N., Hatrup, K., & Landis, R. S. (1993). Item bias indices based on total test score and job performance estimates of ability. *Personnel Psychology Special Issue: Innovations in Research Methods for Field Settings*, *46*, 593-611.
- Smith, L. L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychological Bulletin*, *28*, 754-763.
- Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, *44*, 703-742.
- Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group differences: The concept of item bias. *Psychological Bulletin*, *99*, 118-128.
- Timbrook, R. E., & Graham, J. R. (1994). Ethnic differences on the MMPI-2? *Psychological Assessment*, *6*, 212-217.
- Waller, N. G., Thompson, J. S., & 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.
- Whybrow, P. C. (2005). *American mania*. New York: Norton.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum.
- Zwick, R., & Erickson, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*, 55-66.

Richard Sheppard is a human resources consultant within the corrections industry. Most recently, he conducted research for the California Board of Corrections, including oversight of efforts to develop entry-level examinations used by the state to select adult corrections officers, juvenile corrections officers, and probation officers. His research interests include personnel selection, turnover, performance measurement of corrections and probation personnel, and measurement of test bias.

Kyunghee Han is associate professor of psychology at Central Michigan University. Her broad research interests involve cross-cultural psychology, personality assessment, and quantitative methods. She is an official translator of the Korean MMPI-2 and MMPI-A and currently focuses on standardization of these instruments.

Stephen M. Colarelli is professor of psychology at Central Michigan University. He is interested in the application of evolutionary theory to organizational behavior and human resource management. His most recent book *No Best Way: An Evolutionary Perspective on Human Resource Management* integrates ideas from evolutionary theory and human resource management.

Guangrong Dai is a graduate student of industrial/organizational psychology at Central Michigan University and a researcher at Lominger Limited, Inc. His research interests include leadership competency assessment, cross-cultural human resource management, and application of modern testing theory in noncognitive tests.

Daniel W. King is a professor in the departments of psychology and psychiatry at Boston University and is affiliated with the VA Boston Healthcare System. His research interests emphasize the application of novel quantitative methods to the study of stress, trauma, and their sequelae.