PERSONNEL PSYCHOLOGY
2008, 61, 153–172

THE DIVERSITY–VALIDITY DILEMMA: STRATEGIES FOR REDUCING RACIOETHNIC AND SEX SUBGROUP DIFFERENCES AND ADVERSE IMPACT IN SELECTION

ROBERT E. PLOYHART
University of South Carolina

BRIAN C. HOLTZ
University of Calgary

Pyburn, Ployhart, and Kravitz (this issue, 2008) introduced the diversity–validity dilemma: that some of the most valid predictors of job performance are also associated with large racioethnic and sex subgroup predictor score differences. This article examines 16 selection strategies hypothesized to minimize racioethnic and sex subgroup differences and adverse impact and, hence, balance diversity and validity. Rather than presenting a highly technical review, our purpose is to provide practitioners with a concise summary, paying particular attention to comparing and contrasting the effectiveness of the strategies and reporting new developments. The paper is organized around 4 key questions: (a) Which strategies are most effective for reducing subgroup differences? (b) Which strategies do not involve a validity tradeoff? (c) What are the major new developments in strategies for reducing adverse impact? (d) What are the major new developments in alternative predictor measurement methods (e.g., interviews, situational judgment tests, assessment centers) for reducing adverse impact? We then conclude with recommendations and caveats for how to best balance diversity and validity. These ideas are developed further in Kravitz (this issue, 2008), who considers even broader approaches for solving the diversity–validity dilemma.

Recruiting and selecting competent employees is critical to an organization's competitive advantage. Many organizations also believe that creating a diverse workforce is important for business, social, or ethical reasons. Recruitment and selection are essential mechanisms for increasing diversity, but some of the most valid selection procedures exhibit racioethnic and sex differences such that minority (non-White) and female subgroups

We thank the editor, three anonymous reviewers, David Kravitz, Keith Pyburn, and Fred Oswald for their many helpful suggestions and recommendations.

Correspondence and requests for reprints should be addressed to Robert E. Ployhart, Darla Moore School of Business, University of South Carolina, Columbia, SC 29208; ployhart@moore.sc.edu.

score lower than White and male groups. Pyburn, Ployhart, and Kravitz (2008) refer to this as the diversity–validity dilemma, a dilemma requiring organizations to choose between diversity and optimal prediction. They note that due to legal constraints on overt preferences (particularly the ban on proportional hiring from within each subgroup), organizations may solve this dilemma by using less valid selection procedures. This helps them achieve their diversity goals but with potentially substantial losses of utility.

Practitioners who struggle with these issues should be aware of alternative approaches that might help balance these goals. The purposes of this article are to concisely summarize and organize a large and diffuse research literature, and present 16 strategies that have been proposed to reduce racioethnic (i.e., White vs. Black, Hispanic, or Asian) and sex subgroup differences and adverse impact. We compare and contrast the strategies in their effectiveness and feasibility, providing information to practitioners who can then select those best suited for their particular situation. This paper updates prior reviews of this topic (e.g., Hough, Oswald, & Ployhart, 2001; Sackett, Schmitt, Ellingson, & Kabin, 2001) by presenting several new developments, apparent changes to prior conclusions, and new strategies that warrant consideration. References are provided in tables for readers wanting more detailed information. Our review is organized around key questions faced by practitioners: (a) *Which strategies are most effective for reducing subgroup differences?* (b) *Which strategies do not involve a validity tradeoff?* (c) *What are the major new developments in strategies for reducing adverse impact?* (d) *What are the major new developments in alternative predictor measurement methods for reducing adverse impact?*

Predictor Subgroup Differences

As noted in Pyburn et al. (2008), mean score differences between racioethnic minority (Black, Hispanic, Asian) or female subgroups, and racioethnic majority (White) or male subgroups, contribute to adverse impact. Table 1 presents subgroup differences and validity estimates associated with various predictor constructs and alternative predictor measurement methods. *Predictor constructs* are homogenous and tap a single latent construct (e.g., cognitive ability, Agreeableness). *Alternative predictor measurement methods* (e.g., interviews, assessment centers) measure multiple latent constructs simultaneously to create a single composite score. They are called alternative predictor measurement methods because they are alternatives to traditional multiple-choice tests and because a single method can measure different constructs. For example, one can use assessment centers to measure interpersonal skills and personality.

TABLE 1

Meta-Analytic Standardized Racioethnic and Sex Subgroup Differences and Validities

Predictor	*d*-value (uncorrected)	Criterion-related validity (corrected)
Predictor constructs		
General cognitive ability		.51[b]
White–Black	.99[a]	
White–Hispanic	.58 to .83[a]	
White–Asian	−.20	
Male–Female	.00	
Temperament–Extraversion		.11[c]
White–Black	.10	
White–Hispanic	−.01	
White–Asian	.15	
Male–Female	.09	
Temperament–Conscientiousness		.18[c]
White–Black	.06	
White–Hispanic	.04	
White–Asian	.08	
Male–Female	−.08	
Temperament–Emotional Stability		.13[c]
White–Black	−.04	
White–Hispanic	−.01	
White–Asian	.08	
Male–Female	.24	
Temperament–Agreeableness		.08[c]
White–Black	.02	
White–Hispanic	.06	
White–Asian	.01	
Male–Female	−.39	
Temperament–Openness to Experience		.07[c]
White–Black	.21	
White–Hispanic	.10	
White–Asian	.18	
Male–Female	.07	
Job knowledge		.48[b]
White–Black	.48[d]	
White–Hispanic	.47[d]	
Spatial ability		.51[f]
White–Black	.66[e]	
Psychomotor ability		.35[h]
Male–Female	−1.06[g,†]	
White–Black	−.72[g,†]	
White–Hispanic	−.11[g,†]	
Physical ability: Muscular strength		.23[i]
Male–Female	1.66	
Physical ability: Muscular power		.26[j]
Male–Female	2.10	
Physical ability: Muscular endurance		.23[j]
Male–Female	1.02	

TABLE 1 (continued)

Predictor	*d*-value (uncorrected)	Criterion-related validity (corrected)
Alternative predictor measurement methods		
Biodata		.35[b]
White–Black	.33[k]	
Structured interview		.51[b]
White–Black	.23[k]	
Situational judgment (video)		.22 to .33[m,†]
White–Black	.31[l]	
White–Hispanic	.41[l]	
White–Asian	.49[l]	
Male–Female	−.06[l]	
Situational judgment (written)		.34[n]
White–Black	.40[l]	
White–Hispanic	.37[l]	
White–Asian	.47[l]	
Male–Female	−.12[l]	
Accomplishment record		.17 to .25[o,†]
White–Minority	.24[†]	
Male–Female	.09[†]	
Work sample		.33[p]
White–Black	.52[d]	
White–Hispanic	.45[d]	
Assessment center		.37[b]
White–Black	.60 or less; depending on exercise/dimension[†]	

Notes. Estimates are intended to be representative, not exhaustive (estimates for some groups are not presented due to insufficient data). Positive values indicate the White (or male) group scores higher than the racioethnic minority (or female) group. *Alternative predictor measurement methods* are predictors that are alternatives to traditional multiple choice type testing and usually assess multiple KSAOs. Unless noted, *d*-values come from Hough et al. (2001). Unless noted, all validities are corrected meta-analytic estimates.

[a]Roth, Bevier, Bobko, Switzer, and Tyler (2001).
[b]Schmidt and Hunter (1998).
[c]Hough and Furnham (2003).
[d]Roth, Huffcutt, & Bobko (2003).
[e]Schmitt, Clause, and Pulakos (1996).
[f]Salgado, Anderson, Moscoso, Bertua, and De Fruyt (2003).
[g]Knapp, Sager, and Tremble (2005).
[h]Hunter and Hunter (1984).
[i]Lewis (1989; cited in Hogan, 1991—training criterion).
[j]Lewis (1989; cited in Hogan, 1991—supervisor ratings criterion).
[k]Bobko, Roth, and Potosky (1999).
[l]Nguyen, McDaniel, & Whetzel (2005).
[m]Weekley and Jones (1997).
[n]McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001).
[o]Hough (1984).
[p]Roth, Bobko, & McFarland (2005).
[†]These estimates are based on primary studies and are *not* corrected/meta-analytic estimates.

To allow comparisons across samples and predictors, we report subgroup differences in terms of the *d* statistic (the mean of the majority [or male] group minus the mean of the racioethnic minority [or female] group divided by their pooled standard deviations). This represents differences in standard deviation units (e.g., $d = .50$ indicates the groups differ by .50 standard deviation units). Larger subgroup differences push the predictor score distributions of the subgroups farther apart and make it less likely to hire members of the lower scoring group when using top-down selection. Positive *d*-values indicate the majority or male group scores higher. Table 1 illustrates the diversity–validity dilemma, as well as a number of new findings (or reversals to older findings) for alternative predictor measurement methods that have not previously been discussed but are detailed below.

Strategies for Reducing Subgroup Differences

Table 2 summarizes 16 strategies for reducing subgroup differences and provides information about their effectiveness, implementation, and key references. Effectiveness is defined in terms of how well the strategy *reduces* subgroup differences, frequently in comparison to the sole use of cognitive ability because this has the highest validity but the largest White–Black subgroup difference. The table summarizes an extensive literature and is organized into *categories* based on the underlying similarities of the strategies. Within each category, strategies are roughly rank ordered by effectiveness.

Strategies in Category I use predictors that have smaller subgroup differences than overall cognitive ability. For example, Strategy 1 involves alternative predictor measurement methods such as interviews and assessment centers because they generally demonstrate lower racioethnic subgroup differences than cognitive ability. Strategies in Category II combine or manipulate scores to lower subgroup differences. For example, Strategy 4 uses multiple cognitive and non cognitive predictors to balance high and low subgroup difference predictors. Strategies in Category III attempt to remove construct irrelevant variance from predictor scores (see Messick, 1995). Construct irrelevant variance represents sources of variance (e.g., language, cultural differences) that correlate with subgroup membership and predictor scores, but are not part of the individual differences of interest. Strategies in Category IV allow practice prior to testing, or retesting if the applicant is rejected. Strategies in Category V attempt to foster favorable applicant reactions to assist in recruiting and performance in the selection system.

TABLE 2

Strategies for Reducing Racioethnic and Sex Subgroup Differences and Adverse Impact

Strategy and premise	Effects on reducing subgroup differences	Comments	Key references
Category I. Strategies that use predictors with smaller subgroup differences than cognitive ability			
1. *Use alternative predictor measurement methods (e.g., interviews, work samples, assessment centers, situational judgment tests, biodata).* Using alternative predictor measurement methods will reduce subgroup differences because they measure multiple cognitive and non-cognitive KSAOs, frequently minimize reading requirements, may engender more favorable reactions, and/or are based on job performance tasks for which subgroup differences are smaller.	• Generally effective, but specific reductions are quite variable (Table 1).	• Predictors with smaller cognitive loadings produce smaller differences. • Differences tend to be larger in applicant than incumbent settings. • Magnitudes of reductions are affected by predictor type and subgroup. • Some methods decrease differences for one group but increase them for another. • Validity may be lower than for overall cognitive ability (sometimes marginally). • Lower reliability of alternative predictor measurement methods may attenuate subgroup differences and give the erroneous impression that they are much smaller. • Developing/administering/scoring alternative predictor measurement methods is expensive/time consuming. • Applicant faking may be an issue.	Hough et al. (2001); Schmitt et al. (1996)
2. *Use educational attainment or GPA as a proxy for cognitive ability.* Because GPA and/or educational attainment are related to conscientiousness and motivational constructs, in addition to cognitive ability, using them as proxies for cognitive ability will reduce adverse impact.	• Small to moderate reduction in subgroup differences compared to cognitive ability.	• Subgroup differences increase as educational attainment increases. • Less valid than cognitive ability. • Educational attainment *may* be more useful than GPA. • Research primarily for White–Black differences. • Applicant faking may be an issue when using self-reports.	Berry, Gruys, & Sackett (2006), Roth and Bobko (2000)

TABLE 2 (continued)

Strategy and premise	Effects on reducing subgroup differences	Comments	Key references
3. *Use specific measures of ability.* Specific (narrow) measures of cognitive ability (e.g., verbal, quantitative) have smaller subgroup differences than overall cognitive ability.	• Small to moderate reduction in racioethnic *d*-values compared to overall ability measures. • Male/female differences may be larger than for overall ability and may favor men (quantitative ability) or women (verbal ability).	• With broad measures of performance, validity may be lower than when using overall ability. • Generally useful only with specific criteria (e.g., reading proficiency). • May need to administer more predictors, potentially increasing costs and time for administration and scoring (and if administering several measures of specific abilities, the predictor battery may consequently assesses overall cognitive ability).	Hough et al. (2001)
	Category II. Strategies that combine and manipulate scores		
4. *Assess the full range of KSAOs.* If cognitive ability is one of the most valid predictors but also exhibits the highest subgroup differences, then adding noncognitive predictors that are related to performance but engender smaller subgroup differences may reduce the overall subgroup difference of the predictor battery.	• Generally effective, but the magnitude of reduction depends on predictor validities and intercorrelations.	• Diminishing returns after adding four or more predictors. • The predictor with the highest validity will most determine the composite subgroup difference (when using regression-based weights). • Including a full battery of predictors usually produces higher validity. • Including more predictor KSAOs is expensive/time consuming. • Applicant faking may be an issue. • Must consider issues involved with adding and combining predictors in a battery (e.g., relative importance, incremental importance, incremental validity, and so on).	Bobko et al. (1999), LeBreton, Griepentrog, Hargis, Oswald, and Ployhart (in press), Ryan, Ployhart, and Friedel (1998); Sackett and Ellingson (1997), Sackett and Roth (1996), Schmitt et al. (1996)

TABLE 2 (continued)

Strategy and premise	Effects on reducing subgroup differences	Comments	Key references
5. *Banding and score adjustments.* There is no perfectly reliable predictor; acknowledging this unreliability by creating "bands" from within which scores cannot be empirically distinguished may increase racioethnic minority or female hiring.	• Reductions can be sizeable if using racioethnic minority or female preference within bands; otherwise reductions are small or nonexistent.	• Many factors influence effects of banding, including selection ratio, proportion of racioethnic minority or female applicants, and procedure for hiring within bands. • Racioethnic minority or female preference is usually illegal. • Recent questions about appropriate form of reliability estimate. • May lower validity.	Aguinis (2004), Campion et al. (2001), Murphy, Osten, and Myors (1995), Sackett and Roth (1991)
6. *Explicit predictor weighting.* Rather than simply summing the predictors or using regression-based weights, one may rationally give more weight to predictors with less adverse impact.	• Small to moderate reduction in subgroup differences.	• Greater reduction likely comes from choosing which predictors to put in the battery, rather than differential weighting within the battery. • Validity may be lowered with rationally derived weights. • Applicant faking may be an issue if using non cognitive predictors.	DeCorte (1999), DeCorte and Lievens (2003), Ryan et al. (1998)

TABLE 2 (continued)

Strategy and premise	Effects on reducing subgroup differences	Comments	Key references
7. *Criterion weighting.* Task and/or technical dimensions of performance are more strongly predicted by cognitive ability, whereas contextual/non technical dimensions of performance are more strongly predicted by non cognitive measures like personality. Emphasizing contextual/non technical dimensions will therefore reduce adverse impact through increasing the importance (validity) of non cognitive predictors.	• Small to moderate reduction in adverse impact when weighting contextual or non technical performance.	• Reductions are frequently not large unless selection ratio is high. • Criterion weighting will strongly influence validity, so cannot simply overweight contextual dimensions. • Applicant faking may be an issue if using non cognitive predictors.	Hattrup, Rock, and Scalia (1997), Murphy and Shiarella (1997)
Category III. Strategies that reduce construct irrelevant variance from predictor scores			
8. *Minimize verbal ability requirements to the extent supported by job analysis.* By assessing verbal ability only to the level supported by a job analysis, and/or using video-based predictors, this strategy reduces variance from subgroup differences in verbal ability and may enhance applicant reactions.	• Generally effective but the magnitude is variable.	• Must demonstrate equivalence when developing "lower verbal ability" alternative. • Must ensure verbal ability is not contributing to the validity of the alternative. • Developing/administering/scoring video-based or non written predictor methods is expensive/time consuming. • Verbal ability requirements cannot be lower than the minimum level identified in the job analysis.	Arthur, Edwards, & Barrett (2002), Sacco, Scheu, Ryan, Schmitt, Schmidt, and Rogg (2000), Sackett et al. (2001)

TABLE 2 (continued)

Strategy and premise	Effects on reducing subgroup differences	Comments	Key references
9. *Use "content free" items* that are not more (un)familiar to, or do not serve to advantage, any particular cultural subgroup.	• Small and inconsistent.	• Difficult to write items truly representative of cultures, but equivalent for all cultures. • Little theoretical support in selection contexts.	DeShon, Smith, Chan, and Schmitt (1998), Whitney and Schmitt (1997)
10. *Differential Item Functioning (DIF).* Removing items that demonstrate DIF will reduce subgroup differences.	• Small and inconsistent.	• DIF will favor both White (or male) groups and racioethnic minority (or female) groups for different items. • Very little success in developing theories or explanations of DIF—very empirically driven.	Hough et al. (2001), Sackett et al. (2001)
11. *Sensitivity review panels.* Developers of predictors use what is called a "sensitivity review panel" to examine items and ensure they are appropriate and non-offensive to all relevant subgroups.	• No data on effectiveness.	• They are frequently used in practice, but there is no strong empirical evidence supporting their use. • At present, main benefit is probably for public relations. • May be costly and time consuming to implement sensitivity review panels.	Reckase (1996)
12. *No time limits.*	• No clear reductions.	• White and racioethnic minority groups both improve; White group sometimes improves more. • Extending time limits may be expensive/time consuming.	Sackett et al. (2001)

TABLE 2 (continued)

Strategy and premise	Effects on reducing subgroup differences	Comments	Key references
		Category IV. Strategies that allow practice	
13. *Retesting.* Allowing applicants to reapply for the job will reduce subgroup differences on the predictor and hence adverse impact.	• Small to no reduction.	• Predictor scores tend to improve with retesting for Whites and Blacks. • Whites are more likely to retest, but Whites and Blacks are more likely to retest when in middle of score distribution. • Retesting can be expensive/time consuming.	Sin, Farr, Murphy, & Hausknecht (2004)
14. *Use predictor orientation programs.* Subgroup differences may be partly due to differing amounts of experience or familiarity with testing, so giving practice reduces the differences.	• Small and inconsistent.	• Very little data in selection contexts. • Racioethnic minority groups more likely to participate, so some public relations value. • Using orientation programs is expensive/time consuming.	Sackett et al. (2001)
		Category V. Strategies that foster favorable applicant reactions	
15. *Increasing and retaining racioethnic minority and female applicants.* Increasing or retaining the number of qualified racioethnic minority and female applicants in the pool will increase their hiring numbers and reduce subgroup differences.	• Small reductions.	• Racioethnic minorities and women are more likely to withdraw, but for reasons largely unrelated to the selection system. • Vast majority of research on Blacks and women. • Targeted recruiting and retention is expensive/time consuming.	Ryan, Sacco, McFarland, and Kriska (2000), Tam, Murphy, & Lyall (2004)
16. *Enhance applicant perceptions.* Subgroup differences stem from racioethnic minority or female applicants having less favorable predictor perceptions and motivation; enhancing perceptions will therefore reduce subgroup differences.	• Small reductions.	• Greatest benefit may be for public relations value. • Stereotype threat does not seem to explain subgroup differences. • Providing explanations may be relatively inexpensive, but developing more face-valid predictors can be expensive/time consuming.	Cullen, Hardison, & Sackett (2004), Hausknecht, Day, & Thomas (2004), Farr 2003; Ryan (2001)

Notes. Racioethnic minority = Black, Hispanic, or Asian; SJT = situational judgment test; KSAO = knowledge, skills, abilities, and other constructs; GPA = grade point average; DIF = differential item functioning; and *d*-values = represent standardized mean subgroup differences between White (or male) and racioethnic minority (or female) subgroups.

Question 1: Which Strategies Are Most Effective for Reducing Subgroup Differences?

The most effective *categories* of strategies involve using predictors with smaller subgroup differences (Category I) and combining/manipulating predictor scores (Category II). We consider these "proximal" categories because they most directly address the problem of mean predictor score differences. Strategies that remove construct irrelevant variance (Category III), allow practice (Category IV), or enhance applicant reactions (Category V) are more "distal" and consequently less effective.

In terms of specific strategies, the most effective involve using alternative predictor measurement methods such as interviews and assessment centers (Strategy 1): assessing the entire range of knowledge, skills, abilities, and other constructs (KSAOs; Strategy 4): banding (Strategy 5; but only when using racioethnic minority or female preference): and minimizing the verbal ability requirements of the predictor (Strategy 8; but only to the extent supported by a job analysis). Again, these address the proximal problem of predictor score differences. Interestingly, minimizing verbal ability requirements (Strategy 8) is the only strategy in Category III (removing construct irrelevant variance) that is reasonably effective. Other approaches, such as using educational attainment or grade point average (GPA) as a proxy for cognitive ability (Strategy 2) or explicit predictor weighting (Strategy 6), can be nearly as effective, but their effectiveness is more variable.

Question 2: Which Strategies Do Not Involve a Validity Tradeoff?

Among the most *effective* strategies, the only strategy that does not also reduce validity is assessing the full range of KSAOs (Strategy 4). In fact, this strategy tends to enhance validity. The other *effective* strategies will reduce validity, although sometimes to only a small degree. For example, minimizing verbal ability (Strategy 8) appears to lower validity for some predictors (situational judgment) but not others (assessment centers). It is noteworthy that many of the more distal techniques, such as retesting (Strategy 13), increasing and retaining racioethnic minority and female applicants (Strategy 15), and enhancing applicant reactions (Strategy 16), appear to have little to no effect on validity. Although these distal techniques do not have the same levels of effectiveness as the proximal techniques, they also do not have the validity tradeoff.

Question 3: What Are the Major New Developments in Strategies for Reducing Adverse Impact?

There have been several new developments that we can only summarize here. One development has been to study the consequences of retesting,

or allowing rejected applicants to reapply for the job (Strategy 13). Sin, Farr, Murphy, and Hausknecht (2004) examined the effects of retesting on White–Black subgroup differences over 4 consecutive years. They found rejected White applicants were more likely to retake the predictor. More interesting was an inverted U-shaped relationship between predictor performance and propensity to retake the selection predictor. For both Whites and Blacks, the most likely applicants to retake the predictor were those whose scores were in the middle of the predictor score distribution. Finally, applicants who retook the predictor tended to score higher than applicants who took the predictor for the first time, but subgroup differences did not change even with repeated exposure to the predictor. Therefore, retesting does not appear to offer much potential for reducing racioethnic adverse impact.

A second development examines whether reducing racioethnic minority withdrawal in selection procedures reduces adverse impact (Strategy 15). Tam, Murphy, and Lyall (2004) used archival data from state police officer selection procedures to design a Monte Carlo study in which they manipulated assumptions of the causes of withdrawal and the amount of withdrawal in Black and White subgroups. Relative to the effects of subgroup predictor score differences, they found very small effects of Black withdrawal on adverse impact. Even when withdrawal was related to predictor scores, reducing Black withdrawal would not produce a substantial reduction in adverse impact.

Finally, there have been some new developments attempting to link applicant reactions to subgroup differences (Strategy 16). Meta-analyses have found racioethnic differences in perceptions, with Blacks usually having more negative reactions (Hausknecht, Day, & Thomas, 2004; Ryan, 2001). Most of the recent research on this strategy has explored the effects of stereotype threat. The argument is that racioethnic minorities and women have lower predictor performance in part because they fear confirming negative group stereotypes (e.g., Steele, Spencer, & Aronson, 2002). Despite claims in the popular press, the data are not supportive. A special issue of *Human Performance* (Volume 16(3), 2003) contained four studies that manipulated stereotype threat in simulated selection contexts, and none supported the effect (see Sackett, 2003; Sackett, Hardison, & Cullen, 2004). Cullen, Hardison, and Sackett (2004) found no effect for stereotype threat in two large field studies. Therefore, it appears stereotype threat does little to explain subgroup differences in selection contexts.

Question 4: What Are the Major New Developments in Alternative Predictor Measurement Methods for Reducing Adverse Impact?

Given that use of alternative predictor measurement methods (Strategy 1) is among the most effective strategies and has been an active area

of recent research, we devote some attention to considering new developments in this area. In addition, there have been several caveats and possibly reversals of traditional thinking.

Interviews manifest Black and Hispanic scores about one-quarter of a standard deviation lower than Whites (Huffcutt & Roth, 1998), yet have reasonably high validity. Racioethnic subgroup differences are larger as the cognitive loading of the interview increases. Structured interviews demonstrate higher validity than unstructured interviews. They also exhibit smaller racioethnic differences than unstructured interviews, but the results for sex are less clear (Huffcutt, Conway, Roth, & Stone, 2001). Also unclear is the reduction in subgroup differences that may occur through the use of panels of diverse interviewers.

Situational judgment tests (SJTs) present hypothetical work situations and ask respondents to make judgments regarding a number of possible response options. SJTs exhibit useful criterion-related validity with smaller White–Black differences than cognitive ability, but other racioethnic subgroup differences appear to be larger than previously thought (here we summarize a meta-analysis by Nguyen, McDaniel, & Whetzel, 2005). In particular, Asians score lower than Whites on SJTs but higher on cognitive ability. Hence, to claim SJTs reduce subgroup differences appears to be dependent on the subgroup of interest. Video-based SJTs exhibit slightly smaller subgroup differences than written SJTs but not for all subgroups, and the reduction appears to be smaller (about .10) than early studies indicated. Therefore, SJTs are useful but do not demonstrate universally favorable reductions in mean differences across racioethnic subgroups. The amount of reduction appears to be driven by the cognitive-loading of the SJT. Finally, subgroup differences are higher in applicant settings than incumbent settings for racioethnic differences, but the opposite is true for sex differences.

Assessment centers display high predictive validity and have been thought to have little to no adverse impact. Nevertheless, racioethnic subgroup differences in assessment centers vary as a function of the "cognitive loading" of a given exercise (Goldstein, Yusko, Braverman, Smith, & Chung, 1998; Goldstein, Yusko, & Nicolopoulos, 2001). Therefore, the extent to which assessment centers reduce racioethnic differences is a function of how much they also measure cognitive ability. Statistically controlling for cognitive ability reduced these differences with no substantial loss in validity.

Work samples are traditionally believed to have high predictive validity (Schmidt & Hunter, 1998) and moderate racioethnic subgroup differences (.52 SD, Roth, Huffcutt, & Bobko, 2003). Nevertheless, in applicant samples where range restriction is not a problem, Bobko, Roth, and Buster (2005) found White–Black subgroup differences can be as high as .70 SDs.

Further, a more recent and comprehensive meta-analysis of work samples found the validity to be smaller than previously believed (corrected validity of .33, Roth, Bobko, & McFarland, 2005). Therefore, the subgroup differences associated with work samples may be larger, and validity smaller, than prior research has indicated.

GPA and *educational attainment* have been examined as proxies for cognitive ability (Strategy 2). These measures manifest smaller subgroup differences because they are related to both cognitive ability and motivational constructs. Berry, Gruys, and Sackett (2006) found that selecting on educational attainment would produce lower adverse impact against Blacks and Hispanics than cognitive ability across most selection ratios. Unfortunately, educational attainment has less validity than cognitive ability. Racioethnic adverse impact increased as the amount of education increased, but female adverse impact is lower and not present until setting cuts at 8 years of post secondary education. Roth and Bobko (2000) examined a large number of White–Black differences in GPA across majors and years (sophomore, junior, and senior). Overall, they found a $d = .43$, but the *d*s increased from .21 for sophomores to .78 for seniors. This strategy offers some promise if one can justify lower validity, but we know very little about GPA subgroup differences for different racioethnic groups. There are also numerous operational concerns such as comparability of GPA/educational attainment across schools, courses, majors, as well as faking/misrepresentation of credentials.

Finally, an attempt to reduce reading requirements involves the use of *constructed response options* (Strategy 8) rather than multiple choice options (Arthur, Edwards, & Barrett, 2002). Constructed response answers are similar to "open-ended" responses; they require the applicant to respond in his/her own words rather than choose from a set of responses. The premise is that allowing applicants to respond in their own words and styles will reduce subgroup differences because part of the difference is due to the multiple choice mode. Arthur et al. (2002) examined traditional multiple choice and constructed response cognitive ability predictors on promotional exams for firefighters. The White–Black subgroup differences for the usual multiple choice predictor was .70, but for the constructed response predictor it was only .12 and Blacks scored higher. Edwards and Arthur (2004) recently replicated these findings with a student sample and demonstrated they were partly due to reading load and perceptions of predictors.

Recommendations and Caveats

We have presented 16 strategies that have been proposed to reduce racioethnic and sex subgroup differences and adverse impact. No single

strategy is fully effective or universally best. Different strategies may be more practically feasible and/or effective in different situations. Having considered each strategy in isolation, we now provide an integrated recommendation for what an organization should do to minimize the diversity–validity dilemma.

1. Use job analysis to carefully define the nature of performance on the job, being sure to recognize both technical and non technical aspects of performance.
2. Use cognitive and non cognitive predictors to measure the full range of relevant cognitive and non cognitive KSAOs, as much as practically realistic.
3. Use alternative predictor measurement methods (interviews, SJTs, biodata, accomplishment record, assessment centers) when feasible. Supplementing a cognitive predictor with alternative predictor measurement methods can produce sizeable reductions of adverse impact (if they are not too highly correlated), but the specific reductions are variable (Table 1, Strategies 1 and 4). Using alternative predictor measurement methods is costly but effective because they measure multiple KSAOs, reduce reading requirements, and have higher face validity. Among the best alternative predictor measures are interviews, SJTs, and assessment centers. Nevertheless, the data are unclear about work samples, which may have less validity and larger racioethnic subgroup differences than prior research indicated.
4. Decrease the cognitive loading of predictors and minimize verbal ability and reading requirements to the extent supported by a job analysis. For example, if a job analysis indicates the need for a high school reading level, ensure the predictors do not require a college reading level. Doing so may involve lowering the reading level of instructions and items, allowing constructed response options, or using video formats (but again, only if consistent with the job analysis findings).
5. Enhance applicant reactions. Although this strategy has only a minimal effect on subgroup differences, it does not reduce validity and is almost invariably beneficial from a public relations perspective. Simply using face valid predictors (such as interviews and assessment centers) goes a long way toward enhancing these perceptions. And some approaches are free (e.g., giving explanations for why the selection procedure is being used). Sensitivity review panels may help ensure content validity and legal defensibility.
6. *Consider* banding. We emphasize the word "consider" because this remains a controversial strategy among IO psychologists and will substantially reduce subgroup differences only when there is explicit racioethnic minority or female preference in final hiring decisions. Only in limited instances will it be possible to justify such preferences (noted in Pyburn et al., 2008).

Organizations may wish to implement other strategies as appropriate to their specific situations. Although we find little benefit for their effects on subgroup differences, Kravitz (2008) presents data on several other outcomes where such interventions may have important benefits. It is important to think of the diversity–validity dilemma broadly and to address it broadly by using as many strategies as relevant and realistically feasible.

There are also some caveats to interpreting these research findings:

- For many strategies, we know little about subgroup differences other than White–Black (and to a lesser extent male–female). Given the large number of Hispanics and Asian-Americans in the workforce (and all of the variations within these racioethnic categories), more data are needed about these groups. This research is critical because different subgroups exhibit different amounts of mean score differences and adverse impact.
- There are costs (e.g., time, money, resources, personnel, and so on) associated with every strategy. Practitioners must balance the benefits of each strategy relative to its costs and feasibility. They should also consider the cost of a very effective strategy (e.g., Strategy 1) versus combinations of less effective strategies (e.g., Strategies 13 and 14). In a similar manner, implementing a strategy (e.g., Strategy 4) may introduce other operational issues (e.g., faking, concerns about invasiveness, negative reactions) that must be considered.
- Reducing subgroup differences for one group may exaggerate them for another. For example, Ryan, Ployhart, and Friedel (1998) demonstrated that, relative to selecting solely on cognitive ability scores, selecting solely on personality scores would reduce adverse impact against Blacks and Hispanics but would simultaneously increase adverse impact against women. In a similar manner, Table 2 shows White–Black subgroup differences are reduced when using a video SJT versus a paper SJT, but White–Hispanic and White–Asian differences are increased slightly.
- Practitioners must be cognizant of methodological factors that influence subgroup differences. Subgroup differences tend to be underestimated in incumbent settings because of range restriction. Differences in predictor reliability can distort comparisons of subgroup differences. Therefore, when comparing predictors, reviewing the literature, or conducting a concurrent validation study, realize that many of the same factors attenuating criterion-related validity are also attenuating subgroup differences.

Research on subgroup differences and adverse impact is still evolving, and new findings are constantly being published. A future study could change some of the conclusions we report in this paper, but at present these

strategies and recommendations appear warranted. Research has made important advances in understanding diversity, subgroup differences, individual differences, and validity, and we expect this to continue.

REFERENCES

Aguinis H. (2004). *Test score banding in human resource selection: Legal, technical, and societal issues*. Westport, CT: Quorum.

Arthur W, Edwards BD, Barrett GV. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. PERSONNEL PSYCHOLOGY, *55*, 985–1008.

Berry C, Gruys M, Sackett PR. (2006). Educational attainment as a proxy for cognitive ability in selection: Effects on levels of cognitive ability and adverse impact. *Journal of Applied Psychology, 91*, 696–705.

Bobko P, Roth PL, Buster MA. (2005). Work sample selection tests and expected reduction in adverse impact: A cautionary note. *International Journal of Selection and Assessment,* 1–10.

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

Campion MA, Outtz JL, Zedeck S, Schmidt FL, Kehoe JF, Murphy KR, et al. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. PERSONNEL PSYCHOLOGY, *54*, 149–185.

Cullen MJ, Hardison CM, Sackett PR. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology 89*, 220–230.

DeCorte W. (1999). Weighting job performance predictors to both maximize the quality of the selected workforce and control the level of adverse impact. *Journal of Applied Psychology, 84*, 695–702.

DeCorte W, Lievens F. (2003). A practical procedure to estimate the quality and adverse impact of single-stage selection decisions. *International Journal of Selection and Assessment, 11*, 89–97.

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

Edwards BD, Arthur W. (2004, April). *Race–based subgroup differences on a constructed response paper–and–pencil test*. Poster presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Farr JL. (Ed.). (2003). Stereotype threat effects in employment settings [Special Issue]. *Human Performance*, *16*(3).

Goldstein HW, Yusko KP, Braverman EP, Smith DB, Chung B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. PERSONNEL PSYCHOLOGY, *51*, 357–374.

Goldstein HW, Yusko KP, Nicolopoulos V. (2001). Exploring Black-White subgroup differences of managerial competencies. PERSONNEL PSYCHOLOGY, *54*, 783–807.

Hattrup K, Rock J, 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.

Hausknecht JP, Day DV, Thomas SC. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. PERSONNEL PSYCHOLOGY, *57*, 639–683.

Hogan J. (1991). Physical abilities. In Dunnette MD, Hough LM (Eds.), *Handbook of industrial and organizational psychology,* (Vol. 2, pp. 753–831). Palo Alto, CA: Consulting Psychologists Press.

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

Hough LM, Furnham A. (2003). Use of personality variables in work settings. In Borman WC, Ilgen DR. (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 131–169). New York: Wiley.

Hough LM, Oswald FL, Ployhart RE. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9*, 152–194.

Huffcutt AI, Conway JM, Roth PL, Stone NJ. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86*, 897–913.

Huffcutt AI, Roth PL. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83*, 179–189.

Hunter JE, Hunter R. (1984). Validity and utility of alternative predictors. *Psychological Bulletin, 96*, 72–98.

Knapp DJ, Sager CE, Tremble TR. (2005). *Development of experimental army enlisted personnel selection and classification tests and job performance*. Technical Report 1168, Army Research Institute, Arlington, VA.

Kravitz DA. (2008). The diversity-validity dilemma: Beyond selection—The role of affirmative action. PERSONNEL PSYCHOLOGY*, 61*, 173–193.

LeBreton JM, Griepentrog B, Hargis MB, Oswald FL, Ployhart RE. (2007). A multidimensional approach to evaluating variables in organizational research. PERSONNEL PSYCHOLOGY*, 60*, 475–498.

McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.

Messick S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Murphy KR, Osten K, Myors B. (1995). Modeling the effects of banding in personnel selection. PERSONNEL PSYCHOLOGY, *48*, 61–84.

Murphy KR, Shiarella AH. (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.

Nguyen NT, McDaniel MA, Whetzel D. (2005, April). *Subgroup differences in situational judgment test performance: A meta-analysis*. Paper presented at the 20th annual conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Pyburn KM, Jr., Ployhart RE, Kravitz DA. (2008). The diversity-validity dilemma: Overview and legal context. PERSONNEL PSYCHOLOGY*, 61*, 143–151.

Reckase MD. (1996). Test construction in the 1990s: Recent approaches every psychologist should know. *Psychological Assessment, 8*, 354–359.

Roth PL, Bevier CA, Bobko P, Switzer FS, III, Tyler P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. PERSONNEL PSYCHOLOGY, *54*, 297–330.

Roth PL, 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.

Roth PL, Bobko P, McFarland LA. (2005). A meta-analysis of work sample test validity. PERSONNEL PSYCHOLOGY, *58*, 1009–1037.

Roth PL, Huffcutt AI, Bobko P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology 88*, 694–706.

Ryan AM. (2001). Explaining the Black/White test score gap: The role of test perceptions. *Human Performance, 14*, 45–75.

Ryan AM, Ployhart RE, Friedel L. (1998). Using personality tests to reduce adverse impact: A cautionary note. *Journal of Applied Psychology, 83*, 298–307.

Ryan AM, Sacco JM, McFarland LA, Kriska SD. (2000). Applicant self-selection: Correlates of withdrawal from a multiple hurdle process. *Journal of Applied Psychology, 85*, 163–179.

Sacco JM, Scheu CR, Ryan AM, Schmitt N, Schmidt DB, Rogg KL. (2000). *Reading level and verbal test scores as predictors of subgroup differences and validities of situational judgment tests*. Unpublished manuscript.

Sackett PR. (2003). Stereotype threat in applied selection settings: A commentary. *Human Performance, 16*, 295–309.

Sackett PR, Ellingson JE. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. PERSONNEL PSYCHOLOGY, *50*, 707–721.

Sackett PR, Hardison CM, Cullen MJ. (2004). On interpreting stereotype threat as accounting for African American-White differences on cognitive tests. *American Psychologist, 59*, 7–13.

Sackett PR, Roth L. (1991). A Monte Carlo examination of banding and rank order methods of test score use in personnel selection. *Human Performance, 4*, 279–295.

Sackett PR, Roth L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. PERSONNEL PSYCHOLOGY, *49*, 549–572.

Sackett PR, Schmitt N, Ellingson JE, Kabin MB. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist, 56*, 302–318.

Salgado JF, Anderson N, Moscoso S, Bertua C, De Fruyt F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. PERSONNEL PSYCHOLOGY, *56*, 573–605.

Schmidt FL, Hunter JE. (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.

Schmitt N, Clause CS, Pulakos ED. (1996). Subgroup differences associated with different measures of some common job-relevant constructs. *International review of industrial and organizational psychology* (Vol. 11, pp. 115–139). New York: Wiley.

Sin HP, Farr JL, Murphy KR, Hausknecht JP. (2004). An investigation of Black-White differences in self-selection and performance in repeated testing. Paper presented at the 64th annual meeting of the Academy of Management. New Orleans, LA.

Steele CM, Spencer SJ, Aronson J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In Zanna MP (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379–440). San Diego, CA: Academic Press.

Tam AP, Murphy KR, Lyall JT. (2004). Can changes in differential drop out rates reduce adverse impact? A computer simulation study of a multi-wave selection system. PERSONNEL PSYCHOLOGY, *57*, 905–934.

Weekley JA, Jones C. (1997). Video-based situational testing. PERSONNEL PSYCHOLOGY, *50*, 25–49.

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