Psychological Topics 21 (2012), 3, 399-424

Review article – UDC – 65.012.32
159.944
658.310.8:159.9
159.913:159.94

An Update on the Diversity - Validity Dilemma in Personnel Selection: A Review

Britt De Soete, Filip Lievens

Department of Personnel Management, and Work and Organizational Psychology, Ghent University, Belgium

Celina Druart

Department of Data Analysis, Ghent University, Belgium

Abstract

As globalization increases and labor markets become substantially more diverse, increasing diversity during personnel selection has become a dominant theme in personnel selection in human resource management. However, while trying to pursue this goal, researchers and practitioners find themselves confronted with the diversity-validity dilemma, as some of the most valid selection instruments display considerable ethnic subgroup differences in test performance. The goal of the current paper is twofold. First, we update and review the literature on the diversity-validity dilemma and discuss several strategies that aim to increase diversity without jeopardizing criterion-related validity. Second, we provide researchers and practitioners with evidence-based guidelines for dealing with the dilemma. Additionally, we identify several new avenues for future research.

Keywords: diversity-validity dilemma, personnel selection, strategies for dealing with the diversity-validity dilemma

Introduction

Increasing globalization and international mobility have made contemporary labor markets substantially more diverse. As a result, current organizations often strive to employ an equally diverse workforce. However, achieving diversity often challenges organizations during recruitment and selection stages as ethnic minority group test-takers systematically obtain lower mean scores than majority group test-takers. These ethnic subgroup differences in selection performance can lead to

✉ Britt De Soete, Department of Personnel Management, Work and Organizational Psychology, H. Dunantlaan 2, 9000 Ghent, Belgium. E-mail: *Britt.DeSoete@UGent.be*.

Acknowledgement: This research was supported by a PhD grant from the Fund for Scientific Research – Flanders (FWO).

differential hiring rates for test-takers of different ethnic origin, which is called "Adverse Impact" (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978). Selection systems that display adverse impact exhibit substantially lower hiring rates for ethnic minority group applicants as compared to ethnic majority group applicants. This is associated with severe personal penalties for minority applicants (i.e., hiring chances and psychological consequences) and with costly juridical, ethical, or professional litigations for the organization (Arthur, Edwards, & Barrett, 2002). Increasing diversity by limiting the occurrence of ethnic subgroup differences in selection performance and restricting the potential for adverse impact has proven to be a quest. As can be derived from Table 1, some of the most valid selection instruments display large ethnic score differences in test performance (Hough, Oswald, & Ployhart, 2001; Ployhart & Holtz, 2008; Sackett, Schmitt, Ellingson, & Kabin, 2001). For example, cognitive ability tests have demonstrated to be one of the most valid predictors of job performance, but simultaneously display the largest ethnic subgroup differences in test performance as compared to other selection instruments. This phenomenon is labeled "the diversity-validity dilemma" in personnel selection (Ployhart & Holtz, 2008; Pyburn, Ployhart, & Kravitz, 2008). The dilemma implies that performance and diversity goals do not always go hand in hand in personnel selection, and it hinders organizations that aim to employ valid instruments while at the same time achieving acceptable levels of employee diversity.

Table 1. *Ethnic Subgroup Differences in Selection Test Performance*

Instrument		*Cohen's d*
Cognitive Ability Test		1.00[a]
Language Proficiency Test		0.40-0.76[b]
Big Five Personality	Extraversion	0.16[c]
	Agreeableness	0.03[c]
	Emotional Stability	0.09[c]
	Intellect	0.10[c]
	Conscientiousness	-0.07[c]
Structured Interviews		0.36-0.56[d]
Biodata		0.33[e]
Work Samples		0.52[f]-0.73[g]
Assessment Centers		0.52[h]
Situational Judgment Tests		0.24-0.38[i]

[a] Roth et al. (2001)
[b] Hough et al. (2001)
[c] Foldes et al. (2008)
[d] Roth et al. (2002)
[e] Bobko, Roth, & Potosky (1999)
[f] Roth et al. (2003)
[g] Bobko et al. (2005) and Roth et al. (2008)
[h] Dean et al. (2008)
[i] Whetzel et al. (2008)[1]

[1] To express ethnic performance differences in selection test performance, effect sizes of mean differences are often computed (Cohen's *d*, Cohen, 1994). The use of effect sizes permits to compare subgroup differences over different selection instruments. The *d*-values are obtained by subtracting the mean majority group score by the mean minority group score and dividing this measure by the pooled group standard deviation. Positive *d*-values indicate average test scores advantaging ethnic majority members, while negative *d*-values point to performance differences in favor of ethnic minority members.

During the last decades, this diversity-validity dilemma has been a dominant theme in industrial and organizational psychology research. Scientists have focused on strategies to reduce ethnic subgroup differences in selection performance, while simultaneously maintaining criterion-related validity. In 2008, Ployhart and Holtz published an overview article that categorized the available strategies that attempt to deal with the diversity-validity dilemma. They identified five main clusters of strategies to alleviate the dilemma, including (1) the employment of simulations and other predictors that display smaller ethnic subgroup differences as compared to cognitive ability, (2) reducing irrelevant test variance, (3) statistically combining and manipulating scores, (4) providing coaching and practice, and (5) fostering positive test-taker reactions. Exactly five years of research and knowledge accumulation after the original review of Ployhart and Holtz, we ask ourselves the following: Where are we now with regard to the diversity-validity dilemma and did new strategies emerge to alleviate this quandary?

The current paper aims to answer this question by updating and reviewing the recent literature on (a) the ethnic performance score gap in personnel selection and (b) strategies to deal with the diversity-validity dilemma. Although we broadly adhered to the original cluster structure (Ployhart & Holtz, 2008), new research findings and novel strategies have been added to the framework. By doing so, this review presents researchers and practitioners with a research overview of the current state-of-the-art regarding the diversity-validity dilemma, while at the same time offering a number of evidence-based guidelines for organizations that aim to select a competent as well as ethnically diverse workforce.

"Alternative" Cognitive Ability Measures

Employing alternative measures of cognitive ability as a substitute for or in combination with traditional cognitive ability tests covers an important category of strategies to deal with the diversity-validity dilemma. As current jobs evolve to be more complex and challenging, cognitive ability becomes even more important in the selection process (Gatewood, Feild, & Barrick, 2011). However, cognitive ability measures have repeatedly demonstrated to display large ethnic subgroup differences in test performance (Hough et al., 2001; Ployhart & Holtz, 2008; Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Sackett et al., 2001), which substantially decreases hiring chances for members of ethnic minority groups and increases the potential for adverse impact. Hence, researchers have advised to explore alternative measures of cognitive ability, which maintain validity and additionally exhibit substantially lower subgroup differences in test performance (e.g., Lievens & Reeve, 2012).

The first line of research within this domain has focused on logic-based measurement approaches as alternative measurement formats of cognitive ability (Paullin, Putka, Tsacoumis, & Colberg, 2010). Logic-based measurement

instruments require the application of reasoning skills and aim to measure logical thought processes. The Siena Reasoning Test (SRT), which presents applicants with novel and unfamiliar reasoning problems, serves as a promising example of this approach (Yusko, Goldstein, Scherbaum, & Hanges, 2012). Figure 1 provides an example item of the SRT. The test is time-constraint and consists of 25 or 45 reasoning items. Adequate criterion-related validity (r=.25-.48), and significantly smaller ethnic subgroup differences (d=0.38) as compared to traditional cognitive ability tests have been observed for the SRT (Yusko et al., 2012). Despite their promising psychometric characteristics and practical purposes, questions have also emerged about the effectiveness of logic-based measurement instruments. One concern is that several items have a rather high verbal load, which makes it difficult to apply the instrument to the selection of lower-level functions or in lingual diverse groups (e.g., in which group members have different mother tongues). Another concern is that capturing cognitive ability skills by presenting candidates with logic reasoning tasks may not only change the measurement method: it can also influence the constructs assessed. Consequently, logic-based measurement approaches may capture something different than *g*. As the cognitive load (or *g*-load) of an instrument has commonly been accepted as one of the most influential drivers of ethnic subgroup differences in test performance (Spearman's hypothesis, Jensen, 1998), the potentially lower *g*-load of logic-based measurement instruments would explain why they display lower ethnic score differences.

Figure 1. *Example item of the Siena Reasoning Test (Paullin, Putka, Tsacoumis, & Colberg, 2010)*

A GATH resembles a SHET but is heavier.
A SHET resembles a COUCH but is heavier.
A MUNT resembles a LAMP but is heavier.

Determine if each of the following statements is TRUE or FALSE.

Conclusion 1. A GATH is heavier than a SHET.
Conclusion 2. A COUCH is heavier than a GATH.
Conclusion 3. A LAMP is heavier than a COUCH.

The second strategy within this category aims to improve the point-to-point correspondence of the cognitive predictor with the criterion (i.e., job performance). In this context, Ackerman and Beier (2012) suggest to measure the result of intellectual *investments over time* (typical performance, such as job knowledge tests) rather than to capture maximal intellectual *capabilities at a given point in time* (maximal performance, such as traditional cognitive ability tests). Their argument is based on the fact that maximal prediction only occurs to the extent that predictors and criteria are carefully matched (see also Lievens, Buyse, & Sackett,

2005). However, in most cases maximal performance is assessed during the selection stage (e.g., by traditional cognitive ability tests), whereas typical performance is evaluated on the job. Consequently, Ackerman and Beier (2012) proposed to move away from measuring broad cognitive abilities, which are known for displaying substantial ethnic subgroup differences, and instead rely on knowledge tests as predictors. In a similar vein, a number of researchers plead for contextualizing cognitive ability measures. Contextualization refers to the process of adding circumstantial and situational (i.e., contextual) information to items as opposed to employing general and decontextualized items. In the case of personnel selection, this implies working with business-related cognitive items that confront applicants with realistic organizational issues and questions, instead of using generic cognitive ability items (Hattrup, Schmitt, & Landis, 1992). This approach may reduce implicit cultural assumptions that are often imbedded in generalized cognitive items (Brouwers & Van De Vijver, 2012).

Finally, researchers have suggested using specific cognitive abilities rather than overall cognitive ability measures when dealing with the diversity-validity dilemma as this often results in small to moderate reductions in ethnic subgroup differences (Hough et al., 2001; Ployhart & Holtz, 2008). In this context, the specific concept of "executive functioning", which received substantial research attention lately, seems particularly interesting (e.g., Huffcutt, Goebl, & Culbertson, 2012). Executive functioning relates to monitoring of events, shifting between tasks, dealing with situational and social parameters, and inhibition of tasks. This concept bears resemblances to the specific cognitive demands that are required for effective job performance.

In short, in recent years, a number of alternatives to traditional cognitive ability testing have been proposed. Logic-based measurement methods seem a valid alternative measure for cognitive skills, but in order to enhance our understanding of the drivers of ethnic score differences further research should focus on which characteristics of logic-based instruments cause lower adverse impact. Also the measurement of executive functioning instead of broad cognitive skills, capturing typical versus maximal cognitive performance, and developing contextualized versus decontextualized measures of cognitive ability seem promising with regard to the diversity-validity dilemma. However, more studies are needed before we can make any conclusive remarks on their impact on the reduction of ethnic subgroup differences and their criterion-related validity coefficients.

Use Simulations as Predictors

In their review article concerning strategies to reduce subgroup differences, Ployhart and Holtz (2008; but also other authors, see: Hough et al., 2001; Sackett et al., 2001) explicitly mentioned simulation exercises as one of the best tactics to

avoid or decrease adverse impact. Simulations, such as assessment centers, work samples, and situational judgment tests (SJTs), refer to selection instruments wherein applicants perform exercises that physically and/or psychologically resemble those tasks to be performed on the job (Motowidlo, Dunnette, & Carter, 1990). Assessment centers and work samples are performance assessments that demand applicants to carry out job-related assignments (e.g., setting up a work planning, reprimanding an unmotivated employee, developing and presenting a strategic business plan), whereas SJTs refer to paper-and-pencil or video-based simulations in which applicants are confronted with descriptions of job-related dilemmas and are required to select the most appropriate response out of a set of predetermined options (Motowidlo et al., 1990). Generally, simulations demonstrate adequate criterion-related validity (see Lievens & De Soete, 2012 for an overview). In terms of ethnic subgroup differences in test performance, meta-analytic research has proven that simulations display lower subgroup differences than cognitive ability tests. For assessment centers, Dean, Bobko, and Roth (2008) demonstrated standardized Black-White subgroup differences of 0.52, with White test-takers systematically obtaining higher scores than Blacks. Roth, Huffcutt, and Bobko (2003) found similar effect sizes for work samples (d=0.52), whereas other studies found d-values ranging from 0.70 to 0.73 (Bobko, Roth, & Buster, 2005; Roth, Bobko, McFarland, & Buster, 2008). Finally, Whetzel, McDaniel, and Nguyen (2008) noted small to moderate ethnic subgroup differences in SJT performance (d=0.24-0.38). For nearly all types of simulations, cognitive load (defined as the correlation between test scores on the simulation exercise and cognitive ability test scores, Whetzel et al., 2008) has proven to be the most influencial driver of subgroup differences (for assessment centers: Goldstein, Yusko, Braverman, Smith, & Chung, 1998; Goldstein, Yusko, & Nicolopoulos, 2001; for work samples: Roth et al., 2008; for SJTs: Whetzel et al., 2008).

When critically reviewing these findings, we can draw two conclusions. First, although ethnic subgroup differences on simulations are significantly smaller than those on cognitive ability tests, they are in most cases still substantial (e.g., Roth et al., 2008). As this implies that the risk for adverse impact remains when adding simulations to the selection procedure, continued efforts should be undertaken to develop simulation instruments with minimal adverse impact, while maintaining good validity. Second, our conceptual knowledge of the underlying mechanisms that cause or reduce ethnic subgroup differences in simulation performance is rather limited. As most prior studies in this domain have approached simulations as holistic entities (Arthur & Villado, 2008), alterations in the magnitude of subgroup differences could not be attributed to specific construct or method factors. Fortunately, an increasing number of studies is advocating a more systematic and theory-driven approach for examining subgroup differences in simulation performance (e.g., Arthur, Day, McNelly, & Edens, 2003; Arthur & Villado, 2008; Chan & Schmitt, 1997; Edwards & Arthur, 2007; Lievens, Westerveld, & De Corte, in press). Simulations should be treated as a combination of predictor

constructs, which refer to the behavioral domain being sampled, and predictor methods, which denote the specific techniques by which domain-relevant behavioral information is elicited, collected, and subsequently used to make inferences (Arthur & Villado, 2008, p. 435; Lievens et al., in press). Researchers are recommended to keep specific factors constant when manipulating either predictor constructs or predictor methods in order to increase our theoretical knowledge on the nature of subgroup differences, thereby advocating a "building block" approach rather than a holistic approach. Along these lines, researchers have identified "fidelity" as a key method factor (i.e., building block) of ethnic subgroup differences in performance on simulation-based instruments in addition to cognitive load, which is a key construct factor (e.g., Chan & Schmitt, 1997; Ployhart & Holtz, 2008).

Fidelity can be defined as the extent to which the test situation resembles the actual job situation (Callinan & Robertson, 2000). It can be divided in stimulus fidelity on the one hand, which refers to the fidelity of the presented stimulus material and response fidelity on the other hand, which refers to the fidelity of how participants' responses are collected. For example, when selecting applicants for sales functions, high stimulus and response fidelity are obtained by using video (as opposed to paper-and-pencil) fragments of client interactions and requiring oral (as opposed to written) responses. Regarding the stimulus side, Chan and Schmitt (1997) compared a written SJT to a content-wise identical video SJT, and found significantly lower subgroup differences in performance on the latter variant. Similarly, other studies have demonstrated lower subgroup differences on high stimulus fidelity simulations as compared to lower stimulus fidelity formats, but failed to keep test content and other factors constant (e.g., Schmitt & Mills, 2001; Weekley & Jones, 1997).

On the response side, fidelity has often been neglected as a potential factor of diversity in selection (e.g., Ryan & Greguras, 1998; Ryan & Huth, 2008). Arthur et al. (2002), and Edwards and Arthur (2007) showed that higher fidelity (constructed response or open-ended) response formats generated lower ethnic performance differences than their low fidelity (multiple choice) counterparts. Arthur and colleagues examined this for knowledge tests, but recent studies also investigated the effect of response fidelity on simulations. De Soete, Lievens, Oostrom, and Westerveld (2012) focus on ethnic subgroup differences in performance on constructed response multimedia tests. A technology-enhanced constructed response multimedia test presents applicants with video-based job-related scenes, with a webcam capturing how they act out their response as if they actually take place in the presented situation (De Soete et al., 2012; Lievens et al., in press; Oostrom, Born, Serlie, & van der Molen, 2010, 2011). Preliminary results with regard to diversity have been promising, with constructed response multimedia tests displaying lower ethnic subgroup differences than other commonly used instruments (De Soete et al., 2012).

In sum, employing simulations has demonstrated to be a fruitful strategy in light of the diversity-validity dilemma. Mainly instruments that are characterized by low cognitive load on the one hand and high stimulus as well as response fidelity on the other hand have shown to be effective in reducing ethnic performance differences without impairing criterion-related validity. In an attempt to close the gap between selection research and the fast evolving simulation practice, future research should focus on examining the efficacy of other high stimulus and response fidelity formats that have already entered the practitioner field. On the stimulus side, a number of researchers (e.g., Fetzer, 2012) made the first step in this direction by examining the validity and diversity effects of several innovative computer-based simulation exercises. As broadband internet connectivity becomes more prevalent and multimedia tools become easier to deploy, cartoon SJTs, two-dimensional and three-dimensional graphic simulations, avatar exercises, serious games, and multimedia instruments are increasingly becoming part of the simulation portfolio. Initial results regarding their ethnic subgroup differences and validity have been promising (Fetzer, 2012), but thorough research on their effectiveness is needed.

On the response side, efforts should be made towards comparing higher and lower fidelity formats such as open-ended versus multiple choice response modalities. Furthermore, increased research attempts should be devoted to the discovery of underlying mechanisms of ethnic subgroup differences in simulation performance. Whereas earlier research has already pointed in the direction of cognitive load, fidelity, reading demands (Chan & Schmitt, 1997) and test perceptions (Chan & Schmitt, 1997; Edwards & Arthur, 2007), we need more research on the influence of culture-related communication and response styles (e.g., Gudykunst et al., 1996; Helms, 1992) that may (dis)advantage certain cultural groups according to the employed stimulus or response format. That is, non-Western cultures have been posited to adhere more value on orality, movement, and behavioral aspects, and use more high-context (non-verbal) communication styles as compared to Western societies (Gudykunst et al., 1996; Hall, 1976; Helms, 1992). These cultural differences suggest that test-takers may have more opportunity to express themselves and perform maximally in some response modalities than in others depending on their ethnicity. Finally, culture-based preferences or dispositions for divergent thinking ("There are multiple answers for each problem") versus convergent thinking ("There is only one correct answer", Guilford, 1950) may also influence the effectiveness of test-takers' response strategies (Outtz, Goldstein, & Ferreter, 2006).

Reduce Construct Irrelevant Variance

In their review article on strategies to reduce ethnic subgroup differences in selection performance, Ployhart and Holtz (2008) recommended to reduce all forms of irrelevant variance caused by the predictor measure. Irrelevant construct variance

denotes variance caused by predictor demands that are not related to the criterion measure (i.e., effective future job performance). Irrelevant test demands may be correlated with ethnicity and therefore generate subgroup differences that are unrelated to actual performance differences. In order to increase diversity, it is recommended to ban all irrelevant test demands from the selection procedure. In the current paper, we review the literature and express recommendations on two potential variants of construct irrelevant test variance, i.e., verbal load and cultural load.

An instrument's verbal load can be defined as the extent to which the predictor requires verbal capacities in order to perform effectively (Ployhart & Holtz, 2008). Several studies have demonstrated that ethnic minorities score systematically lower on several measures of verbal ability (Hough et al., 2001). Substantial ethnic subgroup differences have been found in performance on reading comprehension tasks (Barrett, Miguel, & Doverspike, 1997; Sacco et al., 2000). In their meta-analysis on subgroup differences in employment and educational settings, Roth and colleagues (2001) found *d*-values for verbal ability tests ranging from 0.40 (for Hispanic-White comparisons) to 0.76 (for Black-White comparisons), with White respondents receiving higher test scores. In addition, De Meijer, Born, Terlouw, and Van Der Molen (2006) demonstrated that score differences between ethnic minority and ethnic majority applicants on several selection instruments could partly be attributed to (a lack of) language proficiency. These findings confirm that it is advisable to limit the verbal demands of selection instruments strictly to the extent that they are required based on job analysis (Arthur et al., 2002; Hough et al., 2001; Ployhart & Holtz, 2008). As we already discussed, Chan and Schmitt (1997) provided a good example of this strategy by comparing ethnic subgroup differences on a written SJT with a content-wise identical video SJT. The video SJT displayed significantly lower subgroup differences, which was partly caused by the lower reading requirements of the video format as compared to the written format. Along the same lines, several testing firms experiment with cutting-edge technologies to develop selection instruments that impose lower reading demands. Examples are the application of cartoons, two-dimensional and three-dimensional graphic animations, and webcam testing (e.g., Fetzer, 2012; Lievens et al., in press; Oostrom et al., 2010, 2011). Given their low reading and writing demands, technology-enhanced stimulus and response formats seem fruitful avenues for further investigation with regard to the diversity-validity dilemma.

The second strategy to eliminate irrelevant test variance concerns reducing the cultural load of selection instruments. The rationale behind this approach is that test-takers from different cultures adhere to different values, interpretations, and actions, whereby the test-taker's cultural background may differentially influence test performance regardless of the individual's actual capabilities. In fact, Helms (1992, 2012) has repeatedly argued that most selection instruments are developed against a majority cultural background. That is, test developers who belong to the

ethnic majority group use their own culture as a reference framework when creating selection instruments. As this may systematically disadvantage respondents that do not belong to the ethnic majority culture, an option might be to develop culturally equivalent instruments (e.g., Helms, 1992). As opposed to the majority of selection tools, culturally equivalent instruments do not invoke interpretation discrepancies or performance (dis)advantages that are related to ethnicity instead of the capabilities measured (Helms, 1992).

Several attempts have been undertaken to develop instruments that do not impose irrelevant cultural demands. Some researchers tried to incorporate the minority test-takers' culture (by presenting Blacks with cognitive items in a social context in order to integrate their emphasis on social relations, DeShon, Smith, Chan, & Schmitt, 1998), whereas others aimed to develop so-called 'culture-free' measurement tools ('fluid intelligence measures', Cattell, 1971). Unfortunately, thus far these approaches have not resulted in substantial reductions of ethnic subgroup differences. Other techniques to reduce cultural inequity in selection instruments are sensitivity review panels and cognitive interviewing. Reckase (1996) was the first to recommend the use of sensitivity review panels when constructing tools for culturally diverse groups. Subject matter experts regarding cultural groups (and sensitivities) are asked to review all items and evaluate them on their (in)sensitivity towards certain ethnic groups. Potentially offensive stereotypes or expressions are removed to increase test fairness. A similar technique, which requires the cooperation of actual test-takers, is called cognitive interviewing (Beatty & Willis, 2007). Cognitive interviewing can be defined as a qualitative technique to review tests and questionnaires. During the interview, test-takers are asked to think aloud while responding to items or to provide answers to additional questions (e.g., how do you interpret the instructions and items, and what are potential difficulties or ambiguities perceived while completing the instrument). On top of identifying problematic items that generate ethnic subgroup differences, this technique also provides additional insight in the underlying factors of these ethnic score discrepancies. Cognitive interviewing has already proven its merits in the health sector, by increasing the conceptual equivalence of medical questionnaires for ethnically diverse groups (e.g., Nápolez-Springer, Santoyo-Olsson, O'Brien, & Stewart, 2006; Willis & Miller, 2011). In personnel selection, the application of cognitive interviewing is rather scarce. Oostrom and Born (2012) used the technique to discover differences in interpretation between ethnic majority and minority test-takers when conducting a role-play. Results demonstrated that several interpretation discrepancies could be identified, which were mostly explained by differences in language proficiency and attribution styles. However, findings on the effectiveness of the technique with regard to the reduction of subgroup differences are still lacking.

The last tactic to reduce irrelevant cultural predictor variance concerns the identification and removal of culturally biased items through differential item

functioning (DIF, Berk, 1982). The goal of DIF is to detect those items that lead to poorer performance of minority group test-takers as compared to evenly competent majority group test-takers (Sackett et al., 2001). Mostly unfamiliar or verbally difficult items are the focus of attention. Several past studies have found evidence for DIF in tests used in high-stakes contexts (Freedle & Kostin, 1990; Medley & Quirk, 1974; Whitney & Schmitt, 1997). More recently, Scherbaum and Goldstein (2008) found evidence for differentially functioning items in standardized cognitive tests. Imus et al. (2011) successfully applied the DIF technique to biodata employment items in an ethnically diverse sample. Mitchelson, Wicher, LeBreton, and Craig (2009) revealed DIF on 73% of the items of the Abridged Big Five Circumplex of personality traits. Despite some promising results, it should be noted that this technique also has its limitations. The usefulness of DIF is criticized by some scientists due to its unknown effects on validity (Sackett et al., 2001). Furthermore, studies have found little evidence of easily interpretable results, as there are no available theoretical explanations for DIF effects (e.g., Imus et al., 2011; Roussos & Stout, 1996). Finally, differentially functioning items that disadvantage ethnic majority members have regularly been found as well (Sackett et al., 2001). In general, it can be concluded that DIF is often observed in two directions, thereby (dis)advantaging ethnic majority members equally to ethnic minority members. Consequently, its effect on diversity seems limited.

In sum, as verbal and cultural predictor requirements may enhance ethnic subgroup differences in test performance, it is strongly recommended to limit a predictor's verbal and cultural demands strictly, to the extent that they are required for effective job performance. Using video and multimedia during test administration has demonstrated to be effective in reducing the instrument's reading and writing demands and lowering the associated ethnic subgroup differences. To reduce a predictor's cognitive load, cognitive interviewing and DIF have been suggested as promising techniques. However, future studies are needed to examine the underlying causes of DIF, the concrete effects of removing differentially functioning items or cognitive interviewing on the magnitude of ethnic subgroup differences and validity, and the combined effect of DIF and cognitive interviewing.

Provide Coaching Programs and Opportunity for Practice

The fourth strategy to consider for organizations that attempt to alleviate the diversity-validity dilemma is providing coaching programs and opportunity for practice and re-testing to candidates (Ployhart & Holtz, 2008). The underlying assumption is that ethnic subgroups differ in their test familiarity and therefore have differential test taking skills, leading ethnic minority test-takers to perform more poorly in some cases (Sackett et al., 2001). Organizing practice opportunities and offering the possibility to retake the assessment should then allow test-takers to familiarize themselves with the test content and testing situation. Coaching

programs go even one step further and intensively guide potential applicants through the selection process while teaching them test-taking strategies and featuring rigorous exercising. Several studies have demonstrated that practice, retesting, and coaching have small albeit consistently positive effects on test performance (Sackett, Burris, & Ryan, 1989; Sackett et al., 2001). This finding was recently replicated in a meta-analysis by Hausknecht, Halpert, Di Paolo, and Moriarty Gerrard (2007). Results of 107 samples did not only reveal a consistent effect of practice and coaching on subsequent performance, but also specified that a combined approach of coaching and practice leads to the most beneficial results in terms of performance increase.

However, the effect of practice, retesting, and coaching on diversity and criterion-related validity is less straightforward (Hough et al., 2001; Sackett et al., 2001). In most cases, both ethnic majority and minority group test-takers benefit from practice and coaching (Sackett et al., 2001). Schleicher, Van Iddekinge, Morgeson, and Campion (2010) added some important insights to this stream of research by comparing ethnic performance differences after retesting for different types of assessment tools. In general, Whites benefitted more from retesting than Blacks and this effect held stronger for written tests as compared to sample-based tests. Results showed that the effect of retesting on adverse impact ratios was highly dependent on the measure used, so that retesting on sample-based tests could enhance diversity, whereas retesting in the case of written tests had the potential to increase adverse impact (Schleicher et al., 2010). In terms of criterion-related validity, Van Iddekinge, Morgeson, Schleicher, and Campion (2011) discovered no negative influence of retesting.

In short, practice, retesting, and coaching have demonstrated to have small but consistently positive effects on performance, and in some cases a modest reduction in ethnic subgroup differences as a result of these techniques has been observed. To extend our knowledge on coaching and practice effects in light of the diversity-validity dilemma, additional research is required on the moderating variables that trigger retesting and coaching influences. First, it is imperative to investigate the role of test attitudes, test motivation, and perceptions of procedural fairness (Schleicher et al., 2010). Although these mechanisms have been demonstrated to influence learning performance (e.g., Sackett et al., 1989) and the magnitude of subgroup performance differences (e.g., Ryan, 2001), they have not been examined in the context of retesting as a strategy for reducing adverse impact. Second, future studies should differentiate retesting and coaching effects according to the constructs of interest. Up until now, several studies have focused on retesting for cognitive skills (Sackett et al., 2001), which are known to produce substantial ethnic subgroup differences. It would be interesting to investigate whether the retesting strategy is more effective to reduce subgroup differences when focusing on social or interpersonal skills (Roth, Buster, & Bobko, 2011).

Foster Positive Test-Taker Reactions

The fifth strategy to approach the diversity-validity dilemma concerns fostering positive test-taker reactions (Ployhart & Holtz, 2008). The idea is that applicant perceptions may differ across ethnic subgroups, so that ethnic minority group test-takers experience less positive test perceptions which may negatively influence their performance. It is therefore suggested that undertaking interventions to increase positive test perceptions among test-takers in general and ethnic minorities in particular may reduce ethnic subgroup differences in withdrawal intentions and selection test performance (Hough et al., 2001; Ployhart & Holtz, 2008; Sackett et al., 2001). Changing test perceptions can be achieved by altering the selection test's instructional sets or by modifying the items and test format (Sackett et al., 2001).

Several studies have been devoted to the relation between test perceptions and ethnicity. For example, Arvey, Strickland, Drauden, and Martin (1990) were among the first to demonstrate that motivational differences across ethnic subgroups do exist. They found that Whites reported higher test motivation and more belief in selection testing than Blacks, which was related to their performance on ability tests and work sample exercises. Other studies also noticed significantly lower test-taking motivation and higher test anxiety among Black test-takers as compared to Whites (e.g., Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Schmit & Ryan, 1997, but see for an exception: Becton, Feild, Giles, & Jones-Farmer, 2008). In an attempt to apply this knowledge, Chan and Schmitt (1997) compared test perceptions and subgroup differences on a paper-and-pencil SJT and an identical video-based SJT. As has been mentioned before, ethnic subgroup differences were significantly smaller on the latter, which was due to lower reading demands on the one hand and smaller ethnic subgroup differences in face validity perceptions on the other hand. In a similar vein, Edwards and Arthur (2007) demonstrated that lower ethnic subgroup differences on a constructed response knowledge test as compared to a multiple choice variant could be partly attributed to lower subgroup differences in perceived fairness and test-taking motivation.

Another important line of research within this domain focuses on the effect of stereotype threat on the magnitude of ethnic subgroup differences. Stereotype threat comprises the idea that the mere knowledge of cultural stereotypes may affect test-performance (e.g., Steele, 1997, 1998; Steele & Aronson, 1995, 2004). Accordingly, if ethnic minority group test-takers are made aware of negative stereotypes regarding ethnicity and selection test performance, this is suggested to deteriorate their performance (Steele & Aronson, 1995). Steele was the first to propose this hypothesis as an explanation for ethnic differences in performance. He demonstrated that when members of ethnic minority groups enter high stakes testing situations and when they are made aware of the commonly found ethnic group discrepancies, concerns to accomplish poorly arise and performance suffers

(Steele, 1997; Steele & Aronson, 1995). In 2008, Nguyen and Ryan conducted a meta-analysis that demonstrated support for a modest effect of stereotype threat among ethnic minority group test-takers. The size of the effect was a function of the explicitness of the stereotype-activating cues, with moderate cues displaying larger effects than blatant or subtle cues (Nguyen & Ryan, 2008). Another moderator that emerged from earlier studies concerns the extent to which the minority test-taker identified with the domain measured. That is, stereotype threat only occurred for those individuals who regard the test domain as relevant for their self-image (Steele & Aronson, 1995). Despite its popularity in certain lines of research, Steele's hypothesis on stereotype threat has been severely criticized. Sackett has repeatedly expressed his concerns about Steele's research methods and about the misinterpretation and overgeneralization of his research findings (e.g., Sackett, 2003; Sackett, Hardison, & Cullen, 2004; Sackett et al., 2001). In addition, other studies failed to replicate stereotype threat effects (e.g., Cullen, Hardison, & Sackett, 2004; Gillespie, Converse, & Kriska, 2010; Grand, Ryan, Schmitt, & Hmurovic, 2011), thereby questioning the strength of the phenomenon.

To conclude, it seems that fostering positive test-taker perceptions may have in some cases positive albeit small influences on diversity. Additionally, it can enhance the organizational image among potential employees (Ployhart & Holtz, 2008). The most promising strategy in this category regards altering the test format (and accordingly also test fidelity) in order to obtain higher face validity perceptions and test motivation. Further research is needed to shed light on the impact of test-taker perceptions on minority applicant withdrawal (e.g., Schmit & Ryan, 1997; Tam, Murphy, & Lyall, 2004). Regarding the phenomenon of stereotype threat, more research is required to explore to which extent the hypothesis holds in actual applicant situations and which factors perform as moderating influences (e.g., Sackett, 2003).

Use Statistical Approaches

The last category of techniques that deal with the diversity-validity dilemma refers to a number of statistical methods to combine and manipulate selection predictor scores, such as adding non-cognitive predictors to cognitive ones, together with explicit predictor weighting, criterion weighting, and score banding (Ployhart & Holtz, 2008). The following reviews and updates the evidence regarding the effectiveness of these approaches.

The first strategy within this category makes use of non-cognitive predictors that exhibit smaller subgroup differences than cognitive predictors, and combines them with a cognitive predictor into a weighted sum, called a predictor composite score. This is also known as a compensatory strategy, because lower scores on one predictor can be compensated for by higher scores on other predictors. Sackett and Ellingson (1997) have proposed several formulas to estimate the effect size resulting from the combination of predictors with different effect sizes in an equally

or differently weighted composite score. By systematically varying the factors underlying the effect size of the predictor composites (e.g., the intercorrelation of the original predictors), researchers and practitioners can evaluate the potential consequences of different approaches to predictor selection and combination. In a related vein, De Corte, Lievens, and Sackett (2006) described an analytic method that evaluates the outcomes of single- and multi-stage selection decisions in terms of adverse impact and selection quality, as a result of the order in which the predictors are administered (either in the early or later stages of the selection process), and the selection rates at the different stages. Single stage selection decisions are taken after all predictors are administered, whereas multi-stage (or multiple hurdle) selection decisions administer the predictors in several different stages, with only the applicants obtaining a high enough score in one stage, passing to the subsequent stage(s). Although the proposed tools could be used to pursue the development of a set of guidelines for the design of multi-stage selection scenarios that optimize adverse impact and the selection quality, De Corte et al. (2006) and other authors (Sackett & Roth, 1996) warn against such a quest by stating that there are no simple rules to approach hurdle based selection. An illustration of the implied dangers of formulating such rules is provided in a paper by Roth, Switzer, Van Iddekinge, and Oh (2011), who demonstrate that the projected effects on the average level of job performance and adverse impact ratio (i.e., the ratio of the selection rate of the lower scoring applicant subgroup and the selection rate of the higher scoring applicant subgroup, oftentimes used as a measure of adverse impact, AIR) of multiple hurdle selection systems heavily depend on the input values that are used.

In order to rationally develop the weights that are assigned to the elementary predictors (also called predictor weights) to develop predictor composites, De Corte, Lievens, and Sackett (2007, 2008) and De Corte, Sackett, and Lievens (2011) proposed decision aids that can be applied to optimize both adverse impact and the quality of (multi-stage) selection decisions. The proposed decision aids focus at employers that plan selection decisions based on an available set of predictors, and determine the Pareto-optimal predictor weights that lead to Pareto-optimal trade-offs between selection quality and diversity. A specific weighing scheme and corresponding trade-off is called Pareto-optimal when the level on one outcome value (i.e., quality) cannot be improved without doing worse on the other outcome (i.e., AIR). The regression-based predictor composite is one particular Pareto-optimal trade-off, and no other weighing of the predictors can outperform this composite in terms of expected selection quality. However, other Pareto-optimal trade-offs revealed by the decision aid show a more balanced trade-off between the outcomes so that they imply a higher level of AIR than the regression-based composite, for a concession in terms of quality. Furthermore, a similar decision aid was proposed for facilitating decision making in complex selection contexts (Druart & De Corte, 2012). Complex selection decisions handle situations with an applicant pool, several open positions, and applicants that are interested in

at least one of the positions under consideration. Such situations can be encountered in large organizations (e.g., the military) and as admission decisions in educational contexts. Further research should go into the design of more user-friendly decision aids, and user reactions concerning these tools, as suggested by Roth et al. (2011).

Third, the approach that weights different predictors and combines them into a predictor composite score, can be applied to criterion measures as well, and is then called criterion weighting (Ployhart & Holtz, 2008). Criterion weighting is based on the multidimensionality of the criterion space by taking into account task, contextual, and counterproductive behavior. The relative weights assigned to these different criterion dimensions may suggest using alternative weights for the cognitive and non-cognitive predictors within the predictor composite, thereby affecting the minority representation as shown by Hattrup, Rock, and Scalia (1997) and De Corte (1999). In line with the method to obtain Pareto-optimal trade-offs between selection quality and diversity (De Corte et al, 2007), where the amount of selection quality a decision maker indulges to obtain a more favorable AIR is a value issue, criterion weighting reflects an organization's values about the different job performance dimensions. As it is rarely the case that an organization's goal is univariate and thus only considers the maximization of task performance (Murphy, 2010; Murphy & Shiarella, 1997), criterion weighting seems to be a promising method to alleviate the diversity-validity dilemma. However, research that clearly evaluates its merits is scant thus far.

Finally, the last strategy we review within the category of statistical techniques is score 'banding' (Cascio, Outtz, Zedeck, & Goldstein, 1991). Banding involves grouping the applicant test scores within given ranges or bands, and treating the scores within a band as equivalent. The width of the bands is based on the standard error of the difference between scores, and reflects the unreliability in the interpretation of scores. Selection within bands then happens on the basis of other variables that show smaller subgroup differences (Campion et al., 2001). However, banding is controversial, due to several contradictions in the rationale behind this technique. For example, while banding seems to be effective in reducing adverse impact only when using racioethnic minority preferences to select or break ties within a band (Ployhart & Holtz, 2008), this approach is prohibited by law (Cascio, Jacobs, & Silva, 2010). Future research should further investigate other methods than the classical test theory for computing bands, such as item response theory (see Bobko, Roth, & Nicewander, 2005).

Table 2. *Overview of Strategies and their Effectiveness for Dealing with the Diversity-Validity Dilemma*

Strategies per Category	Examples and References	Effectiveness
1. Use alternative cognitive ability measures		
Logic-based cognitive measurement instruments	Siena Reasoning Test (Yusko et al., 2012)	effective
Improve point-to-point correspondence of cognitive predictor with criterion	Typical instead of maximal cognitive performance measures (Ackerman & Beier, 2012)	unknown
	Contextualizing cognitive ability measures (Hattrup et al., 2002)	unknown
Measure specific cognitive abilities	Executive functioning (Huffcutt et al., 2012)	unknown
2. Use simulation exercises as predictors		
Employ assessment centers, work samples, and SJTs	(e.g., Dean et al., 2008; Lievens & De Soete, 2012; Roth et al., 2003; Whetzel et al., 2008)	effective
Increase stimulus fidelity	High stimulus fidelity simulations (e.g., Chan & Schmitt, 1997)	effective
Increase response fidelity	High response fidelity simulations (e.g., De Soete et al., 2012)	effective
Invest in technology-enhanced simulations (graphics, avatars, serious games)	(e.g., Fetzer, 2012)	unknown
3. Reduce construct irrelevant variance		
Reduce verbal load	(e.g., Chan & Schmitt, 1997)	effective
Reduce cultural load	Sensitivity panels (Reckase, 1996)	unknown
	Cognitive interviewing (Beatty & Willis, 2007)	unknown
	Detecting and removing differentially functioning items (e.g., Imus et al., 2011; Sackett et al., 2001; Scherbaum & Goldstein, 2008)	mixed results
4. Provide coaching programs and opportunity for practice		
Provide practice, retesting, and coaching	(e.g., Hausknecht et al., 2007; Schleicher et al., 2010)	mixed results
5. Foster positive test-taker reactions		
Alter test perceptions by increasing fidelity	Increase perceptions of face validity, perceived fairness, and test-taking motivation (Chan & Schmitt, 1997; Edwards & Arthur, 2007)	small effects
Influence the effect of stereotype threat	(e.g., Steele, 1997, 1998; Steele & Aronson, 1995)	mixed results
6. Use statistical approaches to combine and manipulate scores		
Combine different predictor scores into predictor composite	(e.g., De Corte et al., 2006; Sackett & Ellingson, 1997)	effective
Pareto-optimal predictor composites	(e.g., De Corte et al., 2007, 2008; Druart & De Corte, 2012)	effective
Criterion weighting	(e.g., Hattrup et al., 1997; De Corte, 1999)	unknown
Banding	(e.g., Campion et al., 2001; Cascio et al., 2010)	controversial

Discussion

Main Conclusions

The current paper aims to provide an updated overview of strategies to deal with the diversity-validity dilemma (see also Table 2). The conclusion of our review is that there is no easy way to tackle the dilemma. In most cases, organizations have to make a trade-off between several organizational stakes such as performance goals, financial goals, corporate social responsibility aims, employer branding, and diversity. However, a number of strategies have emerged as particularly useful in the context of the diversity-validity dilemma. First, employing logic-based measurement methods to capture cognitive skills seems to be a new and fruitful strategy to reduce ethnic subgroup differences and at the same time identify applicants with job-relevant reasoning capabilities. Second, it seems worthwhile to increase the response fidelity of (simulation) instruments as this does not only enhance the point-to-point matching between predictor and criterion, but also lowers ethnic subgroup differences in some instances. Third, investments in advanced assessment technologies seem to pay off. In fact, initial research findings suggest lower ethnic subgroup differences and good criterion-related validity coefficients for several new multimedia simulations such as cartoon SJTs, graphic simulations, and serious games. Simultaneously, these instruments possess low reading demands, provide the possibility for exercise by means of practice items, and are well received by both ethnic majority group as well as ethnic minority group test-takers – thereby responding to several strategies for reducing ethnic performance differences. Fourth, statistical strategies that take into account workforce diversity as one of the primary goals of selection decisions, besides selection quality, seem to hold promise and gain importance. In particular, decision aids that result in Pareto-optimal trade-offs between selection quality and diversity, in single- and multi-stage, as well as in complex selection situations, have emerged as a particularly effective approach to balance the different outcomes of selection decisions.

Future Research Directions

Across the various avenues for further research that have been pointed out in the current manuscript, we make the following key suggestions for future studies. First, the research domain is in dire need of a more systematic operationalization of ethnicity or race. Currently, there is no consensus on appropriate labels or terminology for certain ethnic groups (Foldes, Duehr, & Ones, 2008). As a result, researchers use their own interpretation of ethnicity, thereby complicating the generalizability of research findings. Second, we strongly recommend researchers to apply the building block approach to study subgroup differences on simulation exercises. It is advised to keep specific factors of the simulation constant when

manipulating either predictor constructs or predictor methods, as this permits us to increase our knowledge of the theoretical drivers of diversity. Third, research on ethnic subgroup differences in personnel selection would greatly benefit from cross-fertilization with other psychology branches. That is, cross-cultural psychology offers research methodologies that could easily be applied to the research domain of adverse impact (Leong, Leung, & Cheung, 2010) and a number of scientists have stressed the potential of including social psychological theories to the study of ethnic score differences in order to take situational and environmental factors into account (Helms, 2012).

References

Ackerman, P.L., & Beier, M.E. (2012). The problem is in the definition: *G* and intelligence in I-O psychology. *Industrial and Organizational Psychology-Perspectives on Science and Practice, 5*, 149-153.

Arthur, W., Day, E.A., McNelly, T.L., & Edens, P.S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125-154.

Arthur, W., Edwards, B.D., & Barrett, G.V. (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.

Arthur, W., & Villado, A.J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435-442.

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

Barrett, G.V., Miguel, R.F., & Doverspike, D. (1997). Race differences on a reading comprehension test with and without passages. *Journal of Business and Psychology, 12*, 19-24.

Beatty, P.C., & Willis G.B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*, 287-311.

Becton, J.B., Feild, H.S., Giles, W.F., & Jones-Farmer, A. (2008). Racial differences in promotion candidate performance and reactions to selection procedures: A field study in a diverse top-management context. *Journal of Organizational Behavior, 29*, 265-285.

Berk, R.A. (1982). (Ed.). *Handbook of methods for detecting test bias.* Baltimore: John Hopkins University Press.

Bobko, P., Roth, P.L., & Buster, M.A. (2005). Work sample selection tests and expected reduction in adverse impact: A cautionary note. *International Journal of Selection and Assessment, 13*, 1-10.

Bobko, R., Roth, P.L., & Nicewander, A. (2005). Banding selection scores in human resource management decisions: Current inaccuracies and the effect of conditional standard errors. *Organizational Research Methods, 8*, 259-273.

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

Brouwers, S.A., & Van De Vijver, F.J.R. (2012). Intelligence 2.0 in I-O psychology: Revival or contextualization? *Industrial and Organizational Psychology-Perspectives on Science and Practice, 5*, 158-160.

Callinan, M., & Robertson, I.T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8*, 248-260.

Campion, M.A., Outtz, J.L., Zedeck, S., Schmidt, F.L., Kehoe, J.F., Murphy, K.R., & Guion, R.M. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology*, *54*, 149-185.

Cascio, W.F., Jacobs, R., & Silva, J. (2010). Validity, utility, and adverse impact: Practical implications from 30 years of data. In J.L. Outtz (Ed.), *Adverse impact. Implications for organizational staffing and high stakes selection* (pp. 271-288). New York, NY: Routledge.

Cascio, W.F., Outtz, J., Zedeck, S., & Goldstein, I.L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4*, 233-264.

Cattell, R.B. (1971). *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin.

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

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

Cohen, J. (1994). The earth is round (p less than .05). *American Psychologist, 49*, 997-1003.

Cullen, M.J., Hardison, C.M., & Sackett, P.R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89*, 220-230.

Dean, M.A., Bobko, P., & Roth, P.L. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology, 93*, 685-691.

De Corte, W. (1999). Weighing 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.

De Corte, W., Lievens, F., & Sackett, P. (2006). Predicting adverse impact and multistage mean criterion performance in selection. *Journal of Applied Psychology, 91*, 523-537.

De Corte, W., Lievens, F., & Sackett, P. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92*, 1380-1393.

De Corte, W., Lievens, F., & Sackett, P. (2008). Validity and adverse impact potential of predictor composite formation. *International Journal of Selection and Assessment, 16*, 183-194.

De Corte, W., Sackett, P., & Lievens, F. (2011). Designing pareto-optimal selection systems: Formalizing the decisions required for selection system development. *Journal of Applied Psychology, 96*, 907-926.

De Meijer, L.A.L., Born, M.P., Terlouw, G., & van der Molen, H.T. (2006). Applicant and method factors related to ethnic score differences in personnel selection: A study at the Dutch police. *Human Performance, 19*, 219-251.

De Soete, B., Lievens, F., Oostrom, J.K., & Westerveld, L. (2012, July). Alternative predictors in personnel selection: Constructed response multimedia tests vs. other instruments. In J. Fontaine & E. Derous (Chairs), *Reducing bias and the achievement gap of minorities in selection procedures in the Low Countries*. Symposium conducted at the International Test Commission, Amsterdam, NL.

DeShon, R.P., Smith, M.R., 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.

Druart, C., & De Corte, W. (2012). Designing pareto-optimal systems for complex selection decisions. *Organizational Research Methods, 15*, 488-513.

Edwards, B.D., & Arthur, W. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology, 92*, 794-801.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). *Uniform guidelines on employee selection procedures*: 29 C.F.R. 1607.

Fetzer, M.S. (2012, April). *Current research in advanced assessment technologies*. Symposium presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Foldes, H.J., Duehr, E.E., & Ones, D.S. (2008). Group differences in personality: Meta-analyses comparing five US racial groups. *Personnel Psychology, 61*, 579-616.

Freedle, R., & Kostin, I. (1990). Item difficulty of 4 verbal item types and an index of differential item functioning for black and white examinees. *Journal of Educational Measurement, 27*, 329-343.

Gatewood, R., Feild, H., & Barrick, M. (2011). *Human resource selection (6th ed.)*. Cincinnati, OH: South-Western.

Gillespie, J.Z., Converse, P.D., & Kriska, S.D. (2010). Applying recommendations from the literature on stereotype threat: Two field studies. *Journal of Business and Psychology, 25*, 493-504.

Goldstein, H.W., Yusko, K.P., Braverman, E.P., Smith, D.B., & 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, H.W., Yusko, K.P., & Nicolopoulos, V. (2001). Exploring black-white subgroup differences of managerial competencies. *Personnel Psychology, 54*, 783-807.

Grand, J.A., Ryan, A.M., Schmitt, N., & Hmurovic, J. (2011). How far does stereotype threat reach? The potential detriment of face validity in cognitive ability testing. *Human Performance, 24*, 1-28.

Guilford, J.P. (1950). Creativity. *American Psychologist, 5*, 444-445.

Gudykunst, W.B., Matsumoto, Y., TingToomey, S., Nishida, T., Kim, K., & Heyman, S. (1996). The influence of cultural individualism-collectivism, self construals, and individual values on communication styles across cultures. *Human Communication Research, 22*, 510-543.

Hall, E.T. (1976). *Beyond culture.* New York, NY: Doubleday.

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.

Hattrup, K., Schmitt, N., & Landis, R.S. (1992). Equivalence of constructs measured by job-specific and commercially available aptitude-tests. *Journal of Applied Psychology, 77*, 298-308.

Hausknecht, J.P., Halpert, J.A., Di Paolo, N.T., & Moriarty Gerrard, M.O.M. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373-385.

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

Helms, J.E. (2012). A legacy of eugenics underlies racial-group comparisons in intelligence testing. *Industrial and Organizational Psychology-Perspectives on Science and Practice, 5*, 176-179.

Hough, L.M., Oswald, F.L., & Ployhart, R.E. (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, A.I., Goebl, A.P., & Culbertson, S.S. (2012). The engine is important, but the driver is essential: The case for executive functioning. *Industrial and Organizational Psychology-Perspectives on Science and Practice, 5*, 183-186.

Imus, A., Schmitt, N., Kim, B., Oswald, F.L., Merritt, S., & Wrestring, A.F. (2011). Differential item functioning in biodata: Opportunity eccess as an explanation of gender- and race-related DIF. *Applied Measurement in Education, 24*, 71-94.

Jensen, A.R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Leong, F.T.L., Leung, K., & Cheung, F.M. (2010). Integrating cross-cultural psychology research methods into ethnic minority psychology. *Cultural Diversity & Ethnic Minority Psychology, 16*, 590-597.

Lievens, F., Buyse, T., & Sackett, P.R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90*, 442-452.

Lievens, F., & De Soete, B. (2012). Simulations. In N. Schmitt (Ed.), *The Oxford handbook of assessment and selection* (pp.383-410). New York, NY: Oxford University Press.

Lievens, F., & Reeve, C.L. (2012). Where I-O psychology should really (re)start its investigation of intelligence constructs and their measurement. *Industrial and Organizational Psychology-Perspectives on Science and Practice, 5*, 153-158.

Lievens, F., Westerveld, L., & De Corte, W. (in press). Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management.*

Medley, D.M., & Quirk, T.J. (1974). Application of a factorial design to study cultural bias in general culture items on national teacher examination. *Journal of Educational Measurement, 11*, 235-245.

Mitchelson, J.K., Wicher, E.W., LeBreton, J.M., & Craig, S.B. (2009). Gender and ethnicity differences on the Abridged Big Five Circumplex (AB5C) of personality traits: A differential item functioning analysis. *Educational and Psychological Measurement, 69*, 613-635.

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

Murphy, K.R. (2010). How a broader definition of the criterion domain changes our thinking about adverse impast. In J.L. Outtz (Ed.), *Adverse impact. Implications for organizational staffing and high stakes selection* (pp. 137-160). New York, NY: Routledge.

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

Nápoles-Springer, A.M., Santoyo-Olsson, J., O'Brien, H., & Stewart, A.L. (2006). Using cognitive interviews to develop surveys in diverse populations. *Medical Care, 44*, 21-30.

Nguyen, H.H.D., & Ryan, A.M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93*, 1314-1334.

Oostrom, J.K., & Born, M.P. (2012). *Using cognitive pretesting to explore causes for ethnic differences on role-plays*. Manuscript submitted for publication.

Oostrom, J.K., Born, M.P., Serlie, A.W., & van der Molen, H.T. (2010). Webcam testing: Validation of an innovative open-ended multimedia test. *European Journal of Work and Organizational Psychology, 19*, 532-550.

Oostrom, J.K., Born, M.P., Serlie, A.W., & van der Molen, H.T. (2011). A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance. *Journal of Personnel Psychology, 10*, 78-88.

Outtz, J., Goldstein, H., & Ferreter, J. (2006, April). *Testing divergent and convergent thinking: Test response format and adverse impact.* Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Paullin, C., Putka, D.J., Tsacoumis, S., & Colberg, M. (2010, April). Using a logic-based measurement approach to measure cognitive ability. In C. Paullin (Chair), *Cognitive ability testing: Exploring new models, methods, and statistical techniques.* Symposium conducted at the 25th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Ployhart, R.E., & Holtz, B.C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153-172.

Pyburn, K.M., Ployhart, R.E., & Kravitz, D.A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology, 61*, 143-151.

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

Roth, P.L., Bevier, C.A., Bobko, P., Switzer, F.S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297-330.

Roth, P.L., Bobko, P., McFarland, L., & Buster, M. (2008). Work sample tests in personnel selection: A meta-analysis of black-white differences in overall and exercise scores. *Personnel Psychology, 61*, 637-661.

Roth, P.L., Buster, M.A., & Bobko, P. (2011). Updating the trainability tests literature on black-white subgroup differences and reconsidering criterion-related validity. *Journal of Applied Psychology, 96*, 34-45.

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

Roth, P.L., Switzer, F.S., Van Iddekinge, C.H., & Oh, I.S. (2011). Toward better meta-analytic matrices: How input values can affect research conclusions in human resource management simulations. *Personnel Psychology, 64*, 899-935.

Roth, P.L., Van Iddekinge, C.H., Huffcutt, A.I., Eidson, C.E., & Bobko, P. (2002). Corrections for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology, 87*, 369-376.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Ryan, A.M. (2001). Explaining the black-white test score gap: The role of test perceptions. *Human Performance, 14*, 45-75.

Ryan, A.M., & Greguras, G.J. (1998). Life is not multiple choice: Reactions to the alternatives. In M.D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 183-202). Mahwah, NJ: Lawrence Erlbaum Associates.

Ryan, A.M., & Huth, M. (2008). Not much more than platitudes? A critical look at the utility of applicant reactions research. *Human Resource Management Review, 18*, 119-132.

Sacco, J.M., Scheu, C.R., Ryan, A.M., Schmitt, N., Schmidt, D.B., & Rogg, I.U. (2000, April). *Reading level* as *a predictor of subgroup differences and validities of situational judgment tests.* Paper presented at the 14th Annual Conference for the Society for Industrial and Organizational Psychology, New Orleans, LA.

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

Sackett, P.R., Burris, L.R., & Ryan, A.M. (1989). Coaching and practice effects in personnel selection. In C.L. Cooper & I.T. Robertson (Eds.), *International review of industrial & organizational psychology* (pp. 145-183). West Sussex, England: John Wiley and Sons.

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

Sackett, P.R., Hardison, C.M., & Cullen, M.J. (2004). On interpreting stereotype threat as accounting for African American-white differences on cognitive tests. *American Psychologist, 59*, 7-13.

Sackett, P.R., & Roth, L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49*, 549-572.

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

Scherbaum, C.A., & Goldstein, H.W. (2008). Examining the relationship between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement, 68*, 537-553.

Schleicher, D.J., Van Iddekinge, C.H., Morgeson, F.P., & Campion, M.A. (2010). If at first you don't succeed, try, try again: Understanding race, age, and gender differences in retesting score improvement. *Journal of Applied Psychology, 95*, 603-617.

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

Schmitt, N., & Mills, A.E. (2001). Traditional tests and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology, 86,* 451-458.

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

Steele, C.M. (1998). Stereotyping and its threat are real. *American Psychologist, 53*, 680-681.

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

Steele, C.M., & Aronson, J.A. (2004). Stereotype threat does not live by Steele and Aronson (1995) alone. *American Psychologist, 59*, 47-48.

Tam, A.P., Murphy, K.R., & Lyall, J.T. (2004). Can changes in differential dropout rates reduce adverse impact? A computer simulation study of a multi-wave selection system. *Personnel Psychology, 57*, 905-934.

Van Iddekinge, C.H., Morgeson, F.P., Schleicher, D.J., & Campion, M.A. (2011). Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. *Journal of Applied Psychology, 96*, 941-955.

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

Whetzel, D.L., McDaniel, M.A., & Nguyen, N.T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance, 21*, 291-309.

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

Willis, G.B., & Miller, K. (2011). Cross-cultural cognitive interviewing: Seeking comparability and enhancing understanding. *Field Methods, 23*, 331-341.

Yusko, K.P., Goldstein, H.W., Scherbaum, C.A., & Hanges, P.J. (2012, April). *Siena Reasoning Test: Measuring intelligence with reduced adverse impact.* Paper presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Received: August 30, 2012