Strategies for dealing with the diversity-validity dilemma in personnel selection: Where are we and where should we go?

Britt De Soete*, Filip Lievens, and Celina Druart

Ghent University, Belgium

ARTICLE INFORMATION
Manuscript received: 12/10/2012
Revision received: 20/12/2012
Accepted: 20/12/2012

Keywords:
Strategies
Diversity
Personnel selection
Validity
Dilemma

ABSTRACT

The diversity-validity dilemma has been a dominant theme in personnel selection research and practice. As some of the most valid selection instruments display large ethnic performance differences, scientists attempt to develop strategies that reduce ethnic subgroup differences in selection performance, while simultaneously maintaining criterion-related validity. This paper provides an evidence-based overview of the effectiveness of six strategies for dealing with the diversity-validity dilemma: (1) using ‘alternative’ cognitive ability measures, (2) employing simulations, (3) using statistical approaches to combine predictor and criterion measures, (4) reducing criterion-irrelevant predictor variance, (5) fostering positive candidate reactions, and (6) providing coaching and opportunity for practice to candidates. Three of these strategies (i.e., employing simulation-based assessments, developing alternative cognitive ability measures, and using statistical procedures) are identified as holding the most promise to alleviate the dilemma. Potential areas in need for future research are discussed.

© 2013 Colegio Oficial de Psicólogos de Madrid. All rights reserved.

*Correspondence concerning this article should be sent to Britt De Soete, Department of Personnel Management, Work and Organizational Psychology, Henri Dunantlaan 2, 9000 Ghent, Belgium. E-mail: de.soete.britt@gmail.com

Keywords:
Strategies
Diversity
Personnel selection
Validity
Dilemma

Estrategias para manejar el dilema validez-diversidad en selección de personal: ¿dónde estamos y hacia dónde deberíamos ir?

RESUMEN

El dilema validez-diversidad ha sido un tema dominante en la investigación y la práctica de la selección de personal. Dado que algunos de los instrumentos de selección más válidos presentan grandes diferencias étnicas en sus puntuaciones, los científicos intentan desarrollar estrategias que reduzcan las diferencias de los subgrupos en la selección mientras mantienen simultáneamente la validez del criterio. Este artículo proporciona una revisión basada en la evidencia de las seis estrategias utilizadas para manejar el dilema validez-diversidad: (1) usar medidas de capacidad cognitiva “alternativas”, (2) emplear simulaciones, (3) utilizar procedimientos estadísticos para combinar las medidas predictores y del criterio, (4) reducir la varianza de los predictores irrelevantes para el criterio, (5) fomentar reacciones positivas en los candidatos y (6) facilitarles preparación y posibilidades de practicar. Tres estrategias (emplear evaluaciones basadas en simulaciones, desarrollar medidas alternativas de capacidad cognitiva y usar procedimientos estadísticos) son las más prometedoras para aminorar el problema. Se discuten áreas potenciales que requieren investigación en un futuro.

© 2013 Colegio Oficial de Psicólogos de Madrid. Todos los derechos reservados.

In personnel selection, some of the most valid selection instruments generally 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 are among the most valid predictors of job performance, but also 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 converge in personnel selection, which hinders organizations that aim to employ valid instruments while at the same time achieving acceptable levels of employee diversity. So, increasing diversity by limiting the occurrence of ethnic subgroup differences in selection performance has emerged as a key issue in the agendas

© 2013 Colegio Oficial de Psicólogos de Madrid. Todos los derechos reservados.

1576-5962/$ - see front matter © 2013 Colegio Oficial de Psicólogos de Madrid. Todos los derechos reservados
DOI: http://dx.doi.org/10.5093/tr2013a2
of both selection researchers and practitioners across the world. The main objective has been to discover strategies that reduce ethnic subgroup differences in selection performance, while simultaneously maintaining criterion-related validity. Note that subgroup differences are not restricted to White-Black differences in the USA but generally refer to selection performance differences between ethnic majority and ethnic minority members in a given country.

Over the years, various selection strategies have been proposed to diminish ethnic subgroup differences in selection test performance, while ensuring similar levels of criterion-related validity. In 2008, Ployhart and Holtz published an overview article that categorized the available strategies for dealing with the diversity-validity dilemma. They identified five main clusters of strategies: (1) the use of simulations and other predictors that display smaller ethnic subgroup differences as compared to cognitive ability, (2) statistically combining and manipulating scores, (3) reducing criterion irrelevant test variance, (4) fostering positive test-taker reactions, and (5) providing coaching and practice. Exactly five years after Ployhart and Holtz’s review, we pose the following question: Where are we now with regard to strategies for dealing with the diversity-validity dilemma and where should we go?

The current paper aims to answer this question by updating the recent research evidence on strategies for dealing with the diversity-validity dilemma. We used the original five-strategy framework as an organizing heuristic for our review (Ployhart & Holtz, 2008). In the end, six broad groups of strategies emerged because for clarity reasons we split up the Ployhart and Holtz’s first strategy (i.e., use of simulations and other predictors that display smaller ethnic subgroup differences as compared to cognitive ability) into two groups, namely (a) the use of alternative cognitive ability measures and (b) the use of simulations. For each strategy, we provide an overview of the current research thus far. We end each section with some domain-specific limitations and avenues for future research. Although there may be partial overlap with the Ployhart and Holtz’s overview, we mainly focused on novel research lines and findings within each category of strategies. A first criterion for inclusion was that the study had to deal with the diversity-validity dilemma or with a strategy to increase diversity or lower ethnic performance differences. To this end, we searched on computerized databases (e.g., Web of Science) using a combination of general keywords such as ‘selection’, ‘diversity’, ‘diversity-validity dilemma’, and ‘subgroup differences’, and keywords that were more closely related to each strategy such as ‘perceptions’, ‘practice’, ‘training’, ‘Pareto’, etc. A second criterion for inclusion deals with the publication year. Studies that were published after 2007 or that were not yet included in the Ployhart & Holtz’s summary were given particular attention. Additionally, we searched for unpublished manuscripts and abstracts of recent conference contributions (e.g., Annual Conference of Society for Industrial and Organizational Psychology). Our review presents scientists and practitioners with a research-based 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.

Use “Alternative” Cognitive Ability Measures

As jobs evolve to be more complex and challenging, assessing cognitive ability continues to gain in the selection process (Gatewood, Feild, & Barrick, 2011). However, as noted above, cognitive ability measures have repeatedly demonstrated to display large ethnic subgroup differences in test performance in US (Hough et al., 2001; Ployhart & Holtz, 2008; Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Sackett et al., 2001) as well as in European contexts (Evers, Te Nijenhuis, & Van der Flier, 2005), which substantially decreases hiring chances for members of ethnic minority groups (i.e., increased potential for adverse impact, which can be defined as less advantageous hiring rates for ethnic minority members as compared to ethnic majority members). Hence, researchers have advised exploring alternative measures of cognitive ability, which maintain validity and additionally exhibit substantially lower subgroup differences in test performance (e.g., Lievens & Reeve, 2012).

A first line of research within this domain has focused on logic-based measurement approaches as alternative measurement formats of cognitive ability (Paulin, 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). 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 some test 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. In fact, 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. Future studies that examine these assumptions are needed to gain insight on the effectiveness of logic-based measurement approaches for cognitive ability.

A 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 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 generic items (Hattrup, Schmitt, & Landis, 1992). This approach may reduce implicit cultural assumptions that are often embedded in generalized cognitive items (Brouwers & Van De Vijver, 2012).

Finally, researchers have suggested assessing 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 “executive functioning”, which has 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 sum, the search for alternative measures of cognitive ability as a substitute for or in combination with traditional cognitive ability tests covers an important first category of strategies for dealing with the diversity-validity dilemma. 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. In addition, 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 holds promise with regard to tackling the diversity-validity dilemma. However, more studies are needed to investigate whether these strategies hold their promise for dealing effectively with the diversity-validity dilemma.

Use Simulations as Additional Selection Procedures

In their review article, Ployhart and Holtz (2008) identified the use of simulation exercises as one of the best strategies to deal with the diversity-validity dilemma. 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 the tasks to be performed on the job (see Lievens & De Soete, 2012 for an overview). 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), whereas SJTs confront applicants with job-related dilemmas (e.g., choosing multiple options and making decisions). This concept bears resemblance to the concept of “executive functioning”, which has received substantial research attention lately (e.g., Chan & Schmitt, 1997; Ployhart & Holtz, 2008). In this context, the specific concept “executive functioning”, which has 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 sum, the search for alternative measures of cognitive ability as a substitute for or in combination with traditional cognitive ability tests covers an important first category of strategies for dealing with the diversity-validity dilemma. 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. In addition, 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 holds promise with regard to tackling the diversity-validity dilemma. However, more studies are needed to investigate whether these strategies hold their promise for dealing effectively with the diversity-validity dilemma.

Use Simulations as Additional Selection Procedures

In their review article, Ployhart and Holtz (2008) identified the use of simulation exercises as one of the best strategies to deal with the diversity-validity dilemma. 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 the tasks to be performed on the job (see Lievens & De Soete, 2012 for an overview). 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), whereas SJTs confront applicants with job-related dilemmas (e.g., choosing multiple options and making decisions). This concept bears resemblance to the concept of “executive functioning”, which has received substantial research attention lately (e.g., Chan & Schmitt, 1997; Ployhart & Holtz, 2008). In this context, the specific concept “executive functioning”, which has 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 sum, the search for alternative measures of cognitive ability as a substitute for or in combination with traditional cognitive ability tests covers an important first category of strategies for dealing with the diversity-validity dilemma. 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. In addition, 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 holds promise with regard to tackling the diversity-validity dilemma. However, more studies are needed to investigate whether these strategies hold their promise for dealing effectively with the diversity-validity dilemma.

Use Simulations as Additional Selection Procedures

In their review article, Ployhart and Holtz (2008) identified the use of simulation exercises as one of the best strategies to deal with the diversity-validity dilemma. 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 the tasks to be performed on the job (see Lievens & De Soete, 2012 for an overview). 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), whereas SJTs confront applicants with job-related dilemmas (e.g., choosing multiple options and making decisions). This concept bears resemblance to the concept of “executive functioning”, which has received substantial research attention lately (e.g., Chan & Schmitt, 1997; Ployhart & Holtz, 2008). In this context, the specific concept “executive functioning”, which has 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 sum, the search for alternative measures of cognitive ability as a substitute for or in combination with traditional cognitive ability tests covers an important first category of strategies for dealing with the diversity-validity dilemma. 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. In addition, 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 holds promise with regard to tackling the diversity-validity dilemma. However, more studies are needed to investigate whether these strategies hold their promise for dealing effectively with the diversity-validity dilemma.

Use Simulations as Additional Selection Procedures

In their review article, Ployhart and Holtz (2008) identified the use of simulation exercises as one of the best strategies to deal with the diversity-validity dilemma. 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 the tasks to be performed on the job (see Lievens & De Soete, 2012 for an overview). 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), whereas SJTs confront applicants with job-related dilemmas (e.g., choosing multiple options and making decisions). This concept bears resemblance to the concept of “executive functioning”, which has received substantial research attention lately (e.g., Chan & Schmitt, 1997; Ployhart & Holtz, 2008). In this context, the specific concept “executive functioning”, which has 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.
differences without impairing criterion-related validity. To close the gap between selection research and the fast evolving simulation practice, we believe that future research should focus on examining the efficacy of other high stimulus and response fidelity formats. Examples are innovative computer-based simulation exercises such as 3D animated SJTs, two-dimensional and three-dimensional graphic simulations, avatar-based SJTs, serious games, and multimedia instruments. Initial results regarding their ethnic subgroup differences and validity have been promising (Fetzer, 2012), but more research on their effectiveness is needed.

Use Statistical Approaches for Predictor and/or Criterion Scoring

A third category of techniques that deal with the diversity-validity dilemma refers to a number of statistical methods to combine and adjust 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).

A first strategy within this category makes use of non-cognitive predictors that exhibit smaller ethnic 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 in the 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 sufficiently high 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 dangers implied in 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 on 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 ethnic minority representation as shown by Hattrup, Rock, and Scala (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 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, although 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 in the USA (Cascio, Jacobs, & Silva, 2010). Future research should investigate other methods than the classical test theory for computing bands, such as item response theory (see Bobko, Roth, & Nicewander, 2005).

Reduce Criterion-irrelevant Variance in Candidates’ Selection Performance

A fourth strategy for reducing ethnic subgroup differences in selection performance consists of eliminating irrelevant variance caused by the predictor measure (Ployhart & Holtz, 2008). Irrelevant criterion variance denotes variance caused by predictor demands that are not related to the criterion (job performance). Irrelevant test
demands may be correlated with ethnicity and therefore generate subgroup differences that are unrelated to actual on-the-job performance differences. In order to increase diversity, it is recommended to eliminate irrelevant test demands from the selection procedure. Below we review recent research regarding two potential sources of criterion irrelevant test variance: verbal load and cultural load.

An instrument’s verbal load can be defined as the extent to which the predictor requires verbal (e.g., understanding, reading, writing texts) 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). For instance, 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 et al. (2001) found $d$-values for verbal ability tests ranging from .40 (for Hispanic-White comparisons) to .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 respondents are partly caused with 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 on the basis of 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 smaller ethnic subgroup differences, which was partly caused by the lower reading requirements of the video format as compared to the written format. Similarly, the recent use of technology-enhanced stimulus and response formats (e.g., two-dimensional and three-dimensional graphic animations, webcam testing, see Fetzer, 2012; Lievens et al., in press; Oostrom et al., 2010, 2011) also aims to lower reading and writing demands.

A 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. For instance, 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). Similarly, there might exist culture-based preferences for divergent thinking (“There are multiple answers for each problem”) vs. convergent thinking (“There is only one correct answer”, Outtz, Goldstein, & Ferreter, 2006). 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). Culturally equivalent instruments aim to avoid interpretation discrepancies or performance (dis)advantages that are related to ethnicity apart from 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 ethnic minority test-takers’ culture, e.g., 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. One exception is a study of McDaniel, Poitka, Legree, Yost, and Weekley (2011). They noticed that there exist White–Black mean differences in the preference for extreme responses on Likert scales. They reasoned that this common finding might affect ethnic subgroup differences on selection procedures that use Likert scales. Therefore, they developed a new scoring approach for SJTs that rely on Likert scales. In an attempt to make the SJT more culture-free they controlled for elevation and scatter in SJT scores. The results of this within-person standardization approach were encouraging: SJTs scored with this technique substantially reduced White–Black mean score differences and also yielded larger validities.

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). 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, & Steward, 2006; Willis & Miller, 2011). In personnel selection, the application of cognitive interviewing is in its infancy. Oostrom and Born (submitted) used the technique to discover differences in interpretation between ethnic majority and minority test-takers in role-play. Results demonstrated that several interpretation discrepancies could be identified, which were mostly explained by differences in language proficiency and attribution styles. Research on the effectiveness of the technique to reduce subgroup differences is needed.

A 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 ethnic minority group test-takers as compared to equally 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. However, 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 few 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.

Taken together, as verbal and cultural predictor requirements may enhance ethnic subgroup differences in test performance, it is recommended to limit a predictor’s verbal and cultural demands strictly to the extent that they are required in the context of 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. In future research, we need to examine the underlying causes of DIF, the specific 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.

**Foster Positive Test-Taker Reactions Among Candidates**

A 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, with ethnic minority group test-takers having less positive test perceptions. In turn, this 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 applicant 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.

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 Becton, Feild, Giles, & Jones-Farmer, 2008, for an exception). Along these lines, 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 smaller subgroup differences in perceived fairness and test-taking motivation (see also Chan & Schmitt, 1997).

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, it 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). Nonetheless, 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. 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 applicant withdrawal among ethnic minorities (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).

**Provide Coaching Programs and Opportunity for Practice to Candidates**

As a sixth strategy to alleviate the diversity-validity dilemma, the provision of coaching programs and the opportunity for practice and retesting to candidates has been suggested (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). Recently, this has been confirmed in a meta-analysis by Hausknecht, Halpert, Di Paolo, and 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 effects of practice, retesting, and coaching on diversity and criterion-related validity are 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 benefited
more from retesting than Blacks and this effect held stronger for written tests as compared to sample-based tests. 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 sum, 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 the effectiveness of the retesting strategy in the context of interpersonal skills (Roth, Buster, & Bobko, 2011).

**Discussion**

**Main Conclusions**

The current paper aimed to provide an updated overview of strategies for dealing with the diversity-validity dilemma. To this end, Table 1 summarizes the main results. As a general conclusion of our review, there does not seem to be an easy way to tackle the diversity-validity quandary. However, throughout our review, a number of strategies have emerged as particularly useful in the context of the diversity-validity dilemma. We present them below.

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 appears to lower ethnic subgroup differences. Third, investments in advanced assessment technologies seem to pay off. In fact, initial research findings suggest lower ethnic

<table>
<thead>
<tr>
<th>Strategies per Category</th>
<th>Examples and References</th>
<th>Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Use alternative cognitive ability measures</strong></td>
<td>Logic-based cognitive measurement instruments (e.g., Hattrup et al., 1992)</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>Improve correspondence of cognitive predictor with criterion (e.g., Ackerman &amp; Beier, 2012)</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>Measure specific cognitive abilities (e.g., Hattrup et al., 1992)</td>
<td>unknown</td>
</tr>
<tr>
<td><strong>2. Use simulation exercises as predictors</strong></td>
<td>Employ assessment centers, work samples, and SJTs (e.g., Dean et al., 2008; Lievens &amp; De Soete, 2012; Roth et al., 2003; Whetzel et al., 2008)</td>
<td>effective</td>
</tr>
<tr>
<td></td>
<td>Increase stimulus fidelity (e.g., Chan &amp; Schmitt, 1997)</td>
<td>effective</td>
</tr>
<tr>
<td></td>
<td>Increase response fidelity (e.g., De Soete et al., in press)</td>
<td>effective</td>
</tr>
<tr>
<td></td>
<td>Invest in technology-enhanced simulations (e.g., Fetzer, 2012)</td>
<td>unknown</td>
</tr>
<tr>
<td><strong>3. Use statistical approaches to combine and manipulate scores</strong></td>
<td>Combine different predictor scores into predictor composite (e.g., De Corte et al., 2006; Sackett &amp; Ellingson, 1997)</td>
<td>effective</td>
</tr>
<tr>
<td></td>
<td>Use Pareto-optimal predictor composites (e.g., De Corte et al., 2007, 2008; Druart &amp; De Corte, 2012)</td>
<td>effective</td>
</tr>
<tr>
<td></td>
<td>Use criterion weighting (e.g., Hattrup et al., 1997; De Corte, 1999)</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>Use banding (e.g., Campion et al., 2001; Cascio et al., 2010)</td>
<td>controversial</td>
</tr>
<tr>
<td><strong>4. Reduce construct irrelevant variance</strong></td>
<td>Reduce verbal load (e.g., Chan &amp; Schmitt, 1997)</td>
<td>effective</td>
</tr>
<tr>
<td></td>
<td>Reduce cultural load (e.g., Sensitivity panels (Beckase, 1996)</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>Scoring (McDaniel et al., 2011)</td>
<td>effective</td>
</tr>
<tr>
<td></td>
<td>Cognitive interviewing (Beatty &amp; Willis, 2007)</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>Detecting and removing DIF in items (e.g., Imus et al., 2011; Sackett et al., 2001; Scherbaum &amp; Goldstein, 2008)</td>
<td>mixed</td>
</tr>
<tr>
<td><strong>5. Foster positive test-taker reactions</strong></td>
<td>Alter test perceptions by increasing fidelity (e.g., Chan &amp; Schmitt, 1997; Edwards &amp; Arthur, 2007)</td>
<td>small</td>
</tr>
<tr>
<td></td>
<td>Combat stereotype threat (e.g., Steele, 1997, 1998; Steele &amp; Aronson, 1995)</td>
<td>mixed</td>
</tr>
<tr>
<td><strong>6. Provide coaching programs and opportunity for practice</strong></td>
<td>Provide practice, retesting, and coaching (e.g., Hausknecht et al., 2007; Schleicher et al., 2010)</td>
<td>mixed</td>
</tr>
</tbody>
</table>
subgroup differences and good criterion-related validity coefficients for several new multimedia simulations. Simultaneously, these instruments pose 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 satisfying 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 in 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 effective approaches for balancing the different outcomes of selection decisions.

Avenues For Future Research

Across the various avenues for future research that have already been pointed out in the current manuscript, we make the following key suggestions for future research. First, this research domain is in need of a more systematic operationalization of ethnicity or race. Currently, there exists little 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, as the current research on ethnic subgroup differences has mainly focused on Black-White performance discrepancies, it is imperative to expand this domain with European research findings and to compare ethnic subgroup differences in American and European settings (e.g., Hanges & Feinberg, 2005; Ruggs et al., 2013; for exceptions in European contexts: De Meijer et al., 2006; De Meijer, Born, Tertulow, & van der Molen, 2008; Ones & Anderson, 2002).

Third, we recommend researchers to apply a building block approach to study subgroup differences on simulation exercises as opposed to a holistic approach (Lievens et al., in press). Specifically, it is advised to conceptualize simulations as a combination of predictor constructs and predictor methods. In order to increase our knowledge of the theoretical drivers of diversity, specific factors of the simulation should be kept constant when manipulating either predictor constructs or predictor methods, as this permits us to increase our knowledge of the theoretical drivers of diversity.

Fourth, 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). Similarly, some scholars have stressed the potential of including social psychological theories (e.g., Helms, 2012) in the study of ethnic score differences.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Financial support

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

References


De Soete, B., Lievens, F., Oostrom, J., & Westerliefde, I. (in press). Alternative predictors for dealing with the diversity-validity dilemma in personnel selection: The...