Alternative Predictors for Dealing with the Diversity–Validity Dilemma in Personnel Selection: The constructed response multimedia test

Britt De Soete*, Filip Lievens*, Janneke Oostrom and Lena Westerveld*****

*Department of Personnel Management, Work and Organizational Psychology, Henri Dunantlaan 2, 9000 Ghent, Belgium. de.soete.britt@gmail.com
**VU University, Amsterdam, The Netherlands
***Police Academy, Apeldoorn, The Netherlands

In the context of the diversity–validity dilemma in personnel selection, the present field study compared ethnic subgroup differences on an innovative constructed response multimedia test to other commonly used selection instruments. Applicants ($N = 245$, 27% ethnic minorities) for entry-level police jobs completed a constructed response multimedia test, cognitive ability test, language proficiency test, personality inventory, structured interview, and role play. Results demonstrated minor ethnic subgroup differences on constructed response multimedia test scores as compared to other instruments. Constructed response multimedia test scores were related to the selection decision, and no evidence for predictive bias was found. Subgroup differences were also examined on the dimensional level, with cognitively loaded dimension scores displaying larger differences.

1. Introduction

One of the key challenges for personnel selection in the 21st century is ensuring and maintaining employee diversity and reducing adverse impact (e.g., differential hiring rates according to group membership; Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978) for juridical, political, economic, social, and ethical reasons. In search for valid selection instruments that permit a diverse inflow, researchers and practitioners are confronted with the diversity–validity dilemma, which implies that some of the most valid selection instruments also tend to display large ethnic subgroup differences in performance (Finch, Edwards, & Wallace, 2009; Hough, Oswald, & Ployhart, 2001; Ployhart & Holtz, 2008; Sackett, Schmitt, Ellingson, & Kabin, 2001). A main objective in personnel selection constitutes addressing the dilemma by developing alternative predictors, which aim to display minor to nonexistent ethnic subgroup differences (Ployhart & Holtz, 2008). The present study contributes to this stream of research by examining ethnic subgroup differences in performance on a constructed response multimedia test (Cucina et al., 2011; Lievens, De Corte & Westerveld in press; Oostrom, Born, Serlie, & Van Der Molen, 2010, 2011).

Constructed response multimedia tests present applicants with video-based job-related scenes (Olson-Buchanan & Drasgow, 2006), with a webcam capturing how they act out their response as if they actually take part in the presented situation. Recent studies on constructed response multimedia tests have focused on criterion-related validity. Constructed response multimedia test scores were significantly related to employment agents' job placement success and learning activities of students (Oostrom, Born, Serlie, & Van Der Molen, 2010; Oostrom, Born, Serlie, & Van Der Molen, 2011). In addition, Lievens et al. (in press) demonstrated scores on the constructed response multimedia test to predict training performance ratings for policemen. Across these studies, the observed validities

International Journal of SELECTION AND ASSESSMENT

of constructed response multimedia test scores for predicting academic, training, and job performance varied between .22 and .29.

However, a key question has remained unanswered so far: How does the constructed response multimedia test perform as an alternative predictor to deal with the diversity–validity dilemma? Hence, the present study provides a first attempt to address this query by comparing ethnic subgroup differences in test performance on the constructed response multimedia test to other frequently employed selection instruments (structured interview, role play, personality scales, and cognitive ability measures). In accordance with the content-method distinction (Arthur & Villado, 2008), we examined ethnic subgroup differences at both the instrument (e.g., method) level and dimension (e.g., content) level. Furthermore, the current study provides some preliminary data on the criterion-related validity and predictive bias of constructed response multimedia test scores.

The present study was situated in the context of applicant selection for entry-level police jobs in the Netherlands. This setting is particularly relevant for examining our objectives considering the special interest of police departments worldwide in hiring a diverse workforce (Metz & Kulik, 2008). After all, as the police corps deals with people of different ethnic backgrounds, it is appealing to employ an equally diverse staff. In addition, given the worldwide occurrence of adverse impact, it is insightful to extend the literature on ethnic subgroup differences and selection procedures from a North American to a European context (Hanges & Feinberg, 2009; Ones & Anderson, 2002).

1.1. Ethnic subgroup differences on selection instruments

The use of cognitive ability tests represents a striking example of the diversity–validity dilemma. Despite being one of the most valid predictors of job performance, several studies revealed cognitive ability tests to demonstrate the largest ethnic subgroup differences as compared to other selection instruments ($d = 1.00$–1.20; Bobko, Roth, & Potosky, 1999; Evers, Te Nijenhuis, & Van der Flier, 2005; Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Roth, Switzer, Van Iddekinge, & Oh, 2011). Other commonly used selection instruments, such as biodata measures ($d = 0.33$; Bobko et al., 1999), job knowledge tests ($d = 0.48$; Roth, Huffcutt, & Bobko, 2003), and employment interviews ($d = 0.25$–0.56; Huffcutt & Roth, 1998; Roth, Van Iddekinge, Huffcutt, Eidson, & Bobko, 2002), show smaller but still substantial ethnic subgroup differences.

As a result, calls have been made to construct 'alternative' selection instruments to deal with the diversity–validity dilemma (i.e., Sackett et al., 2001; Schmitt & Mills, 2001). The search for valid selection procedures that display smaller subgroup differences than traditional cognitive tests led to a renewed interest in simulation-based instruments (e.g., selection tests or exercises that physically or psychologically resemble those tasks to be performed on the job; Motowidlo, Dunnette, & Carter, 1990; Lievens & De Soete, 2012). Simulation-based selection instruments, such as assessment centers (ACs), work samples, and situational judgment tests (SJTs), have demonstrated criterion-related validity coefficients ranging from .19 to .45 (see Lievens & De Soete, 2012, for an overview), and due to their potential to capture a combination of (noncognitive) constructs by realistic measurement methods, several researchers have suggested their usefulness as alternative predictors (i.e., Callinan & Robertson, 2000; Motowidlo & Tippins, 1993; Ployhart & Holtz, 2008). Putting this belief to the test, several studies examined ethnic subgroup differences in ACs ($d = 0.52$; Dean, Bobko, & Roth, 2008), work samples ($d = 0.52$–0.73; Bobko, Roth, & Buster, 2005; Roth, Bobko, McFarland, & Buster, 2008), and SJTs ($d = 0.24$–0.38; Whetzel, McDaniel, & Nguyen, 2008). Meta-analytic research revealed that ethnic subgroup differences on simulations are generally significantly smaller than those on cognitive ability instruments, but they are still substantial and often larger than previously expected (i.e., Roth et al., 2008).

Taken together, the search for alternative predictors has been a dominant theme in personnel selection research and continued efforts should be undertaken to develop selection procedures that ensure a diverse workforce without impairing the selection quality (i.e., Dewberry, 2001; Ployhart & Holtz, 2008). Along these lines, the current study focuses on the constructed response multimedia test as a potentially useful alternative predictor in the domain of personnel selection.

1.2. Constructed response multimedia tests: Research and hypotheses

In search for alternative predictors to deal with the diversity–validity dilemma, we expect constructed response multimedia tests to provide an important contribution. Underlying this expectation is the concept of cognitive load. Spearman's hypothesis, formulated by Jensen (Jensen, 1998; Reeve & Bonaccio, 2009), states that ethnic subgroup differences are primarily a function of the instrument's cognitive load, which is defined as the correlation between the instrument test scores and cognitive ability measures (Whetzel et al., 2008). As cognitive load has repeatedly been identified as one of the main drivers of ethnic subgroup differences (Goldstein, Yusko, Braverman, Smith, & Chung, 1998; Goldstein, Yusko, & Nicolopoulos, 2001; Roth

et al., 2008; Whetzel et al., 2008), developing instruments that possess low cognitive load (e.g., low cognitive test demands) is suggested as an effective strategy to reduce ethnic subgroup differences (Ployhart & Holtz, 2008).

Altering the *stimulus presentation format* has been put forward as a first technique to lower the cognitive load of simulation-based instruments. More specifically, increasing the stimulus fidelity (e.g., the extent to which the stimuli presented by the instrument resemble the stimuli in the actual job situation) generally reduces irrelevant test requirements in terms of reading demands and therefore also cognitive demands. Along these lines, Chan and Schmitt (1997) and Lievens and Sackett (2006) demonstrated that a video-based SJT (relatively high stimulus fidelity) was associated with respectively lower reading demands and lower cognitive load as compared to a content-wise identical paper-and-pencil SJT (low stimulus fidelity). Accordingly, Chan and Schmitt found significantly smaller ethnic subgroup differences in performance on the video-based SJT than on the paper-and-pencil SJT.

Similarly, cognitive load may be reduced by increasing the *fidelity of the response format* (e.g., the extent to which the instrument's response format resembles the response requirements during on-the-job behavior, Bobko & Roth, 2013). Lievens et al. (in press) demonstrated that a (high fidelity) behavioral constructed response format displayed a lower cognitive load as compared to a (low fidelity) written constructed response format. Regarding diversity, higher fidelity response formats have been found to display smaller ethnic subgroup differences than low fidelity response formats for content-wise identical knowledge tests (i.e., constructed response vs. multiple-choice response format; Arthur, Edwards, & Barrett, 2002; Edwards & Arthur, 2007).

In a similar vein, we expect a rather low cognitive load for constructed response multimedia tests because they present applicants with video stimuli of real-life work situations and additionally ask applicants to act out their responses to the given situations as if the job situation actually takes place. More specifically, applicants are required to react immediately and behaviorally (instead of in writing) to the presented stimulus material, which does not permit time to reflect, reread, or correct their response. Along these lines, research in linguistics has shown that behavioral responses require less cognitive resources than written responses (Bourdin & Fayol, 2002).

Taken together, we anticipate a low cognitive load for constructed response multimedia test scores, which we hypothesize to translate into small ethnic subgroup differences in test performance on the constructed response multimedia test as compared to on other selection instruments (Cohen, 1988).

Hypothesis 1: Constructed response multimedia test scores will display small ethnic subgroup differences in test performance.

Recently, the distinction between predictor constructs and predictor methods has emerged as a key development to advance our conceptual understanding of personnel selection procedures (Arthur & Villado, 2008). Predictor constructs denote the behavioral domain captured in the selection procedure (i.e., Knowledge, skills, abilities and other characteristics, KSAOs) whereas predictor methods refer to the techniques that are used to measure these constructs (i.e., role play, SJT). Therefore, examining ethnic subgroup differences at the method (instrument) level represents only one side of the equation. In terms of advancing our understanding of the underlying factors of diversity, it is equally important to study ethnic performance differences at the construct level.

We expect the nature of the dimensions measured to influence the magnitude of the observed subgroup differences. More specifically, as noted above, cognitive load is expected to be a main driver of ethnic performance differences on the dimensions. Various studies have demonstrated that higher cognitive load of the dimensions of simulation-based instruments is associated with larger ethnic subgroup differences (for ACs: Goldstein et al., 1998, 2001; for work samples: Roth et al., 2008; for SJTs: Whetzel et al., 2008). In the present study problem solving, interpersonal sensitivity, forcefulness, and integrity are measured across three selection instruments (e.g., constructed response multimedia test, role play, and structured interview). As problem solving is assumed to have the highest cognitive loading (Goldstein et al., 2001), we hypothesize this dimension to display larger ethnic performance differences than the other dimensions.

Hypothesis 2: Dimensions with higher cognitive loading (e.g., problem solving), defined as the extent to which the dimension is correlated with cognitive ability, will be associated with larger subgroup differences in performance.

A major dilemma in personnel selection is that reductions in ethnic subgroup differences should not come at the cost of criterion-related validity (Ployhart & Holtz, 2008). Accordingly, apart from examining ethnic subgroup differences in scores on the constructed response multimedia test, the present study aims to shed light on the instrument's criterion-related validity and its potential for predictive bias. As the constructed response multimedia test requires test takers to show actual verbal and nonverbal behavior in acting out their response, we expect it to have a high point-to-point correspondence with actual on-the-job behavior. In previous

research, constructed response multimedia test scores were found to be valid for predicting several external outcome measures (see Lievens et al., in press; Oostrom et al., 2010, 2011). Given the absence of an external criterion in this study, we used the selection decision as an internal validation criterion. Note that the constructed response multimedia test had no bearing on this decision. Taken together, this leads to the following hypothesis:

Hypothesis 3: Scores on the constructed response multimedia test will be significantly related to the selection decision outcome.

2. Method

2.1. Sample and procedure

The sample consisted of 245 applicants who applied for an entry-level police job at the Dutch police academy. The applicants' mean age was 25.21 years ($SD = 5.80$). The current study sample consisted of 188 men and 57 women, which is a common gender distribution for police jobs (De Vries & Pettigrew, 1998; Metz & Kulik, 2008). There were 67 ethnic minority[1] members in the applicant pool (e.g., 27%), which is proportionally somewhat more than the percentage of ethnic minorities in the entire Dutch population (e.g., 21%; Statistics Netherlands, 2011). Reflecting the composition of the ethnic minority population in the Netherlands (Statistics Netherlands, 2011), most ethnic minority participants in the sample had a Turkish or Moroccan background.

The selection procedure took off with an administrative screening, during which noneligible candidates (e.g., candidates who did not obtain a high school degree) were not withheld. Next, 245 applicants attended a 2-day selection process, which consisted of a cognitive ability test, a language proficiency test, a personality inventory, a role play, and a structured interview. Finally, applicants completed the constructed response multimedia test, which had no impact on the final hiring decision. During administration of the multimedia test, applicants took place in front of a laptop. The test started with instructions and a practice item. Subsequently, applicants were led through the assessment with a predetermined pace which prevented backtracking. Each item consisted of a video scene that was played once on the laptop screen. At the end of the scene, the character in the video looked into the camera like (s)he was directly addressing the applicant and the scene froze. Next, the applicant was expected to react as if the situation actually took place. Responses were recorded automatically by a webcam that was mounted on the laptop.

2.2. Predictor measures

2.2.1. Constructed response multimedia test: Development

The constructed response multimedia test consists of videotaped item stems that confront applicants with key situations related to the job. To develop the instrument, we followed existing procedures for constructing multimedia SJTs (see Table 1; Chan & Schmitt, 1997; Weekley & Jones, 1997). Four KSAOs were identified to be the focus of the multimedia test, namely problem solving, interpersonal sensitivity, forcefulness, and integrity. The final instrument consists of 24 items, which contain interactions between police officers and civilians/colleagues that entry-level police officers are likely to encounter. There were each time eight scenes specifically designed to trigger behavior related to interpersonal sensitivity, forcefulness, and integrity (totaling 24 items). Problem solving was rated in all 24 scenes. No prior police knowledge was required to answer the multimedia test items.

2.2.2. Constructed response multimedia test: Rating process

A pool of 18 experienced assessors received a half-day frame-of-reference training, practice, and feedback. After test administration, each candidate's responses (24 webcam vignettes) were randomly assigned to two selection officers. They used behaviorally anchored rating scales (BARSs) for evaluating the responses. After viewing all eight vignettes per dimension, an overall dimension rating was given on a 5-point rating scale (1 = *poor* to 5 = *excellent*) for interpersonal sensitivity (i.e., reassures and provides help, shows concern for others), forcefulness (i.e., does not hesitate to confront others, discusses consequences of behavior), and integrity (i.e., confronts in the case of power abuse or inappropriate behavior). Problem solving (i.e., provides arguments for decisions) was rated after viewing all 24 vignettes. The one-way random intraclass correlations (ICC[1,2]) were .77 for problem solving, .80 for interpersonal sensitivity, .68 for forcefulness, and .80 for integrity. Intercorrelations between the dimension scores ranged between .38 and .69.

2.2.3. Cognitive ability test

A computer adaptive test of inductive reasoning skills, which required applicants to find the underlying principle in a configuration of letters or numbers, was used for measuring cognitive ability (CEBIR, 2013). As this test was the publisher's property, we obtained only applicants' final scores and were not able to compute internal consistencies on the item level.

2.2.4. Language proficiency test

Language proficiency was measured by four tests of general language proficiency and three tests of specific lan-

Table 1. Development of the constructed response multimedia test

Step	Description of actions
1	• A thorough job analysis was undertaken to determine those KSAOs that are relevant for entry-level police jobs. • Four KSAOs were identified to be the focus of the multimedia test, namely *problem solving*, *interpersonal sensitivity*, *forcefulness*, and *integrity*.
2	• We conducted interviews with 15 police officers and sergeants (3 women, 12 men; 4 ethnic minorities, 11 ethnic majorities) to gather critical incidents relevant for police jobs. • Redundant incidents were identified and removed from the pool. • Nonredundant incidents were grouped into categories. • Incidents were evaluated by a sensitivity review panel on their language and cultural sensitivity.
3	• Item stems were written based on the 70 incidents that survived the former step. • Police officers and sergeants reevaluated the scenarios to remove items that were not realistic in an entry-level police environment or did not sufficiently capture the four KSAOs. As such, 20 items were eliminated.
4	• A pilot test was conducted among 228 candidates (165 men, 62 women; 19 minorities, 208 majorities) to assess the difficulty level of the remaining 50 items. • Items that were not able to discriminate between applicants or were too costly to film (i.e., a car crash) were eliminated. • From this set of 31 items, 24 were randomly chosen (e.g., from the available items per dimension category) to compose the constructed response multimedia test.
5	• There were eight scenes specifically designed to trigger behavior related to *interpersonal sensitivity*. For example, in one scene a pregnant woman parks her car on a spot reserved for people with disabilities while she does not possess the corresponding parking permit, thereby violating the law. Subsequently, the applicant is asked to take the role of police officer and react on this situation as if it really takes place. • Eight different scenes aimed to trigger behavior related to *forceful* behavior (i.e., a man reacts aggressively when the police accuses him of nuisance after several neighbor protests). • The final eight scenes were designed to trigger behavior related to *integrity* (i.e., a thankful civilian offers a police officer a present). • *Problem solving* was rated in all 24 scenes.

guage abilities (ICE, 2005). The internal consistency of all language test scores was satisfactory ($\alpha = .75$). Hence, we computed a composite score for language proficiency.

2.2.5. Personality inventory

Personality was measured by the M5Q, which is a measure of the Big Five personality factors (Klinkenberg & Van Leeuwen, 2003; Van Leeuwen, 2000). Each factor is measured by a 10-item scale. An example item is 'I enjoy talking to people' (extraversion). The test manual reported good internal consistencies (.80–.86) and test–retest reliabilities (.80–.94) for each of the five scales.

2.2.6. Structured interview

Each candidate was invited for a 1-hr structured (level 3; see Huffcutt & Arthur, 1994) behavior description interview conducted by a psychologist. All interviewers received an internal training period of up to 3 months during which they assisted senior interviewers while conducting the interview. The interview aimed to measure the four earlier mentioned KSAOs, which resulted in scores from *very poor* (1) to *outstanding* (7) on problem solving, interpersonal sensitivity, forcefulness, and integrity.

2.2.7. Role play exercise

Finally, a role play was included in the selection procedure. It aimed to elicit behavior related to sensitivity, problem solving, and forcefulness, which were the same job-related KSAOs as were the focus of the multimedia test and the structured interview (except for integrity). Each candidate took part in a role play that simulated a conflict situation in which the applicant was assigned a moderating role between the conflicting parties in order to constructively settle the argument. All assessors were experienced selection officers and had attended a comprehensive training seminar in accordance with the Guidelines and Ethical Considerations for Assessment Center Operations (International Task Force on Assessment Center Guidelines, 2000). Each applicant was rated by two assessors who used 7-point BARS (1 = *poor* to 7 = *outstanding*). The interrater agreement (ICC[1,2]) equaled .93.

2.2.8. Overall dimension scores

Overall dimension scores across instruments were calculated for problem solving, interpersonal sensitivity, forcefulness, and integrity by computing the mean of the standardized instrument dimension scores across the constructed response multimedia test, structured interview, and role play.

2.3. Criterion measure

As participants' test scores on the constructed response multimedia test had no impact on the final selection

decision, we used the selection decision (e.g., being hired vs. not being hired) as an internal validation criterion. The selection decision was based on applicants' scores on all selection instruments (except for the multimedia test) combined with subjective judgments of the selection board. Of the 245 applicants who participated in our study, 56 were selected (e.g., a selection ratio of 23%), of which 41 were ethnic majority members and 15 ethnic minority members.

3. Results

3.1. Descriptive statistics

Descriptive statistics, intercorrelations, and Cronbach's alphas for all study variables are presented in Table 2. This table shows the constructed response multimedia test to be positively correlated with language proficiency, interview ratings, role play performance, extraversion, agreeableness, and conscientiousness, and negatively correlated with neuroticism. Note that performance on the constructed response multimedia test was not correlated with cognitive ability, which is in line with our expectations. Finally, test scores of the constructed response multimedia test were correlated with age, $r = .23$, $p < .001$, so that older (e.g., more experienced) applicants obtained higher ratings. There was no significant relationship with gender, $r = -.01$, $p = .884$. Further, Table 2 demonstrates that scores on the multimedia test, structured interview, and role play are significantly correlated with the final selection decision.

3.2. Test of Hypothesis 1

According to Hypothesis 1, the constructed response multimedia test was expected to display small ethnic subgroup differences. To examine ethnic performance differences in selection test performance, effect sizes of mean differences were 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 ethnic majority group score by the mean ethnic minority group score and dividing this measure by the pooled group standard deviation. Positive *d*-values indicate average test scores advantaging ethnic majority members, whereas negative *d*-values point to the opposite.

Table 3 shows an overview of the effect sizes associated with each selection instrument. As expected, the largest differences in test performance were found to be associated with cognitive measures: the uncorrected *d*-values for cognitive ability and for language proficiency equaled 0.42 and 0.56, respectively. The role play

Table 2. Means, standard deviations, reliabilities, and intercorrelations of study variables

	Variable	*N*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13
1	Ethnicity	245	—	—													
2	Gender	245	—	—	−.06												
3	Age	245	25.16	5.85	.03	−.12											
4	Cognitive ability	243	−9.30	159.82	−.19**	−.01	.06	(.55)									
5	Language proficiency	117	109.08	7.71	−.25**	−.14	.12	.42**	(.75)								
6	Neuroticism	231	105.14	13.15	.07	.14*	−.16*	−.03	−.10	(.83)							
7	Extraversion	228	178.71	12.85	−.05	.05	−.13*	.11	.01	−.43**	(.76)						
8	Openness	231	156.30	14.23	−.10	.05	.15*	.20**	.06	−.23**	.42**	(.69)					
9	Agreeableness	231	173.26	12.17	−.20**	.11	.10	.03	.06	−.47**	.31**	.24**	(.72)				
10	Conscientiousness	231	181.84	12.49	.08	−.17*	.16*	−.08	.03	−.61**	.37**	.16*	.36**	(.82)			
11	Structured interview	230	17.03	2.06	−.07	.03	.22**	.15*	.06	−.28**	.25**	.19**	.18**	.25**	(.64)		
12	Role play	235	11.11	2.92	−.13	.05	.10	.11	.05	−.17*	.23**	.28**	.07	.04	.61**		
13	Constructed response multimedia test	239	34.74	7.21	−.06	−.01	.23**	.11	.22*	−.21**	.15*	.11	.13*	.20**	.41**	.39**	
14	Selection decision	245	—	—	−.01	.00	−.04	−.03	−.04	−.12	.11	.01	.15*	.11	.36**	.25**	.24**

Notes: Ethnicity, gender, and selection decision are dummy coded (men = 0, women = 1; ethnic majority member = 0, ethnic minority member = 1; not hired = 0, hired = 1). Cronbach's alphas are in parentheses. *Correlation is significant at the .05 level (two tailed). **Correlation is significant at the .01 level (two tailed). Overall instrument scores for the interview, role play, and constructed response multimedia tests constitute the sum scores of the instrument dimension scores that are measured per instrument.

Table 3. Subgroup differences on different instruments

	Instrument	Ethnic majority			Ethnic minority			t	p	d (uncorrected)	d (corrected)
		N	M	SD	N	M	SD				
1	Cognitive ability	176	9.34	150.24	67	–58.27	174.48	2.80	.01	0.42	0.57
2	Language proficiency	84	110.30	7.15	33	105.97	8.32	2.81	.01	0.56	0.65
3	Neuroticism	166	104.61	12.70	65	106.51	14.27	–0.99	.32	–0.14	–0.15
4	Extraversion	163	179.10	12.91	65	177.72	12.75	0.73	.47	0.11	0.13
5	Openness	166	157.18	15.11	65	154.06	11.51	1.69	.09	0.23	0.28
6	Agreeableness	166	174.78	12.32	65	169.40	10.97	3.07	.00	0.46	0.54
7	Conscientiousness	166	181.19	12.43	65	183.49	12.61	–1.26	.21	–0.18	–0.20
8	Structured Interview	170	17.12	2.16	60	16.80	1.74	1.03	.31	0.16	0.20
9	Role play	169	11.34	2.86	66	10.52	3.01	1.96	.05	0.28	0.29
10	Constructed response multimedia test	176	35.01	7.20	63	33.98	7.26	0.96	.34	0.14	0.15

Notes: Uncorrected effect sizes are calculated by dividing mean score differences by the pooled standard deviation. Corrected effect sizes represent the effect sizes corrected for attenuation (Bobko, Roth, & Bobko, 2001) with ICC(1,2)s as reliability measures for the constructed response multimedia test and role play and internal consistencies as reliability measure for all other instruments. Positive d-values indicate performance differences in favor of ethnic majority members, negative d-values point to the opposite. Overall instrument scores for the interview, role play, and constructed response multimedia tests constitute the sum scores of the instrument dimension scores that are measured per instrument.

displayed ethnic subgroup differences of moderate size ($d = 0.28$). For the constructed response multimedia test, an uncorrected d-value of 0.14 was found, thereby providing support for Hypothesis 1. Note that the structured interview was associated with subgroup differences of similar magnitude as those on the constructed response multimedia test ($d = 0.16$).

3.3. Test of Hypothesis 2

Hypothesis 2 stated that dimensions with higher cognitive loading, defined as the extent to which dimension scores are correlated with cognitive ability test scores (Whetzel et al., 2008), will be associated with larger ethnic subgroup differences. To put this hypothesis to the test, the cognitive loading, as well as ethnic subgroup differences, were compared among overall dimension scores.[2] Table 4 shows that the dimension with the largest cognitive loading (e.g., problem solving) demonstrates the largest ethnic subgroup differences. The d-value for problem solving equaled 0.45, indicating a moderately large performance difference in favor of White test takers. To examine whether the ethnic performance differences on problem solving are due to the dimension's cognitive load, we tested for an indirect effect of ethnicity on problem solving through cognitive ability. For small to moderate samples, it is advised to examine indirect effects by bootstrapping procedures (Hayes, 2009; Preacher & Hayes, 2004; Shrout & Bolger, 2002). By extracting 5,000 bootstrapped samples from the dataset based on random sampling with replacement and computing the indirect effect of ethnicity on the dimension scores through cognitive ability for each sample, 95% confidence intervals (CIs) were calculated. A direct effect of ethnicity on overall problem solving score was observed, $t(218) = -3.62$, $p = .000$, also when cognitive ability was controlled for, $t(214) = -3.13$, $p = .002$. In addition, an indirect effect of ethnicity on problem solving through cognitive ability was found (*estimate* = –.05, *SE* = .03, *lower CI* = –.11, *higher CI* = –.01, $p < .05$). Taken together, these findings demonstrate a partial mediation and thereby emphasize the role of cognitive load in dimensional subgroup differences. In addition, we tested Hypothesis 2 by using the method of correlated vectors (Jensen, 1998; Reeve & Bonaccio, 2009). The correlation between the dimensional g-loading and standardized mean ethnic performance differences vectors was computed. An uncorrected correlation of $r = .44$ was found, indicating that the magnitude of the subgroup differences (in favor of majority members) on the dimensions increases as the cognitive load of the dimension enhances. In sum, these results largely support Hypothesis 2.

Table 4 further shows that the noncognitively loaded dimensions show the smallest ethnic subgroup differences in overall dimension scores as forcefulness and

Table 4. Subgroup differences at dimension by instrument and at dimension level

	Constructed response multimedia test	Structured interview	Role play	Composite of three	Composite of three
	d	*d*	*d*	*d*	*g*-loading
Problem solving	0.27	0.27	0.51	0.45	.19**
Interpersonal sensitivity	−0.04	0.00	0.00	−0.02	.09
Forcefulness	−0.13	0.07	0.15	0.05	.11
Integrity	0.35	0.16	—	0.34	.06
Overall	0.14	0.16	0.28		

Note: Effect sizes are calculated by dividing mean score differences by the pooled standard deviation. Positive *d*-values indicate performance differences in favor of ethnic majority members, negative *d*-values point to the opposite. Composite effect sizes were calculated based on the formula of Sackett and Ellingson (1997) for equally weighted multipredictor composites. The cognitive loading (*g*-loading) refers to the correlation between overall dimension scores and performance on the cognitive ability test. *Correlation is significant at the .05 level (two tailed). **Correlation is significant at the .01 level (two tailed).

Table 5. Logistic regression for selection decision on ethnicity and performance on the constructed response multimedia test

Predictor	*B*	*SE*	*Wald*	*df*	*p*	*Exp(B)*
Constant	−3.87	.99	15.43	1	.00	.02
Ethnicity	−1.23	2.06	.36	1	.55	.29
Constructed response multimedia test	.07	.03	7.86	1	.01	1.08
Ethnicity × Constructed Response Multimedia Test	.04	.06	.41	1	.52	1.04

Notes: Ethnicity is dummy coded (ethnic majority member = 0, ethnic minority member = 1). Cox & Snell $R^2 = .06$, Nagelkerke $R^2 = .09$.

interpersonal sensitivity were associated with *d*-values of 0.05 and −0.02, respectively. One exception is integrity, which was associated with moderate ethnic subgroup differences in overall performance ($d = 0.34$) despite its low cognitive loading ($r = .06$, $p = .354$).

3.4. Test of Hypothesis 3

To test Hypothesis 3, we examined the criterion-related validity of the constructed response multimedia test. Table 2 reveals a validity coefficient of $r = .24$ ($p = .000$) for predicting the selection decision. In order to make a judgment on differential validity, we compared validity coefficients of the constructed response multimedia test for both ethnic groups. For ethnic minority members and ethnic majority members, we observed validity coefficients of $r = .30$ ($p = .018$) and $r = .22$ ($p = .004$), respectively, which were not significantly different ($z = 0.26$, $p = .397$).

We also tested for differential prediction using the Cleary model (Cleary, 1968). A logistic regression with ethnicity, constructed response multimedia test scores, and their interaction as predictors and the selection decision as dependent variable revealed solely an effect of constructed response multimedia test scores on the selection decision, whereas ethnicity and the interaction of ethnicity and multimedia test performance displayed no significant effect on the selection decision (see Table 5). These results should be interpreted with caution given the small sample size.

4. Discussion

Organizations and researchers are nowadays challenged to develop selection instruments that ensure work staff diversity without impairing selection quality. In the context of the diversity–validity dilemma in personnel selection, previous studies have shown mixed success in their attempts to develop valid instruments that reduce ethnic subgroup differences (i.e., Roth et al., 2008). The aim of the current field study was to provide a first attempt in examining whether an innovative simulation, namely a constructed response multimedia test, displays minor ethnic score differences without impairing criterion-related validity. Results demonstrated that the constructed response multimedia test in the present study displays small ethnic subgroup differences as compared to other commonly used selection instruments. Furthermore, these performance differences were found to be partly attributable to the cognitive load of the test dimensions measured. Additionally, performance on the constructed response multimedia test significantly predicted the selection decision outcome and we found no evidence of differential prediction or differential validity.

At a practical level, the present findings combined with the predictive validity evidence found in previous studies (i.e., Oostrom et al., 2010, 2011) suggest that the constructed response multimedia test may be a valuable alternative predictor in diverse applicant settings. This might be particularly relevant for police force selection as previous studies have revealed rather low

predictive validity coefficients for cognitive ability tests in police contexts (Dayan, Kasten, & Fox, 2002; Hirsh, Northrop, & Schmidt, 1986; Pynes & Bernardin, 1989; Salgado et al., 2003) and called for alternative instruments to assess interpersonal skills (i.e., Hirsh et al., 1986)

Contrary to our hypothesis, the dimension of integrity showed a substantial ethnic score difference ($d = 0.34$), which could not be explained by the dimension's cognitive saturation. Two possible explanations may account for this finding. First, the results may be attributable to the particular demographic composition of the present study's ethnic minority group, which deviates from other (US) research samples. Different ethnic groups may diverge on their definitions of the concept of integrity. Second, this finding may result from the operationalization of integrity. According to Van Iddekinge, Taylor, and Eidson (2005), integrity is a multidimensional construct and the dimensions vary in the ethnic differences that they display ($d = -0.08$ to 0.77). The particular operationalization of integrity may explain why some researchers found negligible subgroup differences (i.e., Ones & Viswesvaran, 1998) whereas the present and other studies have found significant ethnic performance discrepancies for integrity measures (i.e., Van Iddekinge et al., 2005). Particularly in the context of law enforcement occupations, follow-up research is necessary to identify the underlying reasons for subgroup differences on integrity dimensions.

As opposed to the majority of US diversity selection studies, the present study was conducted in a European selection setting. As the US ethnic minority group composition differs from the European, it is worthwhile to compare findings on ethnic subgroup differences in selection test performance. Table 6 contrasts European findings on subgroup differences combined with the present study's ethnic performance differences with their commonly found US equivalents for Black, Hispanic, and Asian minorities. European and US findings seem to be in line for the structured interview, role play, language proficiency, and most personality scales. Cognitive ability seems to differ somewhat, with adverse impact potential in Europe to appear slightly lower than in North American settings. Furthermore, the present study shows quite large score differences on agreeableness with ethnic minorities scoring consistently lower, which is in contrast to most European and US research findings (but for an exception, see Weekley, Ployhart, & Harold, 2004). However, as European meta-analyses on ethnic subgroup differences are mostly lacking thus far and considering the sample size of the present study, caution is in order when drawing conclusions from this comparison. Therefore, we encourage further European research on subgroup differences in selection contexts to confirm and expand the present findings.

Table 6. Subgroup differences between Whites and ethnic minorities

	Instrument	Europe		United States			
		Present study ($N_{minority} = 67$)	Overall	Blacks[a]	Hispanics[a]	Asians[a]	
		d	*d*	*d*	*d*	*d*	
1	Cognitive ability	0.42	0.26[b]–1.39[c]	0.99–1.20[e]	0.58–0.83	–0.20	
2	Language proficiency	0.56	No data available	0.55[f]–0.76[f]	0.40[f]	No data available	
3	Neuroticism	–0.14	0.06[b]	0.04	0.01	–0.08	
4	Extraversion	0.11	0.00[b]	0.10	–0.01	0.15	
5	Openness	0.23	0.04[b]	0.21	0.10	0.18	
6	Agreeableness	0.46	–0.05[b]	0.02	0.06	0.01	
7	Conscientiousness	–0.18	–0.12[b]	0.06	0.04	0.08	
8	Structured interview	0.16	0.16[b]	0.23	No data available	No data available	
9	Role play	0.28	0.15–0.21[b]	0.03–0.40[f]	No data available	No data available	
10	Situational judgment test (video)	No data available	0.38[d]	0.31	0.41	0.49	
11	Constructed response multimedia test	0.14	No data available	No data available	No data available	No data available	

Notes: Effect sizes are calculated by dividing mean score differences by the total group standard deviation. Positive *d*-values indicate performance differences in favor of ethnic majority members, negative *d*-values point to the opposite. [a]Based on Ployhart and Holtz (2008) unless stated differently. [b]Based on De Meijer, Born, Terlouw, and Van der Molen (2008). [c]Based on Te Nijenhuis, De Jong, Evers, and Van Der Flier (2004). [d]Based on De Meijer (2008). [e]Based on Jensen (1998). [f]Based on Hough et al. (2001).

Some limitations of the present study should be mentioned. A first limitation concerns the sample. The present study's sample contains 245 applicants, and should therefore be perceived as a first promising attempt to examine the effectiveness of the constructed response multimedia test as an alternative predictor. Future research is needed to expand the present findings. Additionally, the ethnic minority sample is characterized by its heterogeneous nature regarding ethnic background, which is common in European settings (i.e., De Meijer, Born, Terlouw, & Van der Molen, 2008). Although the current study already provides a first important overview among subgroup differences between native Dutch and immigrant applicants, the ethnic minority sample was rather small in order to differentiate among ethnic groups. However, we made a first attempt by replicating the analyses solely for the Turkish and Moroccan subgroup, which is the largest immigrant group in Western Europe. Results were in the same line for the constructed response multimedia test ($d = 0.15$ vs. $d = 0.14$ for the full group), although subgroup differences on the cognitive ability test were more pronounced ($d = 0.79$ vs. $d = 0.42$ for the full group). Future research should strive to examine performance differences on constructed response multimedia tests for various ethnic groups by differentiating according to ethnicity or cultural similarity (e.g., Schwartz, 2004). A second limitation relates to the construct-related validity of our measures. We made an effort to differentiate between methods and constructs in explaining subgroup differences on the selection instruments. Yet, it should be noted that the correlations between dimensions were moderate, which is consistent with prior research (e.g., Lievens & Conway, 2001). To this end, we computed overall dimension scores across instruments. Third, the present study made use of an internal criterion measure because external criteria data were not available. Additionally, the observed criterion-related validity coefficients in the present study are likely to be underestimations of the actual values, as there was no possibility to correct for range restriction. Future studies should expand this line of research by simultaneously examining diversity and validity criteria.

Although the present research results are promising for the use of constructed response multimedia tests as alternative predictors, follow-up research is necessary to replicate these findings in larger samples with different ethnic compositions, and for constructed response multimedia tests that capture different dimensions. Additionally, future studies should examine the underlying mechanisms of ethnic subgroup differences on constructed response multimedia test scores. The present study already provided a first attempt in clarifying the role of cognitive load. Given the constructed response format, other potential drivers of ethnic subgroup differences on constructed response multimedia test scores are culture-related preferences for specific communication styles and test motivation (i.e., Chan & Schmitt, 2004; Gudykunst et al., 1996; Helms, 1992).

Acknowledgement

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

Notes

1. As there is no straightforward definition of 'ethnic minority member', proxy variables are often used. The present study defines ethnic minority/majority status based on applicants' self-reported ethnicity.
2. A caveat should be added. The dimension correlations between instruments are moderate. That is, problem solving scores on the constructed response multimedia test correlated .36 ($p < .001$) with scores on the role play and the interview. Interpersonal sensitivity scores on the constructed response multimedia test correlated .22 ($p = .001$) and .25 ($p < .001$) with sensitivity scores on the role play and the interview, respectively. Forcefulness on the constructed response multimedia test correlated .32 ($p < .001$) with the role play and .32 ($p < .001$) with the interview. Integrity on the constructed response multimedia test correlated .15 ($p = .026$) with scores on the interview.

References

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.

Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on Black–White mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology*, 66, 91–126. 10.1111/peps.12007.

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of d for range restriction and unreliability. *Organizational Research Methods*, 4, 46–61.

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, 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.

Bourdin, B., & Fayol, M. (2002). Even in adults, written production is still more costly than oral production. *International Journal of Psychology*, 37, 219–227.

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

CEBIR (2013). CEBIR tests. Available at www.cebir.be (accessed 4 July 2013).

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. (2004). An agenda for future research on applicant reactions to selection procedures: A construct-oriented approach. *International Journal of Selection and Assessment*, 12, 9–23.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

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

Cucina, J. M., Busciglio, H. H., Thomas, P. H., Callen, N. F., Walker, D. D., & Goldenberg Schoepfer, R. J. (2011). Video-based testing at U.S. Customs and border protection. In N. T. Tippins, S. Adler, & A. I. Kraut (Eds.), *Technology-enhanced assessment of talent* (pp. 338–354). San Francisco, CA: Jossey-Bass.

Dayan, K., Kasten, R., & Fox, S. (2002). Entry-level police candidate assessment center: An efficient tool or a hammer to kill a fly? *Personnel Psychology*, 55, 827–849.

De Meijer, L. A. L. (2008). *Ethnicity Effects in Police Officer Selection: Applicant, assessor, and selection-method factors*. Unpublished doctoral dissertation, Erasmus University, Rotterdam.

De Meijer, L. A. L., Born, M. P., Terlouw, G., & Van der Molen, H. T. (2008). Criterion-related validity of Dutch police-selection measures and differences between ethnic groups. *International Journal of Selection and Assessment*, 16, 321–332.

De Vries, S., & Pettigrew, T. F. (1998). Effects of ethnic diversity: The position of minority workers in two Dutch organizations. *Journal of Applied Social Psychology*, 28, 1503–1529.

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.

Dewberry, C. (2001). Performance disparities between whites and ethnic minorities: Real differences or assessment bias? *Journal of Occupational and Organizational Psychology*, 74, 659–673.

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.

Evers, A., Te Nijenhuis, J., & Van der Flier, H. (2005). Ethnic bias and fairness in personnel selection: Evidence and consequences. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *Handbook of personnel selection* (pp. 306–328). Oxford, UK: Blackwell.

Finch, D. M., Edwards, B. D., & Wallace, J. C. (2009). Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology*, 94, 318–340.

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.

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.

Hanges, P. J., & Feinberg, E. G. (2009). International perspectives on adverse impact: Europe and beyond. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 349–373). New York: Routledge.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76, 408–420.

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

Hirsh, H. R., Northrop, L. C., & Schmidt, F. L. (1986). Validity generalization results for law-enforcement occupations. *Personnel Psychology*, 39, 399–420.

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., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, 79, 184–190.

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

ICE (2005). *Handleiding DIGIBO 2003 versie 2.0*. Lienden: Bureau ICE.

International Task Force on Assessment Center Guidelines (2000). Guidelines and ethical considerations for assessment center operations. *Public Personnel Management*, 29, 315–331.

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

Klinkenberg, E. L., & Van Leeuwen, A. E. (2003). *Voortgangsverslag ontwikkeling M5Q- IWSP [Progress report of development M5Q-IWSP]*. Culemborg, Nederland: Meurs Personeelsadvies.

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology*, 86, 1202–1222.

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

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, 91, 1181–1188.

Lievens, F., De Corte, W., & Westerveld, L. (in press). Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management*, doi: 10.1177/0149206312463941.

Metz, I., & Kulik, C. T. (2008). Making public organizations more inclusive: A case study of the Victoria police force. *Human Resource Management*, 47, 369–387.

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.

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

Olson-Buchanan, J. B., & Drasgow, F. (2006). Multimedia situational judgment tests: The medium creates the message. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 253–278). Mahwah, NJ: Lawrence Erlbaum.

Ones, D. S., & Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology*, 75, 255–276.

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

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.

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.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods Instruments & Computers*, 36, 717–731.

Pynes, J. E., & Bernardin, H. J. (1989). Predictive validity of an entry-level police officer assessment center. *Journal of Applied Psychology*, 74, 831–833.

Reeve, C. L., & Bonaccio, S. (2009). Measurement reliability, the Spearman-Jensen effect and the revised Thorndike model of test bias. *International Journal of Selection and Assessment*, 17, 61–68.

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., 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.

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

Sackett, P. R., Schmitt, N., Ellingson, J. E., & 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.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology*, 88, 1068–1081.

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.

Schwartz, S. H. (2004). Mapping and interpreting cultural differences around the world. In H. Vinken, J. Soeters, & P. Ester (Eds.), *Comparing cultures: Dimensions of culture in a comparative perspective* (pp. 43–73). Leiden, Nederland: Brill Academic Publishers.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7, 422–445.

Statistics Netherlands (2011). *Figures on population by origin*. Available at http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLEN&PA=37325eng&D1=0-2&D2=0&D3=0&D4=0&D5=0-1,3-4,139,145,210,225&D6=4,9,%28l-1%29-l&HD=090611-0858&LA=EN&HDR=G3,T&STB=G5,G1,G2,G4 (accessed 22 October 2012).

Te Nijenhuis, J., De Jong, M. J., Evers, A., & Van Der Flier, H. (2004). Are cognitive differences between immigrant and majority groups diminishing? *European Journal of Personality*, 18, 405–434.

Van Iddekinge, C. H., Taylor, M. A., & Eidson, C. E. (2005). Broad versus narrow facets of integrity: Predictive validity and subgroup differences. *Human Performance*, 18, 151–177.

Van Leeuwen, A. E. (2000). *Constructie van de M5Q voor IWSP [Development of the M5Q for IWSP]*. Culemborg, Nederland: Meurs Personeelsadvies.

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

Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement, and subgroup differences. *Human Performance*, 17, 433–461.

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.