9-2015

Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg

Wilfried DE CORTE
Ghent University

Lena WESTERVELD

DOI: https://doi.org/10.1177/0149206312463941

Citation

LIEVENS, Filip; DE CORTE, Wilfried; and WESTERVELD, Lena. Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. (2015). *Journal of Management*. 41, (6), 1604-1627. Research Collection Lee Kong Chian School Of Business.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/5513

Journal of Management
Vol. XX No. X, Month XXXX xx-xx
DOI: 10.1177/0149206312463941
© The Author(s) 2012
Reprints and permission: http://www.
sagepub.com/journalsPermissions.nav

Understanding the Building Blocks of Selection Procedures: Effects of Response Fidelity on Performance and Validity

Filip Lievens
Wilfried De Corte
Ghent University
Lena Westerveld
Politieacademie

This study aims to advance our conceptual understanding of selection procedures by exploring the effect of response fidelity (i.e., written constructed response vs. behavioral constructed response) on test performance, validity, and applicant perceptions. Stimulus fidelity (multimedia stimulus) was kept constant. In a field experiment, 208 applicants for entry-level police officer jobs completed a multimedia situational judgment test with written constructed responses and behavioral responses. We hypothesized the behavioral response mode (a) to be a better predictor of police trainee performance one year later, (b) to be less cognitively saturated, (c) to exhibit higher personality (extraversion) saturation, and (d) to be perceived more positively in terms of media richness. Results suggested support for these hypotheses, although most effect sizes were not large. Implications for a building block approach to personnel selection procedures are discussed.

Keywords: *personnel selection; response fidelity; assessment center; situational judgment test*

Acknowledgments: The authors would like to acknowledge Paul R. Sackett and Britt De Soete for their helpful comments on a prior version of this article. Partial funding for this study came from the SIOP Foundation via the Small Grant Program.

Corresponding author: Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

E-mail: filip.lievens@ugent.be

Traditionally, the staffing domain has been associated with selection procedures such as cognitive ability tests, personality inventories, biographic information blanks, interviews, or assessment centers. This focus on selection procedures as holistic entities is understandable because this is how these procedures are used in operational selection practice (Schmidt & Hunter, 1998). Apart from this continued applied emphasis, in recent years calls have been made to examine and compare selection procedures less "as a whole" because such research provides little conceptual insight into their contributing factors (Arthur, Day, McNelly, & Edens, 2003; Arthur & Villado, 2008; Christian, Edwards, & Bradley, 2010). That is, when a given selection procedure is more valid than another one, it is impossible to conclusively determine the factors responsible for this result.

Therefore, recent research has tried to advance our conceptual understanding of selection procedures by making a distinction between predictor constructs and predictor methods (Arthur & Villado, 2008; Christian et al., 2010). Predictor constructs refer to the behavioral domain being sampled, whereas predictor methods denote the specific techniques by which domain-relevant behavioral information is elicited, collected, and subsequently used to make inferences (Arthur & Villado, 2008, p. 435). In addition, predictor methods might be further distinguished into different methods for presenting the stimulus information (stimulus formats) and for capturing candidates' responses (response formats) (Thornton & Cleveland, 1990; Weekley, Ployhart, & Holtz, 2006). Basically, a selection procedure is then conceptualized as a specific combination of these and other "building blocks."

Such a building block approach to personnel selection has conceptual and practical appeal. Conceptually, an advantage of a building block approach is that it allows isolating and understanding specific method factors (e.g., degree of fidelity in stimulus and response formats) across a given construct that lead to the best selection outcomes (keeping all other things constant). At a practical level, when different fidelity levels of method factors are manipulated, it enables researchers to scrutinize whether an investment in more costly predictor methodology pays off in terms of validity. In addition, a building block approach opens a window of opportunity for creatively developing hybrids of existing selection procedures by combining different building blocks into a new selection procedure.

Despite its promise for selection theory and practice, research that has isolated the effects of specific predictor methods across a given construct on selection outcomes is scarce (Ployhart, Schneider, & Schmitt, 2006). So far, only the effects of stimulus format fidelity are relatively well understood (Chan & Schmitt, 1997; Lievens & Sackett, 2006; Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000). For a building block approach to be successful and operational, it is clear that the effects of other predictor methodology components such as response format fidelity should also be scrutinized. Currently, however, our understanding of this aspect is limited (Edwards & Arthur, 2007; Ryan & Huth, 2008).

Therefore, this study focuses on response fidelity as a specific predictor method factor and explores the effects of response modes (i.e., written constructed response vs. behavioral response) on key selection outcomes such as predictive validity, construct-related validity, and applicant perceptions, while keeping the stimulus fidelity (multimedia stimulus) constant. To increase this study's external validity we decided to experiment with the response modalities in an actual selection setting. We set up a field experiment among applicants for police officer jobs and used a predictive validation design to examine their effects.

Fidelity as a Key Aspect of Predictor Methodology in Selection

The few prior studies that have aimed to isolate the impact of method factors on test performance and validity have focused on fidelity because fidelity is seen as a crucial determinant in establishing behavioral consistency and validity. Fidelity denotes the extent to which the assessment task and context mirror those actually present on the job (Callinan & Robertson, 2000; Goldstein, Zedeck, & Schneider, 1993). Most research has manipulated the degree of stimulus fidelity within a specific selection procedure, revealing that higher levels of stimulus fidelity lead to beneficial effects in terms of validity and diversity. For example, situational judgment tests (SJTs) with high stimulus fidelity (video-based items) were found to lead to higher validity (Christian et al., 2010; Lievens & Sackett, 2006), lower Black-White subgroup differences (Chan & Schmitt, 1997), and more favorable applicant reactions (Richman-Hirsch et al., 2000) than SJTs with low stimulus fidelity (paper-and-pencil items). Note that in these primary studies the test content was held constant so that the differences observed could be attributed to the differing stimulus presentation methods.

However, stimulus fidelity represents only one determinant of the behavioral consistency paradigm (Goldstein et al., 1993; Schmitt & Ostroff, 1986). Maximizing the point-to-point correspondence with the criterion, response fidelity seems at least as important (Funke & Schuler, 1998; Ryan & Greguras, 1998; Sackett, 1987). Unfortunately, response fidelity has received only scant research attention (Arthur, Edwards, & Barrett, 2002; Edwards & Arthur, 2007; Funke & Schuler, 1998). Along these lines, Ryan and Greguras (1998) noted: “Response domains are often neglected in test constructions . . . little systematic research . . . has occurred in a framework that provides guidance on the effects of response format across content domains” (p. 190). In two recent studies, Arthur and colleagues found that a written constructed response format (i.e., writing in the answer) substantially reduced subgroup differences and yielded more favorable applicant perceptions of African American candidates as compared to a multiple choice response format of the same knowledge test. Similar criterion-related validity results were found across the two formats. Funke and Schuler (1998) discovered that a written constructed response SJT format was a better predictor of performance in a high-fidelity simulation (a role-play) than a multiple choice SJT response format. In addition, this response fidelity manipulation exerted more effects than the stimulus fidelity manipulation (oral vs. video-based).

The conclusions of these two prior studies are restricted to response modes at the relatively low end of the response fidelity continuum (multiple choice vs. written constructed response). Clearly, it would be worthwhile to examine the effects of response modes across a wider range of the fidelity continuum. Therefore, this study contrasts lower levels of response fidelity (e.g., written constructed response mode) to higher levels of response fidelity (e.g., behavioral response mode). Hypotheses about the effects of response fidelity on predictive validity, construct-related validity, and applicant perceptions are posited.

Hypothesized Effects of Response Fidelity

Thornton and Rupp (2006) discussed possible differences between low-fidelity and high-fidelity simulation methods. They argued that candidate responses to low-fidelity simulation methods might primarily reflect candidates' behavioral intentions. That is, in low-fidelity simulation methods, candidates typically indicate what they intend to do but are not required to act out these intentions orally and in behavior. Similarly, Motowidlo and Beier (2010) posited that such low-fidelity simulation methods measure people's procedural knowledge about the costs and benefits of engaging in specific courses of action (see also Motowidlo, Hooper, & Jackson, 2006). Conversely, a behavioral response mode requires applicants to translate such procedural knowledge into actual behavior (Lievens & Patterson, 2011; Lievens & Sackett, 2012). In this mode, candidates display not only verbal behavior but also nonverbal and paralinguistic behavior (body language, voice tone, and inflection) and show their emotions (Borkenau, Mauer, Riemann, Spinath, & Angleitner, 2004; DeGroot & Motowidlo, 1999). These differences between the behavioral response mode and the written constructed one suggest that these response modes might correlate only moderately and that the actual samples of behavior gathered via the behavioral response mode might lead to more fidelity and a better match with criterion behavior as presented on the job. In this study, performance as a police trainee serves as criterion. The multimedia test presented interpersonal situations (e.g., with citizens or colleagues) that entry-level police officers are likely to encounter in their traineeship/job. Assessors then rated applicants' responses on various job-related dimensions such as forcefulness, communication, or sensitivity. Clearly, assessors might benefit from being able to observe whether in the interpersonal situations applicants' verbal, nonverbal, and paralingual behavior is actually forceful, communicative, or sensitive (instead of relying only on the written responses of the applicants). Thus,

> *Hypothesis 1:* Scores on a multimedia test with a behavioral response mode at the time of selection will have higher validity for predicting performance as a police trainee one year later than scores from a written constructed response mode.

Apart from predictive validity, we also investigate the response mode's effect on a test's correlation with cognitive ability and personality (see Whetzel, McDaniel, & Nguyen, 2008). So, we examine whether the response modality affects the test's saturation with either cognitive ability or personality. Saturation refers to how a given construct (e.g., cognitive ability or personality) influences a complex multidimensional measure (e.g., multimedia SJT) (Lubinski & Dawis, 1992; Roth, Bobko, McFarland, & Buster, 2008). So far, no studies have compared levels of response fidelity in terms of their cognitive and personality saturation. However, an examination of especially the cognitive saturation of different response modalities is important as it affects the status of the selection procedure as an alternative predictor method (see Ployhart & Holtz, 2008). In fact, cognitive saturation has been found to be the major driver of subgroup differences in assessment center exercises (Dean, Bobko, & Roth, 2008; Goldstein, Yusko, Braverman, Smith, & Chung, 1998), work samples (Roth et al., 2008), employment interviews (Huffcutt & Roth, 1998), and SJTs (Whetzel et al., 2008).

We expect some cognitive saturation under each of the two response formats. Individuals higher in cognitive ability should be better able to consider various courses of action. In addition, verbal ability might help in formulating one's reply, regardless whether this response is written or behavioral. That said, we do posit that the cognitive loading of written constructed responses will be higher than that of behavioral responses. Three reasons underlie this logic. One reason is that research in linguistics has shown that written production requires more cognitive resources (e.g., in terms of verbal ability and reasoning) than responding orally (Bourdin & Fayol, 2002). Another reason is that people have somewhat more time to reflect (producing a more thoughtful reaction) and think about what words to use when they write down their response. Third, research on SJT response instructions (McDaniel, Hartman, Whetzel, & Grubb, 2007) has revealed that SJTs with behavioral tendency instructions (i.e., asking people hypothetically what they would do) produce a lower cognitive loading than knowledge-based instructions (i.e., asking whether they know what to do) because responses on behavioral tendency instructions reflect typical performance: what one would do rather than knowledge of what one should do. One might expect similar effects when comparing written constructed responses to behavioral ones.

In terms of the personality saturation, we expect that applicant scores on the response mode where they are required to show (verbal and nonverbal) behavior will correlate more strongly with self-ratings on extraversion. People high on extraversion are sociable, gregarious, assertive, talkative, energetic, and active (Barrick & Mount, 1991). People with high levels of extraversion also prefer social situations in which they can interact with others (Goldberg, 1990). As verbal and nonverbal responses simulate better the nature of interpersonal exchanges than written ones, we expect that extraversion will enable applicants to score better in the behavioral response format. In the interview domain, there exists some support for these assertions as Huffcutt, Weekley, Wiesner, DeGroot, and Jones (2001) found that extraverted people obtained higher ratings in behavior description interviews. Thus,

Hypothesis 2a: Scores on a multimedia test with a written constructed response mode will be more cognitively saturated than scores from a behavioral response mode.

Hypothesis 2b: Scores on a multimedia test with a behavioral response mode will have higher extraversion saturation than scores from a written constructed response mode.

A final set of hypotheses deals with the effects of response fidelity on applicant perceptions. Research in this area is again scarce. For example, Ryan and Huth (2008: 125) lamented this dearth of research by positing that "one question somewhat ignored by the applicant reactions area is how response format influences perceptions." An exception is the Edwards and Arthur (2007) study that showed that the response mode (written constructed vs. multiple-choice format) of a knowledge test made a difference in terms of procedural fairness.

Although applicant perceptions are typically conceptualized in terms of procedural fairness dimensions (e.g., job relatedness), in this study we do not expect that different response modes associated with a multimedia test will affect such traditional procedural fairness dimensions because multimedia tests (as compared to the knowledge tests used in Edwards & Arthur, 2007) typically receive already high job relatedness ratings (e.g., Kanning, Grewe, Hollenberg, & Hadouch, 2006; Lievens & Sackett, 2006). On conceptual grounds, we do expect that changing the response mode of such a multimedia test might affect other applicant perceptions. In particular, we posit an effect of response fidelity on applicants' perceptions of media richness. It is relevant and important to apply a communication theory such as media richness theory in this study because response modes differ from each other in how they enable people to communicate information.

Media richness theory deals with how information is conveyed and communicated, with ambiguity serving as the central concept (Daft & Lengel, 1984; Fulk & Boyd, 1991; Sitkin, Sutcliffe, & Barrios-Choplin, 1992; Trevino, Lengel, & Daft, 1987). Communication media are then ordered along a continuum of media richness on the basis of their capacities to resolve this ambiguity. That is, richer media can be distinguished from leaner media on the basis of four specific factors: (1) opportunity for two-way communication, (2) ability to convey a multiplicity of cues (verbal, nonverbal, and paralinguistic), (3) ability to convey a sense of personal focus, and (4) the use of natural language. Essentially, these factors refer to the medium's ability to carry a variety of data (e.g., aural cues, visual cues, text cues) and the medium's ability to carry symbolic information (e.g., emotions) about the individuals who are communicating. The basic premise of media richness theory is that communication media are most effective when they match the ambiguity level of the task. In other words, the medium should fit the type of message. Richer media (e.g., face-to-face contact, video) should be used when the ambiguity is high, whereas leaner media (e.g., text) suffice when ambiguity is low. Potosky (2008) further proposed that aspects of and perceptions about a medium are related to people's beliefs about how well they can express themselves and to potential higher/lower frustration during the communication process. Hence, perceptions of media richness bear resemblance to positive/negative judgments about "voice"-related procedural fairness dimensions (e.g., opportunity to perform perceptions).

In this study, the job-related situations in multimedia tests typically contain a high degree of ambiguity because they present practical contextualized problems to applicants that are ill defined and incomplete in information and have multiple possible solutions (Weekley et al., 2006). Faced with such high ambiguity and with communicating through less rich media (i.e., written mode), we expect that applicants will be frustrated by what they are able to express as they cannot get their "full" message across. In other words, communicating via a written medium might impede applicants from matching their replies with the ambiguity level inherent in multimedia stimuli. Conversely, we expect that a behavioral response mode will enable applicants to reply better to multimedia stimuli because the behavioral response mode allows them to use multiple communication channels (verbal, nonverbal, and paralinguistic) and emotions. Thus,

Hypothesis 3: Applicant perceptions of media richness will be significantly higher for a multimedia test with a behavioral response mode than for the same test with a written constructed response mode.

Method

Sample and Procedure

The police academy agreed that we could run a field experiment (see the following) for a period of 4 months in 2008. In that period, each day four randomly chosen applicants were asked to complete this study's multimedia test with the different response formats. This produced a sample of 208 applicants (126 men, 82 women; mean age = 22.8 years). These individuals applied for an entry-level police officer job in the Netherlands. There were

13 minority candidates. Most of them were from Moroccan and Turkish backgrounds. None of the candidates had prior police work experience. All candidates had finished high school. To examine this sample's representativeness, we compared its characteristics to the entire population of applicants in 2008. No significant differences in terms of gender, age, or ethnicity were found.

Candidates were screened on the basis of their resume. Next, eligible candidates were invited to take part in a 2-day selection procedure in the recruitment and selection center of the police academy. During those 2 days, candidates completed various tests (verbal reasoning tests, language proficiency tests, physical ability tests, structured behavioral description interview, role-play exercise). Some candidates (N = 86) also completed a personality inventory. Of the 208 candidates that participated in the field experiment, 81 candidates were hired, producing a selection ratio of about 35%.

The multimedia test with the different response formats was always completed at the end of the 2-day selection process. Applicants were informed that this test was still in its experimental stages and had no bearing on hiring decisions. Results were also not disseminated. Prior to starting the actual test, applicants were given a practice multimedia item. Next, candidates answered 24 video scenes (12 scenes in a written constructed response and 12 scenes in a behavioral response mode, see design described in the following). Finally, participants completed a posttest questionnaire for measuring their perceptions of the different response fidelity modes. Anecdotal evidence (comments of participants and their general test-taking attitude) showed that participants completed the tests with a similar motivational level as the other tests.

Although we would have wished to run the field experiment a bit longer, a 4-month window was the maximum because this experiment demanded substantial resources and efforts from the police academy. For instance, the police academy had to organize half-day assessor training sessions for this new type of selection procedure. Experienced assessors of the selection center of the police academy attended these sessions during their working hours. Next, all vignettes of the 208 candidates (webcam as well as written vignettes, totaling 4,992 vignettes) of this field experiment had to be rated by two of these trained assessors during working hours. These ratings were used only for research purposes.

Design of Field Experiment

In the two formats, applicants were "guided" through the administration by a narrator as instructions and video scenes were on the PC. So, there was a predetermined pace and people could not backtrack. Applicants were presented with the scenes on the screen of a laptop one by one. At the end of each scene, the character in the scene looked into the camera as if he or she was directly addressing the applicant and the scene froze. That was the point when the two formats differed. In *written constructed responses*, candidates wrote down their answer in a text balloon. In *behavioral responses*, candidates responded to the character. Their behavioral reply was captured by a webcam mounted on the PC.

A within-subjects design was used. So, each applicant received 12 video scenes that had to be answered by writing down the response and 12 video scenes that had to be answered by responding via verbal and nonverbal behavior. To avoid confounds (scene effects, order effects, fatigue effects, and practice effects), various sets of the multimedia test were constructed with the use of the same set of the original 24 scenes so that the specific scene and the order of the scene were counterbalanced. In addition, half of the applicants provided written responses to the first 12 scenes and behavioral responses to the last 12 scenes, whereas the opposite was true for the other half of the applicants.

Development of Multimedia Test

We followed existing procedures for developing the multimedia test (see Chan & Schmitt, 1997; Weekley & Jones, 1997). First, a thorough job analysis was conducted to determine the Knowledge, Skills, Abilities, and Other Characteristics (KSAOs) related to entry-level police officer jobs. Nine key KSAOs were identified. The general aim of the SJT was to present situations related to four of these dimensions, namely, interpersonal sensitivity, forcefulness, integrity, and communication.

Second, interviews were held with 15 police officers and police sergeants (3 women, 12 men; 4 minorities) to gather critical incidents. Redundant incidents were removed and the remaining incidents were grouped in categories. A sensitivity review panel (Ployhart & Holtz, 2008) also screened the incidents for language and cultural sensitivity. After this stage, 70 nonredundant critical incidents were left.

Third, these critical incidents served as basis for writing item stems and item options. These items were then given back to the interviewed police officers and sergeants. Only item stems and item options that were rated as realistic and that were considered to trigger the KSAOs targeted were kept, resulting in a set of 50 items.

Fourth, these 50 items were tested in a sample of 228 candidates (165 men, 62 women; 19 minorities; mean age = 24.08 years, *SD* = 6.78). This pilot was used to gather indications on the difficulty of the items and the endorsement frequency of the item options. Items (e.g., car accident) that were too costly to film were also eliminated. Thirty-one items survived this pilot.

Finally, these 31 scenarios were filmed using professional actors and a recording company. The recording took place at different sites. In this study, 24 video-based items were selected from this available set (i.e., these video-based items were randomly selected from the available items per dimension category) and edited in a multimedia test. In its final form, this study's multimedia test consisted of short, videotaped vignettes of key situations (about 1 to 2 minutes) that police officers are likely to encounter. All situations were interactions between police officers and either citizens or colleagues. There were eight scenes that triggered behavior related to interpersonal sensitivity, eight scenes related to forcefulness, and eight scenes related to integrity. For none of the scenes, prior knowledge about police work was needed. Examples were situations related to domestic disturbance (for triggering forcefulness), civilians needing help or being in distress (for triggering sensitivity), and the misuse of position or power (for triggering integrity). For instance, the domestic disturbance scenario related to a phone call from a man who tells the police there is a lot of noise at his neighbor's house. The man thinks there is a fight going on. Therefore, the police go to the neighbor's house. However, the neighbor does not want the policemen to come in, shouts at them ("get lost"), and threatens them.

Rating Process

A group of 20 experienced selection officers (5 males and 15 females) received a half-day training prior to rating the vignettes. This training included a lecture (on the basis of frame-of-reference training), practice, and feedback. In the actual study, the written and webcam vignettes were randomly distributed to these assessors who then used behavioral anchored rating scales (BARS) for evaluating the applicants. The anchors of these scales were specifically developed to reflect key behaviors to be elicited from the specific stimuli in the situations. Depending on the scene presented, they rated candidates either on interpersonal sensitivity (example behavioral anchors included: listens to the other, offers help), forcefulness (e.g., is clear about what needs to happen, discusses consequences of behavior), or integrity (e.g., does not accept presents, whistle blows). Communication (e.g., provides arguments for actions/opinions, communicates clearly and comprehensively) was assessed in each of the 24 scenes. The BARS used were the same across the two response modes. Each response was rated by two assessors on a 5-point scale, with 1 = *poor* and 5 = *excellent*. For example, a response related to the domestic disturbance scenario that was rated positively (on forcefulness) was a response of an applicant who tells the angry man that he or she has to take a look inside the house one way or the other. Conversely, an example response that was negatively rated was one of an applicant who tells the angry man that he or she will stop by later. Apart from the dimension ratings, we also computed a composite rating that was the average of the dimension ratings.

Other Measures

Verbal reasoning tests. Two computer adaptive tests were used for measuring verbal reasoning. The first test required applicants to find the underlying principle in a configuration of letters (mean test time of about 30 minutes). In the second test, candidates were presented with verbal analogies (mean test time of about 12 minutes). In this study, we used a composite score that was the average of the two tests (the tests correlated .315). As these verbal reasoning measures were the publisher's property, we received only candidates' final scores and were not able to compute internal consistencies. Prior research mentioned in the test manual indicated internal consistencies around .90 (CEBIR, 2009).

Personality inventory. Extraversion was measured with a 10-item scale from the M5Q. This is a measure of the Big Five personality factors (Klinkenberg & Van Leeuwen, 2003; Van Leeuwen, 2000). Given that the M5Q is the property of the publisher, we received only candidates' standardized composite scores on this scale. Hence, we were not able to compute the reliability of the scale. In the test manual, an internal consistency reliability of .85 and a test-retest reliability of .94 were mentioned.

Posttest questionnaire. Upon completion of the multimedia test, candidates were asked to fill out a short questionnaire about their perceptions of the response modes. Among other questions, this inventory measured applicants' perceptions of media richness, satisfaction, and excitement. All items were rated twice (for each of the two response modes) on a 5-point scale, ranging from 1 = *strongly disagree* to 5 = *strongly agree*. To measure media richness we used six items of Webster and Trevino (1995). An example item was: "During the test I was able to convey multiple types of information (verbal and nonverbal)." Applicants' satisfaction (four items; e.g., "Participation in the test has been a positive experience") and excitement (five items; e.g., "I found the test to be exciting") with the response modes were measured with scales developed by Richman-Hirsch et al. (2000). All internal consistencies of the scale scores in the posttest questionnaire were satisfactory (above .70).

Criterion Measures

In this study, three criterion measures were used. The first criterion was an internal criterion measure. As participants' scores on the response modes to the multimedia test had no bearing on the selection decision, the actual selection decision (i.e., being hired vs. not being hired) served as the first criterion. This hiring decision was based on participants' scores on all administered selection procedures (with the exception of candidates' multimedia test scores on the different response modes) and subjective judgments made by the selection board.

The other two criterion measures were external criterion measures. We gathered two training performance measures, namely, training performance ratings and training performance test scores. Upon being hired, the selectees received training in the police academy for 2 years. This training involved lectures, simulations, workshops, assignments, and on-the-job training. We were able to collect training performance ratings because at the start of the training each hired candidate was paired with an experienced coach/mentor. Coaches had at least weekly contact with the trainees. None of them was familiar with candidates' scores on the multimedia test. One year after the hired candidates had enrolled into the police academy, coaches were sent an online survey and were asked to rate the hired candidates on task (i.e., each of the dimensions targeted, quantity of work, and quality of work) and contextual performance (i.e., work dedication) dimensions using the relative percentile method (Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996; Goffin, Jelley, Powell, & Johnston, 2009). In the relative percentile method, raters are asked to assign percentile scores to ratees on the performance dimensions targeted. Goffin and colleagues developed this method to reduce rating inflation as raters are told that the reference group to be used for the ratings consists of the average trainee (i.e., a percentile score of 50). Prior research showed that the relative percentile method had higher criterion-related validity than conventional absolute rating formats (Goffin et al., 1996, 2009). In this study, we received usable responses of 64 of the 81 hired candidates (79% response rate; after four reminders). The internal consistency of the training performance ratings was .94 (a one-factor solution also emerged from an exploratory factor analysis). Therefore, this study uses a composite training performance rating. As 24 trainees were each rated by two trainers, we were also able to compute the interrater reliability of the ratings of this subset of trainees. Interrater reliability was .62. When two ratings of a candidate were available, the average was used in the analyses.

As a second training performance measure, we collected training performance test scores. To obtain a standardized measure of their performance and progress in the training program, trainees were required to participate in performance tests (after about 12 months). For instance, in one performance test (work sample), trainees were required to actually patrol on a large square while being observed by trained raters. These performance tests were not conducted by the police academy but by an external organization with trained raters. None of these raters was familiar with candidates' scores on the multimedia test. In this study, we were able to retrieve overall performance test scores of 75 of the 81 hired candidates (93%) from the archives of this external organization. The missing scores belonged to individuals who already had left the police academy. The performance test scores were standardized within administration. The two training performance criteria (training test scores and training performance ratings) correlated .28 ($p < .05$).

Range Restriction

As noted previously, 81 of the 208 candidates were hired, with the hiring decision being based on people's scores on all selection procedures administered (except the multimedia test scores related to the response modes) and other unmeasured variables (e.g., unquantified subjective judgments made by the selection board). Therefore, in the range restriction typology of Sackett and Yang (2000), scenario 2d (unrestricted variance of test scores known; unmeasured decision variable, which is a composite of the measured selection procedure scores, and some unmeasured additional selection process) applied. It is inappropriate to correct for this type of range restriction with indirect (Schmidt, Oh, & Le, 2006) or multivariate range restriction correction formulas (Ree, Carretta, Earles, & Albert, 1994). Instead, the Heckman (1976, 1979) two-stage procedure as well as the maximum likelihood approach as proposed by Sackett and Yang (2000) should be used. However, it was not possible to accurately model the hiring decision on the basis of the available measured variables because these two approaches produced diverging and counterintuitive results, indicating that the assumptions underlying the modeling process were not met (see Winship & Mare, 1992). Hence, we could not make range restriction corrections, and the validities of this study should be regarded as conservative. Note that the scarce studies that were able to successfully apply these specific range restriction correction approaches were all based on simulated (instead of actual) data.

Although we were not able to statistically correct for range restriction, we did examine the potential effects of range restriction. To this end, we inspected the correlations between people's scores on the operational predictors and their hiring decision. Scores on the two predictors related to this study's hypotheses (i.e., verbal reasoning and extraversion) did not correlate substantially with the hiring decision (.07 and .11, respectively). The hiring decision was driven by scores on two predictors that were not related to our hypotheses (i.e., behavioral description interview and role-play exercise). This is important because it indicates that our inability to correct for this specific type of range restriction does not affect finding support for some of the hypotheses.

Table 1
Reliability Results Broken Down by Response Mode (*N* = 208)

	ICC (1, 2)		Cronbach's alpha	
	Written	Behavioral	Written	Behavioral
Communication	$.71_a$	$.70_a$	—	—
Sensitivity	$.71_a$	$.75_a$	$.67_a$	$.69_a$
Forcefulness	$.70_a$	$.64_a$	$.76_a$	$.79_a$
Integrity	$.60_a$	$.65_a$	$.60_a$	$.67_a$
Total	$.81_a$	$.78_a$	$.80_a$	$.83_a$

Notes: Dashes indicate that internal consistency reliability could not be computed because this dimension was not rated on the basis of separate items (i.e., behavioral anchors). Within a row, values for intraclass correlations (ICCs) and alphas with the same subscripts do not differ significantly from each other. We used the formula in Feldt (1980) for testing the statistical difference of two dependent Cronbach's alphas. For testing the statistical difference of two dependent ICCs, we inspected the overlap in confidence intervals around the ICCs.

Results

Preliminary Analyses

We started by conducting preliminary analyses to rule out a series of alternative explanations for possible differences among the two response formats. First, we examined whether there were any effects of the set of videos that participants received. As noted earlier, we constructed different sets (all consisting of the same 24 scenes, but in a different order and with different response modes required) to avoid possible confounds (order, scene, fatigue, practice effects). There was no significant effect of set across the different response modes.

Second, we examined whether assessors' agreement in rating applicants' responses differed across the two formats. Table 1 presents intraclass correlation coefficients (ICC 1, 2) for assessors' dimension ratings and for the composite rating. In both response formats, agreement was satisfactory. There were negligible differences between the two formats.

Third, we computed the internal consistency reliability of assessors' ratings to determine whether scenes that trigger a similar dimension provide a consistent measurement of that specific dimension and whether there are enough scenes for such a consistent measurement to occur. Table 1 shows that the internal consistencies were acceptable (only a more ambiguous dimension such as integrity scored a bit lower). Consistency in measurement was somewhat higher for scores in the behavioral response mode. However, the differences among the response modes were not significant. Logically, internal consistency was highest for the overall scores (mean of the dimension ratings).

Fourth, we conducted multiple group measurement invariance analyses using EQS (Bentler, 1995). In these analyses, we examined whether the measurement structure underlying candidate scores on the behavioral mode was equivalent to the measurement structure underlying candidate scores on the written constructed mode. Assessors' dimension ratings of candidate performances on each of the scenes served as indicator variables. Prior to testing the measurement invariance, we specified a baseline model. In particular, the model

Table 2
Tests of Measurement Invariance for Multiple Dimension Mode Underlying Multimedia Test Scores Across Response Modes (*N* = 208)

	χ^2	*df*	RNI	CFI	IFI	RMSEA	90% CI of RMSEA
Equal number of factors	271.148**	204	.940	.953	.955	.056	.036-.073
Equal factor loadings	299.347**	213	.926	.940	.942	.063	.045-.078
Equal factor covariances	308.127**	216	.922	.936	.938	.064	.047-.079
Equal error variances	315.591**	228	.926	.939	.941	.061	.043-.076

Notes: RNI = Relative Noncentrality Index; CFI = Comparative Fit Index; IFI = Incremental Fit Index; RMSEA = root mean square error of approximation; CI = confidence intervals.
***p* < .01.

of interest was a confirmatory factor analysis model wherein each of the scene ratings was specified to load on their designated dimension factor, implying that the respective scenes indeed triggered the relevant dimension. This multiple-dimension model was contrasted to a one-dimension model. In each of the two response modes, the multiple-dimension model provided a good fit to the data, whereas the one-dimension model produced a poor fit. In addition, in the multiple-dimension model all of the scenes had significant factor loadings on the dimension they were purported to measure. This suggests each dimension could be adequately rated via its respective scenes. After establishing the multiple-dimensions model as the baseline model, we continued with testing the measurement invariance of this multiple-dimension model across participants' response mode scores. To this end, we conducted an increasingly restrictive series of measurement invariance tests (form, metric, and error invariance; Vandenberg & Lance, 2000). Results of these measurement invariance analyses are presented in Table 2. There was evidence of form, metric, and error invariance for the multiple-dimensions model across response mode scores. This implies that the measurement structure (factor structure, use of the intervals of rating scale) underlying candidate scores on the multimedia test was invariant across the response modes.

Finally, we examined whether there were significant differences among applicants in their satisfaction and excitement of the two formats. There were neither differences in satisfaction, $M_{\text{Behavioral}} = 3.27$, $SD = .76$; $M_{\text{Written}} = 3.33$, $SD = .75$, $t(180) = -1.82$, $p = .07$; nor in excitement, $M_{\text{Behavioral}} = 3.19$, $SD = .71$; $M_{\text{Written}} = 3.18$, $SD = .70$, $t(180) = .13$, $p = .90$.

Descriptive Statistics

Table 3 presents the means and standard deviations of applicants' scores on the multimedia test in the two response modes. Applicants scored significantly higher in the written response mode as compared to the behavioral response mode ($d = -.23$). This makes sense as it is generally more difficult to show actual behavior instead of simply writing it down. Significant differences were found for scores on two dimensions: sensitivity and forcefulness, with effect sizes being small to moderate ($d = -.20$ and $d = -.30$, respectively). The correlations between scores associated with the two response modes were significant,

Table 3
Descriptive Statistics of Response Modes (*N* = 208)

	Behavioral		Written				
	M	*SD*	*M*	*SD*	*d*	*t*	*p*
Communication	2.82	.73	2.90	.70	–.11	–1.44	.15
Sensitivity	2.70	.68	2.84	.70	–.20	–2.52	.01
Forcefulness	2.66	.69	2.87	.65	–.30	–3.45	.00
Integrity	3.11	.64	3.19	.62	–.13	–1.64	.10
Total	2.82	.56	2.95	.54	–.23	–3.02	.00

varying from .21 to .39 across the four dimensions (see Table 4). Overall, the correlation between total scores from the two response modes was .41 (.52 when corrected for unreliability using the ICCs for the overall ratings of Table 1). This correlation might be somewhat lower than expected due to the within-subjects design used (i.e., the correlation between candidates' scores on the two response modes is based on different video-based scenes).

We also examined whether there were differences by gender and ethnicity in candidate scores on the response modes. In the written response mode, female applicants outperformed male applicants on the overall score (*d* = .13) and on all dimension scores. However, the difference on the overall score between male and female candidates was not significant. In the behavioral response mode, male applicants significantly outscored female applicants overall (*d* = –.31) and on all dimensions with the exception of sensitivity. Regarding ethnicity, the majority group scored significantly better than the minority group in both response modes (*d* = .43 and *d* = .44 in the written and behavioral response mode, respectively). However, these ethnicity results should be interpreted with extreme caution in light of the small number of minorities (*N* = 13) in the sample.

Test of Hypothesis 1

According to Hypothesis 1, scores on a multimedia test with a behavioral response mode would obtain higher predictive validity than test scores based on a written constructed response mode. As shown in Table 4, the correlation between the overall score on the behavioral response mode and the three criteria was significant (.31, .30, and .26). Conversely, the correlation between the overall written constructed response score and these criteria was not significant (.01, .08, and .19, respectively). Next, we tested whether the respective dependent correlation coefficients differed significantly from each other. In light of the aforementioned challenges in setting up this field experiment with actual candidates and assessors (see Sample and Procedure section), our sample size was not large and therefore the statistical power for testing differences between these correlation coefficients was low. Therefore, we used a more liberal alpha level of .10. We also computed effect size measures. We used Cohen's (1988) rule of thumb defining *d* = .20 (i.e., accounting for 1% of variance) as a small effect. The validity of applicants' test scores across the response mode

Table 4
Construct-Related and Criterion-Related Validity Results

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Behavioral mode (*N* = 208)																
1. Communication	2.82	.73	.57**													
2. Sensitivity	2.70	.68	.67**	.47**												
3. Forcefulness	2.66	.69	.58**	.43**	.52**											
4. Integrity	3.11	.64	.88**	.76**	.83**	.77**										
5. Total	2.82	.56														
Written mode (*N* = 208)																
6. Communication	2.90	.70	.38**	.31**	.25**	.27**	.37**									
7. Sensitivity	2.84	.70	.35**	.32**	.25**	.22**	.36**	.68**								
8. Forcefulness	2.87	.65	.28**	.19**	.21**	.30**	.31**	.62**	.48**							
9. Integrity	3.19	.62	.23**	.17*	.20**	.39**	.30**	.59**	.45**	.50**						
10. Total	2.95	.54	.38**	.31**	.28**	.36**	.41**	.89**	.81**	.79**	.77**					
Other measures																
11. Verbal reasoning (*N* = 208)	–26.83	77.99	.10	.00	–.04	.07	.04	.19**	.10	.14*	.11	.16*				
12. Extraversion (*N* = 104)	.00	1.00	.38*	.36**	.31**	.17	.38**	.03	–.04	.04	–.01	.01	–.02			
Criteria																
13. Selection decision (*N* = 208)	.39	.49	.15*	.15*	.10	.06	.14*	.03	.00	–.01	–.07	–.02	.11	.19		
14. Training performance test (*N* = 75)	–.08	.61	.30**	.27*	.21	.18	.30**	.09	.13	–.09	.12	.08	.13	.15	—	
15. Training performance rating (*N* = 64)	6.14	1.94	.21	.25*	.12	.26*	.26*	.08	.25*	.09	.19	.19	.19	.25	—	.28*

Notes: Dashes indicate that the correlation could not be computed because one of the variables (selection decision) was a constant in the analyses with the two training performance criteria: Training performance criterion data were available only from people being hired.

*p < .05

**p < .01

scores for predicting the selection decision and the training test scores was significantly different at $p < .05$, $t(205) = 4.21$, and $p < .10$, $t(72) = 1.82$, respectively. The validity difference between the response mode scores was not significant for predicting the training performance ratings, $t(61) = .47$, *ns*. The effect sizes associated with the differences in predictive validity of the response mode scores showed that the behavioral response mode scores explained 9.7% (= $.31^2 – .01^2$, with rounding), 8.5% (= $.30^2 – .08^2$), and 2.8% (= $.26^2 – .19^2$) more variance in the three criteria, respectively.

As another way of testing Hypothesis 1, we investigated the incremental validity of the response modes over one another. To this end, we conducted hierarchical regression analyses with each of the criteria as dependent variables. In the first set of analyses, we entered the cognitive ability (verbal reasoning) test score first (see Schmidt & Hunter, 1998), followed by the overall score on the written constructed response mode. The overall behavioral response mode score was entered as the last block. The behavioral response mode score explained an additional variance of 2.8% ($p < .05$) in the selection decision, 8.3% ($p < .05$) in the training test scores, and 3.4% (*ns*) in the training performance ratings. In the second set of analyses, cognitive ability was again entered first, but the order of the last two blocks was reversed (the overall behavioral response mode score was entered before the written constructed response one). In that case, the overall score on the written constructed response mode explained negligible amounts of incremental variance (at most 1.0%) in each of the three criteria.

In short, for two of the three criteria, the behavioral response mode score significantly outperformed the written constructed one in terms of predictive validity and incremental validity. However, given the small *N* and the low power, it is important to interpret these results cautiously. Therefore, we interpret these results as *suggestive* of the possibility of an effect and of support for Hypothesis 1.

Test of Hypothesis 2

Our second set of hypotheses dealt with the cognitive and personality saturation of the response mode scores. Table 4 shows that only the overall written response mode score had a significant correlation with cognitive ability (verbal reasoning), $r = .16$, $p < .05$. The correlation between the overall behavioral response mode score and cognitive ability (verbal reasoning) was not significant, $r = .04$. The difference in correlation coefficients was significant only at the more liberal $p < .10$, $t(205) = 1.68$. The effect size associated with this difference showed that cognitive ability (verbal reasoning) explained 2.6% (= $.16^2 – .04^2$) more variance in the written response mode score than in the behavioral response mode score. In light of the small *N* and the low power, we use the same cautionary language as we did with Hypothesis 1 and interpret these results as suggestive of support for Hypothesis 2a.

Table 4 further shows that the overall behavioral response mode score correlated significantly with extraversion ratings ($r = .38$, $p < .01$). Conversely, the correlation between the overall written response mode score and extraversion equaled .01 (*ns*). The correlations between extraversion ratings and the respective response mode scores were significantly different from each other, $t(106) = –3.86$, $p < .01$. In terms of effect size, extraversion explained 14.2% (= $.38^2 - .01^2$) more variance in the behavioral response mode score than in the written response mode score. This result lends clear support to Hypothesis 2b.

Test of Hypothesis 3

Hypothesis 3 posited applicant perceptions of media richness to be significantly higher for the behavioral response mode than for the written constructed one. In support of Hypothesis 3, the response modes differed significantly in terms of media richness perceptions, with the behavioral response mode ($M = 3.08$, $SD = .65$) receiving significantly higher media richness ratings than the written response mode, $M = 2.95$, $SD = .60$, $t(183) = 3.12$, $p < .01$, $d = .21$.

Discussion

Main Conclusions

The distinction between content and method has emerged as an important recent trend in personnel selection research (Arthur & Villado, 2008). This study fits in this trend. Instead of focusing on entire selection procedures such as assessment centers or situational judgment tests, this study starts to explore the isolated impact of differences in response modality as a key method factor on a series of substantive selection outcomes.

As a first important result, our results suggest that only the behavioral response mode emerged as a significant predictor of performance as a police trainee. Apparently, in this kind of job only behavioral samples of applicants provide assessors with enough cues to make significant predictions about their future performance. Consistent with the validity results, candidates also perceived the behavioral format to be richer for communicating their reply than the written format. In line with media richness theory, candidates might have perceived the behavioral response mode to be a closer match to the stimulus format (multimedia) as it enabled them better to convey (non)verbal information, paralinguistic cues, and emotions and use more natural language. Moreover, the incremental validity analyses showed that the behavioral response mode explained between 2.8% and 8.3% of additional variance over and above the written response mode. Although caution is warranted in interpreting these results due to the small sample size and low power, we also believe these effects of response fidelity are not negligible as these results pertain to two high-fidelity formats (the same multimedia test with either a written or behavioral format). By way of a comparison, would one expect a traditional open-ended in-basket to explain similar portions of additional variance above a multiple-choice PC in-basket? In addition, differences in measurement/factor structure and reliability cannot explain the validity differences between the response modes because these modes were measurement equivalent and produced similar reliabilities.

More broadly, these predictive validity results are consistent with additive models of the impact of different information channels (Archer & Akert, 1980; Gesn & Ickes, 1999). According to these additive models, adding extra sources of information (verbal, nonverbal, and paralinguistic behavior) pays off and leads to higher fidelity and validity. Additive models have also received support in interview research. For instance, interviewees' visual and aural cues were found to provide valid information over and above the verbal content of their answers (Burnett, Fan, Motowidlo, & DeGroot, 1998; DeGroot & Motowidlo, 1999).

Second, the different cognitive/personality saturation of the two response modes suggests that a method factor such as response fidelity does not reflect only error variance. Instead, the response mode used might change the *substantive* (construct-related) relationships of a test with other tests. Accordingly, this study begins to answer calls to shed light on method variance in selection procedures by explicitly modeling sources of substantive variance in predictor scores (e.g., Ployhart, 2006; Schmitt, 1994). In particular, this study indicated that the correlates of the two response modalities were somewhat different. The behavioral response mode had a higher loading on the personality trait of extraversion. Although this finding should be interpreted with caution due to the smaller sample of candidates taking the personality inventory, it provides preliminary evidence that candidates who are sociable, gregarious, assertive, talkative, energetic, and active obtain higher ratings in the behavioral response mode. Thus, a behavioral response mode might be more relevant when extraversion-related traits matter on the job. Although the multimedia test generally exhibited a low correlation with cognitive ability (verbal reasoning), this correlation became significant for the written response mode scores. So, using a written response mode places a multimedia test a bit less in the “alternative” (noncognitive) domain. These results contribute to the literature on alternative test formats. So far, the stimulus format (video vs. text) had already been identified as an important driver behind the cognitive saturation of SJTs (Lievens & Sackett, 2006). This study suggests that the choice of the response mode might also affect the test’s cognitive saturation. Again, caution is warranted regarding these interpretations because the differences in cognitive (verbal) saturation found in this study were small. So, future studies are needed to replicate this effect in larger samples and with a broader general mental ability measure (consisting of domains other than the verbal domain).

Although not the main focus of our study, this study also revealed an interesting effect of response fidelity on gender subgroup differences. In the behavioral response mode, male candidates significantly outperformed female candidates, whereas such a result was not found in the written constructed response mode (with the exception of sensitivity). One reason might be that behavioral responses give away an applicant’s gender and other characteristics (e.g., physical attractiveness), thereby invoking potential stereotypes among the assessors. This result maps well into results of an assessment center study conducted in a similar male-dominated setting (UK army) where females were rated higher only on more communal dimensions (Anderson, Lievens, Van Dam, & Born, 2006). It is important to note, though, that this study does not enable one to determine whether potential subgroup (gender, ethnicity) differences are due to response mode differences or performance differences among candidates. Only when one holds the response mode constant (by transcribing the webcam vignettes) can one disentangle these two explanations. So, future research with larger samples is needed to shed light on the degree of gender differences in response mode scores. Such an examination is especially important in jobs (e.g., police officer job positions) that have traditionally been held by males.

Implications for Selection Theory and Research

Arthur and Villado (2008) called for programmatic research on the separate effects of predictor methods. This study's results about response fidelity provide a step in this direction by filling in blank cells in the predictor methodology map. In the end, similar predictor methodology studies should enable us to increase our conceptual understanding of a variety of different stimulus and response formats and their effects on selection outcomes (Ployhart et al., 2006; Thornton & Rupp, 2006). Knowledge of the effects of these and other building blocks of selection procedures might then serve as much-needed conceptual and evidence-based guidance for developing new selection procedures and for better understanding method variance in selection procedures.

Apart from this main implication, we highlight two other avenues of research that we see as important. First, future studies should scrutinize whether response fidelity provides additional insight into the diversity-validity dilemma. On the one hand, we anticipate subgroup differences might be lower for behavioral response format scores as compared to written constructed ones. One reason is that this study found that the written response mode scores had a higher cognitive (verbal) saturation. As another reason, the behavioral response mode emphasizes orality and social relations, which are typically highly valued in African American populations (Chan & Schmitt, 1997; Helms, 1992). On the other hand, it is also possible that a behavioral response format leads to more subgroup differences in applicant scores because this format gives away the candidate's gender and ethnicity. As noted, empirical research that compares a variety of response modes (written, only oral, and behavioral) by keeping candidate performance constant is needed to put these different arguments to the test.

Second, the link between faking and response fidelity deserves attention in future research. Currently, faking has been studied mainly at the level of selection procedures. For instance, we know that assessment center exercises are less prone to faking than employment interviews (McFarland, Yun, Harold, Viera, & Moore, 2005), which in turn seem less fakable than personality inventories (Van Iddekinge, Raymark, & Roth, 2005). Unfortunately, in this line of research the underlying constructs typically varied across the selection procedures. An intriguing issue for future research is to examine faking effects at a more fine-grained level. For instance, are scores on the behavioral response mode less fakable than scores on the written one (keeping other things constant)? Several arguments (e.g., more spontaneous reaction, less prone to test-taking skills) hint in this direction but they need to be empirically tested.

Implications for Selection Practice

From a practical perspective, this study provides some guidance for predictor methodology choice when designing selection procedures. In particular, it is important to determine whether investments in more costly higher response fidelity formats (behavioral response mode) pay off as compared to less expensive lower fidelity counterparts (written constructed response mode). So far, this utility question was answered by comparing the validity of low-fidelity simulations (SJTs) to that of high-fidelity simulations (assessment center exercises; Lievens & Patterson, 2011). This study begins to provide a more fine-grained answer to this utility issue by keeping stimulus fidelity constant and varying response fidelity. Our results suggest that lowering response fidelity results in lower predictive validity. Of course, these gains should be weighed against the estimated costs in developing and administering the response modes. As the same multimedia test was used as stimulus material in this study, this cost is constant across the response modes. In this setting, it took one assessor on average 40 minutes to score behavioral replies to 24 scenes (i.e., an average of 100 seconds per response), whereas it took one assessor on average 35 minutes to score the written replies to 24 scenes (i.e., an average of 87.5 seconds per response), indicating a 14% increase in assessor time in the behavioral response mode. Apart from these assessor time costs, the cost of creating webcam vignettes should also be factored in.

This study is also relevant for practitioners who are developing selection procedures that take advantage of newer technologies within their assessment tools. As broadband Internet connectivity becomes more prevalent and multimedia tools become easier to deploy, practitioners are increasing the pace of their innovations and experimentation with response formats. In fact, as a by-product of this study's experimental design, initial insights were generated related to the validity of two such new selection procedures (a multimedia test with a written constructed response format and a multimedia test with a behavioral response format). These selection procedures are hybrids between SJTs and assessment center exercises. These selection procedures provide practitioners with a web-based, standardized, job-related, and attractive platform for candidates to show their skills. Furthermore, they answer calls for "new" instruments that capture interpersonal job performance aspects (Dayan, Kasten, & Fox, 2002; Gowing, Morris, Adler, & Gold, 2008; Oostrom, Born, Serlie, & Van der Molen, 2010). On the basis of this study's encouraging results, the police academy is considering implementing the multimedia SJT with a behavioral response mode in the future.

Limitations

A first limitation relates to the setting of this study. Our study was conducted in a police officer selection context in the Netherlands. Although it is rare that one is able to set up a field experiment with actual applicants in a real selection context, this also came with some disadvantages and practical constraints. As already discussed, we would have wished our initial sample size to be larger. Therefore, we present our results as *suggestive* instead of as *conclusive*. In addition, we call for future studies to replicate our results in larger samples. Further, training performance served as criterion among a relatively small sample of hired trainees. So, future studies should replicate our results with job performance as criterion.

Second, our study dealt with a multimedia test with primarily interpersonal situations. Future research should examine whether our results generalize to other tests and job situations (e.g., decision-making situations). As emotions, facial expressions, voice inclinations, and posture play a predominant role in interpersonal interactions, this might have increased discrepancies between the written versus behavioral response mode scores. Behavioral response mode scores may result in higher criterion-related validity for jobs that have a strong interpersonal component, whereas this might not be the case for jobs where performance is usually written.

Conclusion

How candidates are required to respond to tests is an essential component of selection procedures. Yet, there is little understanding of this building block of selection procedures. This study contrasted a behavioral response mode to a written constructed one. The results are *suggestive* of the behavioral response mode being more valid for predicting police trainee performance one year later, less cognitively saturated, more correlated with extraversion, and higher on media richness than a written constructed response mode. On a broader level, these encouraging results call for more studies that examine the isolated impact of method factors on the substantive relationships of test performance and validity.

References

Anderson, N., Lievens, F., Van Dam, K., & Born, M. P. 2006. A construct-driven investigation of gender differences in a leadership-role assessment center. *Journal of Applied Psychology,* 91: 555-566.

Archer, D., & Akert, R. M. 1980. The encoding of meaning: A test of three theories of social interaction. *Sociological Inquiry,* 50: 393-419.

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

Arthur, W., Jr., 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., Jr., & 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.

Barrick, M. R., & Mount, M. K. 1991. The big 5 personality dimensions and job-performance: A meta-analysis. *Personnel Psychology,* 44: 1-26.

Bentler, P. M. 1995. *EQS, structural equations program manual, program version 5.0*. Encino, CA: Multivariate Software.

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. 2004. Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology,* 86: 599-614.

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

Burnett, J. R., Fan, C., Motowidlo, S. J., & DeGroot, T. 1998. Interview notes and validity. *Personnel Psychology,* 51: 375-396.

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

CEBIR 2009. *Testgids Maart 2009, versie 1.6* [Test Guide March 2009, version 1.6]. Retrieved from http://www.cebir.be/Cebir.php

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.

Christian, M. S., Edwards, B. D., & Bradley, J. C. 2010. Situational judgement tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology,* 63: 83-117.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Daft, R. L., & Lengel, R. H. 1984. Information richness: A new approach to managerial behavior and organization design. *Research in Organizational Behavior,* 6: 191-233.

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.

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.

DeGroot, T., & Motowidlo, S. J. 1999. Why visual and vocal interview cues can affect interviewers' judgments and predict job performance. *Journal of Applied Psychology,* 84: 986-993.

Edwards, B. D., & Arthur, W., Jr. 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.

Feldt, L. S. 1980. A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika,* 45: 99-105.

Fulk, J., & Boyd, B. 1991. Emerging theories of communication in organizations. *Journal of Management,* 17: 407-446.

Funke, U., & Schuler, H. 1998. Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment,* 6: 115-123.

Gesn, P. R., & Ickes, W. 1999. The development of meaning contexts for emphatic accuracy: Channel and sequence effects*.* *Journal of Personality and Social Psychology,* 77: 746-761.

Goffin, R. D., Gellatly, I. R., Paunonen, S. V., Jackson, D. N., & Meyer, J. P. 1996. Criterion validation of two approaches to performance appraisal: The Behavioral Observation Scale and the relative percentile method. *Journal of Business and Psychology,* 11: 23-33.

Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnston, N. G. 2009. Taking advantage of social comparisons in performance appraisal: The relative percentile method*.* *Human Resource Management,* 48: 251-268.

Goldberg, L. R. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology,* 59: 1216-1229.

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, I. L., Zedeck, S., & Schneider, B. 1993. An exploration of the job analysis-content validity process. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations:* 2-34. San Francisco, CA: Jossey-Bass.

Gowing, M. K., Morris, D. M., Adler, S., & Gold, M. 2008. The next generation of leadership assessments: Some case studies. *Public Personnel Management,* 37: 435-455.

Heckman, J. J. 1976. The common structure of statistical models of truncated, sample selection and limited dependent variables, and a simple estimator of such models. *Annals of Economic and Social Measurement,* 5: 475-492.

Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica,* 47: 153-161.

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

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

Huffcutt, A. I., Weekley, J. A., Wiesner, W. H., DeGroot, T. G., & Jones, C. 2001. Comparison of situational and behavior description interview questions for higher-level positions. *Personnel Psychology,* 54: 619-644.

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. 2006. From the subjects' point of view—Reactions to different types of situational judgment items. *European Journal of Psychological Assessment,* 22: 168-176.

Klinkenberg, E. L., & Van Leeuwen, A. E. 2003. *Voortgangsverslag ontwikkeling M5Q- IWSP* [Progress Report Development M5Q- IWSP]. Culemborg, the Netherlands: Meurs Personeelsadvies.

Lievens, F., & Patterson, F. 2011. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced level high-stakes selection. *Journal of Applied Psychology,* 96: 927-940.

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., & Sackett, P. 2012. The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology,* 97: 460-468.

Lubinski, D., & Dawis, R. V. 1992. Aptitudes, skills, and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology*, vol. 3: 1-59. Palo Alto, CA: Consulting Psychologists Press.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. 2007. Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology,* 60: 63-91.

McFarland, L. A., Yun, G. J., Harold, C. M., Viera, L., & Moore, L. G. 2005. An examination of impression management use and effectiveness across assessment center exercises: The role of competency demands. *Personnel Psychology,* 58: 949-980.

Motowidlo, S. J., & Beier, M. E. 2010. Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology,* 95: 321-333.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. 2006. Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology,* 91: 749-761.

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.

Ployhart, R. E. 2006. The predictor response model. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement and application:* 83-105. Mahwah, NJ: Lawrence Erlbaum.

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

Ployhart, R. E., Schneider, B., & Schmitt, N. 2006. *Staffing organizations: Contemporary practice and research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Potosky, D. 2008. A conceptual framework for the role of the administration medium in the personnel assessment process. *Academy of Management Review,* 33: 629-648.

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. 1994. Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology,* 79: 298-301.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. 2000. Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology,* 85: 880-887.

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.

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

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

Sackett, P. R. 1987. Assessment-centers and content validity: Some neglected issues. *Personnel Psychology*, 40: 13-25.

Sackett, P. R., & Yang, H. 2000. Correction for range restriction: An expanded typology. *Journal of Applied Psychology,* 85: 112-118.

Schmidt, F. L., & Hunter, J. E. 1998. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin,* 124: 262-274.

Schmidt, F. L., Oh, I. S., & Le, H. 2006. Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology,* 59: 281-305.

Schmitt, N. 1994. Method bias: The importance of theory and measurement. *Journal of Organizational Behavior,* 15: 393-398.

Schmitt, N., & Ostroff, C. 1986. Operationalizing the "behavioral consistency" approach: Selection test development based on a content-oriented strategy. *Personnel Psychology,* 39: 91-108.

Sitkin, S. B., Sutcliffe, K. M., & Barrios-Choplin, J. R. 1992. A dual-capacity model of communication media choice in organizations. *Human Communication Research,* 18: 563-598.

Thornton, G. C., & Rupp, D. E. 2006. *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum.

Thornton, G. C., III, & Cleveland, J. N. 1990. Developing managerial talent through simulation. *American Psychologist,* 45: 190-199.

Trevino, L. K., Lengel, R. H., & Daft, R. L. 1987. Media symbolism, media richness, and media choice in organizations: A symbolic interactionist perspective. *Communication Research,* 14: 553-574.

Vandenberg, R. J., & Lance, C. E. 2000. A review and synthesis of the measurement equivalence literature: Suggestions, practices, and recommendations for organisational research. *Organizational Research Methods,* 3: 4-70.

Van Iddekinge, C. H., Raymark, P. H., & Roth, P. L. 2005. Assessing personality with a structured employment interview: Construct-related validity and susceptibility to response inflation. *Journal of Applied Psychology,* 90: 536-552.

Van Leeuwen, A. E. 2000. *Constructie van de M5Q voor IWSP* [Construction of the M5Q for IWSP]. Culemborg, the Netherlands: Meurs Personeelsadvies.

Webster, J., & Trevino, L. K. 1995. Rational and social theories as complementary explanations of communication media choices: Two policy-capturing studies. *Academy of Management Journal,* 38: 1544-1572.

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

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. 2006. On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests:* 157-182. San Francisco, CA: Jossey-Bass.

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.

Winship, C., & Mare, R. D. 1992. Models for sample selection bias. *Annual Review of Sociology,* 18: 327-350.