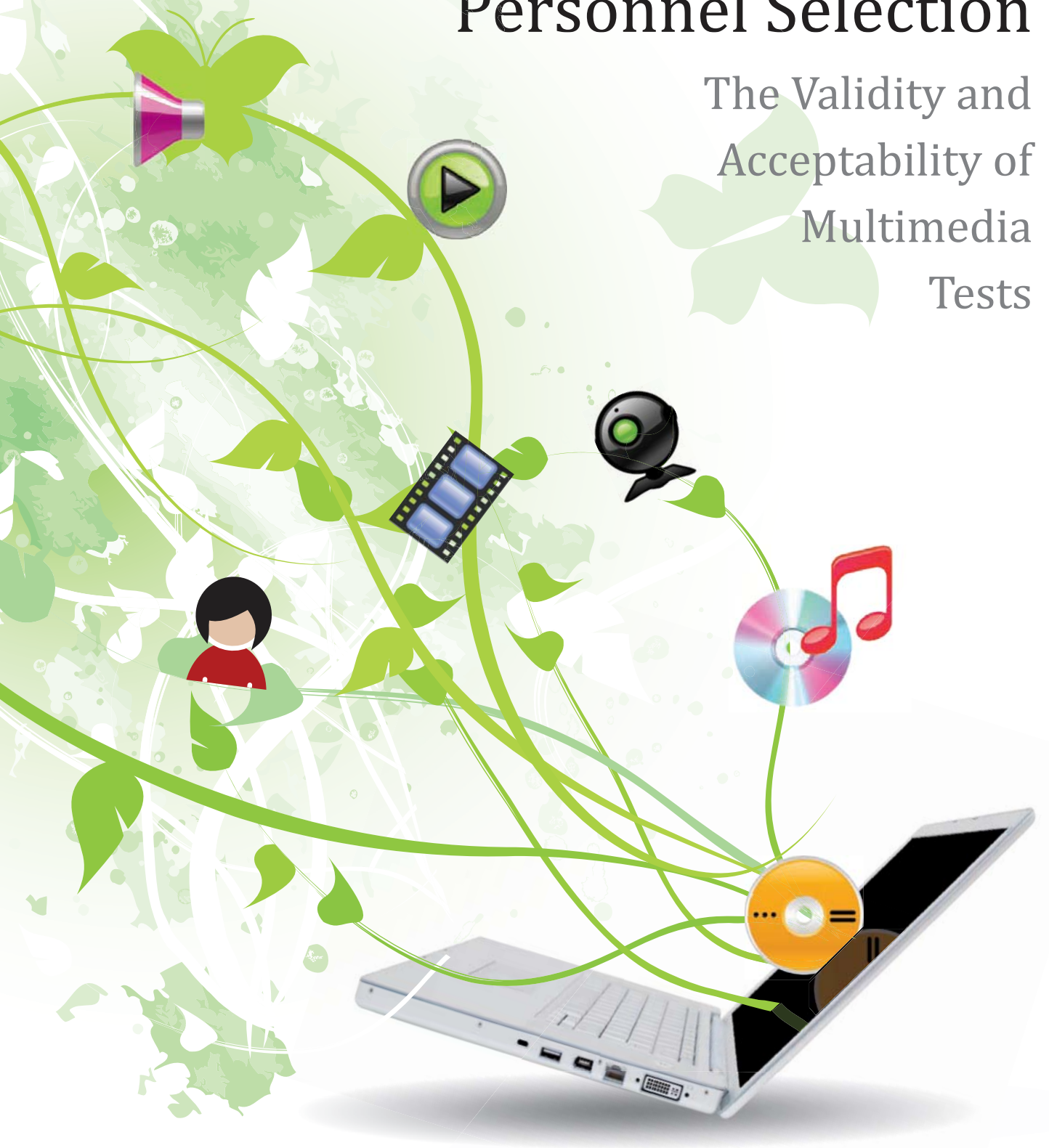
New Technology in Personnel Selection

The Validity and Acceptability of Multimedia Tests



Janneke K. Oostrom

New Technology in Personnel Selection:

The Validity and Acceptability of Multimedia Tests

Janneke K. Oostrom

The research presented in this dissertation was supported in part by funding from GITP International BV. The opinions expressed by authors are their own and do not necessarily reflect the views of GITP International BV.

ISBN: 978-90-76269-85-6

Cover designed by Joost P. L. Modderman
Lay out by Janneke K. Oostrom
Printed by Ipskamp Drukkers

New Technology in Personnel Selection:
The Validity and Acceptability of Multimedia Tests

Nieuwe technologie in personeelsselectie:
de validiteit en acceptatie van multimedia tests

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam

op gezag van de
rector magnificus

Prof.dr. H.G. Schmidt

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
vrijdag 1 oktober 2010 om 11.30 uur

door

Janneke Karina Oostrom
geboren te Krimpen aan den IJssel

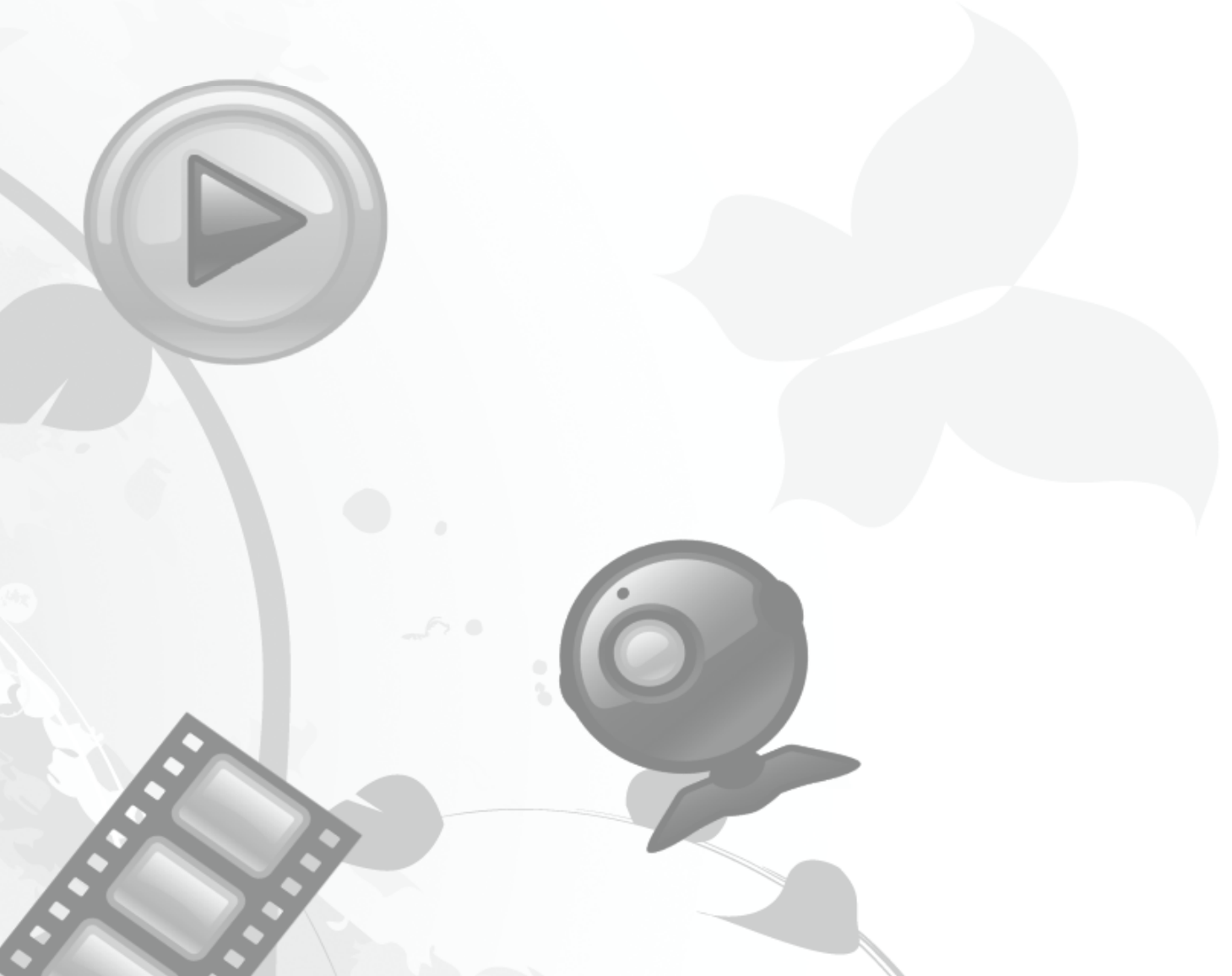
Promotiecommissie

Promotoren:	Prof.dr. M.Ph. Born
	Prof.dr. H.T. van der Molen
Overige leden:	Prof.dr. A.B. Bakker
	Prof.dr. F. Lievens
	Prof.dr. H. van der Flier
Copromotor:	Dr. A.W. Serlie

Contents

Chapter 1

General introduction



The advances in technology of the last fifty years, specifically the advent of the computer, its continuous improvements in functionality and capacity, and the growth of the internet, have affected almost every aspect of psychological testing in personnel selection practices. Since the 1960s, traditional psychological tests with paper-and-pencil formats are already being converted to computerized formats (Bartram, 1994). Yet, new technology provides more possibilities than simply changing the test medium. For instance, it also provides the opportunity to dynamically select the items to be presented and to use a variety of stimulus materials (Olson-Buchanan & Drasgow, 1999). Recently, researchers and practitioners are using new technology for the delivery of so-called multimedia tests, which include audio and video fragments (Lievens, Van Dam, & Anderson, 2002).

The present dissertation presents five empirical studies on multimedia tests and is aimed to address both theoretical and practical questions concerning their validity and acceptability. In this introductory chapter, first, a short overview of the history of computerized testing is given. Second, past research regarding multimedia testing within the domain of personnel selection is discussed. Finally, the research aims of the following five chapters of this dissertation are presented.

History of Computerized Testing

Already in the 1960s, some visionary test developers realized that psychological tests could be efficiently and adequately administered via computers (Bartram & Bayliss, 1984). During the late 1960s some of the earliest systems were designed to automate the scoring procedures of psychological tests (Bartram, 2006).

In the 1970s, the first computer adaptive tests were launched (e.g., Brown & Weiss, 1977; Kreitzberg, Stocking, & Swanson, 1978; Weiss, 1973). These tests were administered via the computer and made use of item response theory to administer test items that matches applicants' ability level as determined by their performance on previous items. However, early versions of computer adaptive tests were subject to a number of constraints, such as high initial hardware and developmental software costs (Kreitzberg et al., 1978).

In the 1980s, the first personal computers were introduced, which marks the beginning of current approaches to psychological testing (Sands, Waters, & McBride, 1997). Since the introduction of personal computers, test developers started to develop computerized versions of paper-and-pencil tests. However, these early adaptations from paper-and-pencil tests to computerized tests were merely a change in test medium. In order to take advantage of the potential of the computer for test delivery, test developers tried to create an enhanced value through computerized testing (McBride, 1997). Thus, the aim no longer was to simply transfer paper-and-pencil tests to electronic page-turner versions, but to create so-called innovative computerized tests. However, the possibilities for the development of innovative computerized tests were restricted, as computers at that time were rather expensive, and the storage, software, and multimedia capabilities still were limited.

In the 1990s, computers finally became equipped with the necessary capabilities for the development of innovative computerized tests (Drasgow & Mattern, 2006). Graphical user interfaces, large memory capacities, sound and video cards, and the beginning of the widespread use of the internet opened the door for the development of various innovative computerized tests. Thus, technology was able to provide test developers with diverse opportunities to improve psychological testing, for example by including multimedia (i.e. video clips), by modifying the format and the scoring of the test (i.e. adaptive testing), by altering the way applicants respond to the test items (i.e. multiple choice or constructed responses), or by developing tests that measure individual differences that were difficult or impossible to measure with paper-and-pencil tests, such as communication skills, teamwork, leadership, critical thinking, and creativity (McHenry & Schmitt, 1994). Current publications have provided a range of examples of innovative computerized tests that vary on one or more of these dimensions, such as computer-based realistic job previews (e.g., Highhouse, Stanton, & Reeve, 2004), multimedia situational tests (e.g., Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000), computerized in-basket exercises (e.g., Wiechmann & Ryan, 2003), and virtual reality tests (e.g., Aguinas, Henle, & Beaty Jr, 2001). The present dissertation focuses on computerized tests which use multimedia, so-called multimedia tests.

Research on Multimedia Tests

In a multimedia test, applicants are usually presented with a variety of challenging job-related situations. The situation then freezes at an important moment and applicants are asked to evaluate a number of courses of action and indicate how they would act in this particular situation (Weekley & Ployhart, 2006). This type of multimedia test is called a multimedia or video-based situational judgment test (SJT). Recently, another innovative multimedia test has entered personnel selection practices, namely a webcam test. A webcam test can be conceptualized as a multimedia SJT with a constructed-response item format. In a webcam test, applicants are presented with situations through the use of video clips and are then asked to act out their response, while being filmed by a webcam (Lievens, Peeters, & Schollaert, 2008).

Research regarding the psychometric properties of multimedia tests is until this moment scarce. The present dissertation aims to fill this void by presenting five empirical studies on multimedia tests. In the following paragraphs an overview of prior research on both multimedia SJTs and webcam tests will be provided with respect to six criteria against which personnel selection tests need to be assessed, namely validity, reliability, adverse impact, acceptability, cost effectiveness, and ease of use (Cook, 2009). The present dissertation will address the most central of these criteria, namely validity and acceptability.

Validity

A valid psychological test is one that measures what it claims to measure and that predicts something useful (Cook, 2009). There are several types of validity that are

relevant in selection contexts. Regarding multimedia tests, the most relevant types of validity are criterion-related validity, incremental validity, and construct validity. Research findings concerning these three types of validity are discussed in the following paragraphs.

Criterion-related validity. This type of validity refers to the degree to which a test estimates an important behavioral criterion, external to the test itself (Nunnally, 1978). Within the personnel selection domain the most important criteria are job performance and to a lesser extent academic performance. Various studies have confirmed the validity of multimedia SJTs in both the prediction of job and academic performance (Lievens & Sackett, 2006; Weekley & Ployhart, 2005). In a meta-analysis, Salgado and Lado (2000) demonstrated that multimedia SJTs are good predictors of job performance, with an average observed validity of .25 and an average corrected validity of .49.

Research regarding the criterion-related validity of multimedia SJTs with constructed-response item formats still is very scarce. Because the manner of responding in multimedia SJTs with constructed-response item formats (that means that respondents have to act out their response) more closely resembles actual work conditions than does the manner of responding in a multimedia SJT with multiple choice formats, multimedia SJTs with constructed-response item formats were expected to be better indicators of future performance (Motowidlo, Dunnette, & Carter, 1990). Indeed, Funke and Schuler (1998) found that the manner of responding moderates the criterion-related validity of SJTs, leading to the highest criterion-related validity for tests with orally-given free responses. However, an important drawback in the study of Funke and Schuler is their criterion measure, namely role play behavior. Performance in a role play exercise is inherently different from job performance. For this reason, the present dissertation will examine the criterion-related validity of multimedia SJTs that have a constructed-response item format, with actual job and academic performance as the criteria to predict.

Incremental validity. This form of criterion-related validity refers to whether a selection test adds to the prediction of a criterion above what is predicted by other selection tests (Hunsley & Meyer, 2003). Personnel selection procedures typically include measures of cognitive ability and personality (F. L. Schmidt & Hunter, 1998). The value of multimedia tests would therefore increase if they showed incremental validity over these traditional predictors. In a meta-analysis, Salgado and Lado (2000) demonstrated that multimedia SJTs add substantial validity over cognitive ability measures (ΔR^2 = .10). Similarly, Lievens and Sackett (2006) reported that a multimedia SJT which aimed to measure interpersonal and communication skills had incremental validity over and above a cognitive composite measure and a work sample in the prediction of students' grades on interpersonally oriented courses (ΔR^2 = .11).

Despite the fact that several authors have welcomed research regarding the incremental validity of multimedia SJTs with constructed-response item formats (e.g.,

Lievens et al., 2008; Olson-Buchanan & Drasgow, 2006), until now to our knowledge no studies have examined the incremental validity of this type of multimedia test. For that reason, the present dissertation will be the first to investigate the incremental validity of multimedia SJTs with constructed-response item formats over and above several traditional personnel selection tests.

Construct validity. This type of validity refers to the extent to which a psychological test relates to other measures based on theoretically derived hypotheses (Carmines & Zeller, 1979). The construct validity of SJTs remains hard to pin down. According to Stemler and Sternberg (2006) SJTs measure practical intelligence, which is the ability to adapt to, shape, and select everyday environments. Other researchers argue that situational judgment tests reflect a number of constructs that are related to job performance, such as cognitive ability, personality, and job experience (Weekley & Jones, 1999). Along these lines, meta-analyses have shown that SJTs have an average observed correlation of .31 with cognitive ability (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Of the personality dimensions, emotional stability has been found to have the highest observed correlation (r = .31) with SJT performance, followed by conscientiousness (r = .26) and agreeableness (r = .25; McDaniel & Nguyen, 2001). However, almost all construct validity evidence until now has been restricted to paper-and-pencil SJTs. The test medium and the response format of an SJT is expected to affect the construct validity (McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006). For example, multimedia SJTs are expected to reduce the cognitive load of an SJT primarily by reducing the reading demands. Chan and Schmitt (1997) demonstrated that reading comprehension indeed is uncorrelated with test performance on a multimedia SJT. The construct validity of multimedia tests evidently needs further examination. Therefore, the present dissertation will address several questions regarding the construct validity of multimedia tests.

Reliability

Reliability refers to the degree to which a measure of individuals differences, such as a psychological test, produces the same or similar results on different occasions, or by different observers, or by similar parallel tests (Streiner & Norman, 1995). Regarding multimedia SJTs, the most widely used measure of reliability is the internal consistency as indexed by coefficient alpha. However, estimating the internal consistency of SJTs is often problematic and not very relevant, because most SJTs assess multiple constructs, such as personality dimensions and cognitive ability (McDaniel & Whetzel, 2005). For this reason, multimedia SJTs typically present low internal consistencies. For example, Chan and Schmitt (1997) report an internal consistency coefficient of .55 for a multimedia SJT developed to assess a series of for work habits and interpersonal skills. However, in construct-driven multimedia SJTs the internal consistency is a relevant form of reliability, as these types of multimedia tests are developed to measure one specific construct. Studies on construct-driven multimedia SJTs report adequate levels of internal consistency. For example, De Meijer, Born, Van

Zielst, and Van der Molen (in press) report an internal consistency of .69 for a multimedia SJT developed to measure the integrity of police officers. Regarding multimedia SJTs with constructed-response item formats, such as webcam tests, good internal consistencies have been reported. Stricker (1982), for example, reported internal consistencies between .74 and .82 for a multimedia SJT in which participants had to write down their responses.

Besides internal consistency, the inter-rater reliability is also an important index of the reliability of multimedia SJTs with constructed-response item formats. In general, good inter-rater reliabilities have been reported for multimedia SJTs with constructed-response item formats. For example, regarding a multimedia SJT with orally-given free responses, Funke and Schuler (1998) found an average correlation of .79 between three assessors. Walker and Goldenburg (2004, as cited in Olson-Buchanan & Drasgow, 2006) reported inter-rater reliabilities ranging from .67 to .78 for a multimedia SJT with a constructed-response item format designed for the selection of border patrol officers. Since reliability is a prerequisite for validity, the present dissertation will indirectly address the reliability of multimedia tests.

Adverse impact

Adverse impact refers to a substantially different rate of selection in hiring, which works to the disadvantage of a minority group (Ironson, Guion, & Ostrander, 1982). Although cognitive ability tests have been found to be the best predictor of job performance (e.g., F. L. Schmidt & Hunter, 1998), these tests also have been found to produce the highest adverse impact with respect to ethnic minority groups (e.g., Hunter & Hunter, 1984). As many organizations believe it is important for business and ethical reasons to create a diverse workforce, researchers are searching for valid predictors of job performance that result in less adverse impact than cognitive ability tests. Using multimedia tests instead of paper-and-pencil tests has been suggested as one of the strategies to reduce adverse impact. The use of multimedia reduces the reading demands, and subsequently the cognitive load of the test (Ployhart & Holtz, 2008). Chan and Schmitt (1997) demonstrated that a multimedia SJT indeed resulted in less adverse impact compared to a paper-and-pencil version of the same SJT.

As far as we know, the adverse impact of multimedia SJTs with constructed-response item formats has not yet been examined. SJTs with a multiple-choice format have been suggested to measure participants' knowledge of what should be done in the particular job-related situation (Motowidlo, Brownlee, & Schmit, 2008). Both knowledge and cognitive ability are cognitive constructs. In contrast, SJTs with constructed-response item formats have been suggested to measure applicants' actual skills, because participants have to create and enact their own answer (Funke & Schuler, 1998). For this reason, using a constructed-response item format may even further reduce the cognitive loading of an SJT. Further research is needed to investigate whether the use of a constructed-response item format indeed affects the cognitive loading of an SJT and subsequently results in less adverse impact. Although this question is worth studying, it is not addressed in the present dissertation.

Acceptability

In order to realize the benefits of the use of computerized tests in selection contexts, such tests have to be acceptable to applicants. Measuring how applicants react to selection tests has been found to be relevant for applicants themselves and for organizations. Previous studies have demonstrated that applicant reactions are related to intentions to accept the job, the likelihood of litigation against the outcome of the selection procedure, and perceived organizational attractiveness (Anderson, Lievens, Van Dam, & Ryan, 2004; Chan & Schmitt, 2005; Gilliland, 1993; Ryan & Ployhart, 2000). Multimedia tests provide a realistic job preview to the applicant and therefore are expected to be more attractive for applicants in terms of interest and motivation than traditional paper-and-pencil tests (Stricker, 1982). Richman-Hirsch, Olson-Buchanan, and Drasgow (2000) demonstrated that compared to a paper-and-pencil test, a multimedia version of the same test indeed yielded more positive reactions. The multimedia assessment was perceived as more content valid, more face valid, more enjoyable and led to more satisfaction with the assessment process. Chan and Schmitt (1997) demonstrated that participants rate the face validity of a multimedia SJT significantly more positive than the face validity of a paper-and-pencil SJT.

Much of the research on applicant reactions to computerized selection instruments has been rather descriptive and comparative, rather than explanatory (e.g., Kanning, Grewe, Hollenberg, & Hadouch, 2006; Reynolds, Sinar, & McClough, 2000; Richman-Hirsch et al., 2000). Theory is lacking on why applicants evaluate different selection instruments in a different manner (Anderson, 2003). For example, differences in test anxiety, computer anxiety or openness to experience are likely to influence applicant reactions to multimedia tests, yet have only been included in a few studies (e.g., Wiechmann & Ryan, 2003). The present dissertation aims to shed light on the nature of applicant reactions by examining relationships between several testing-related individual differences and applicant reactions to compurized selection tests.

Cost-effectiveness

A disadvantage of multimedia tests may involve the production costs. Scripts must be written for the scenarios, the scenarios need to be filmed by a professional crew and with professional actors, and the recordings need to be edited. Dalessio (1994) estimated the costs of multimedia SJTs from $2000 to $3000 per minute of filming. In addition, the administration costs of multimedia SJTs are higher because technological investments have to be made for administering multimedia SJTs. However, the cost effectiveness of any selection test is not only determined by the development and administration costs involved, but also by its criterion-related validity (Cronbach & Gleser, 1965). As described above, there is meta-analytic evidence that supports the criterion-related validity of multimedia SJTs (Salgado & Lado, 2000).

In multimedia tests with constructed-response item format, applicants' responses are rated afterwards. This scoring method is also quite costly. However, as an important advantage of computer technology lies in the automatic scoring, perhaps in

the future the costs of scoring could be reduced by using voice-recognition software or other automatic scoring possibilities (Powers, Burstein, Chodorow, Fowles, & Kukich, 2002). As the cost effectiveness of a selection test is partly determined by its criterion-related validity, the present dissertation will indirectly address the cost effectiveness of multimedia tests.

Ease of use

Ease of use refers to how conveniently the test fits into an organization's selection system (Cook, 2009). Larger organizations apparently believe multimedia tests to fit in conveniently, as they have rushed to incorporate new technology, including multimedia tests, into their selection systems (Anderson, 2003). In general, test administrators are more comfortable with giving applicants selection tests that are perceived as job-related (Shotland, Alliger, & Sales, 1998). Therefore, test administrators should also feel comfortable giving multimedia tests, since these tests have been found to be perceived as more job-related than their paper-and-pencil counterparts (e.g., Chan & Schmitt, 1997). Shotland et al. (1998) demonstrated that managers were impressed with the ease of administration and scoring of a multimedia test, and appreciated the fact that limited involvement was required on their part. An important advantage of multimedia tests over alternative selection instruments such as role play exercises is that multimedia tests can be administered over the internet, which allows to test large groups of applicants at once and on various locations (Lievens et al., 2008). Because the ease of use of multimedia test has already been confirmed in the literature, this is not addressed in the present dissertation.

Specific Research Aims

As described in the previous paragraph, important questions regarding the validity and acceptability of multimedia tests still pertain. This dissertation presents five empirical studies on the criterion-related validity, incremental validity, construct validity, and acceptability of multimedia tests. Chapters 2, 3, and 4 address the criterion-related validity and the incremental validity of multimedia tests. Chapters 2 and 4 address the construct validity of multimedia tests. Chapters 5 and 6 address the acceptability of multimedia tests. An overview of the specific research purpose of each chapter is presented below.

In **chapter 2** the criterion-related validity and the construct validity of a webcam test aimed to measure interpersonally oriented leadership skills are examined. In particular its relationship with personality, cognitive ability, previous job experience, and academic performance is examined in a sample of psychology students. In addition, the incremental validity of the webcam test over and above a cognitive ability test and a personality questionnaire in the prediction of academic performance is investigated.

In **chapter 3** the criterion-related validity of a webcam test aimed to measure the effectiveness in the core task of employment consultants, that is advising job seekers, is investigated. Furthermore, it will be examined whether the webcam test is able to

explain unique variance in job performance over and above a job knowledge test. The study is conducted among consultants of an employment agency who participated in a certification process.

In **chapter 4** we sought to extend Motowidlo, Hooper, and Jackson's (2006b) work on the implicit trait policy (ITP) theory. ITP theory assumes that there are stable differences in individuals' implicit beliefs about the effectiveness of different levels of trait expression which affect judgments of the effectiveness of SJT response options (Motowidlo, Hooper, & Jackson, 2006a). In a sample of assessment candidates it is examined whether a multimedia SJT for leadership skills is able to measure implicit trait policies for targeted traits and whether these implicit trait policies are able to predict leadership behaviors.

In **chapter 5** the relationship of several testing-related and general individual differences with students' perceived job relatedness of a computerized cognitive ability test and a multimedia SJT are examined. Previous studies have shown that test content and test characteristics affect the perceived job relatedness of selection instruments (e.g., Chan & Schmitt, 1997), but there is still substantial variance in these perceptions that remains unexplained. This chapter examines whether anxiety (test anxiety and computer anxiety), self-evaluations (test-taking self-efficacy, core self-evaluations, and subjective well-being), and personality (agreeableness, emotional stability, and openness to experience) are able to explain variance in job relatedness perceptions.

In **chapter 6** the nature of applicant reactions and their relationship with test performance are examined by drawing upon the applicant reaction model of Chan, Schmitt, Sacco, and DeShon (1998). Furthermore, in this chapter applicants' pretest and posttest face validity perceptions, predictive validity perceptions, and fairness perceptions regarding a paper-and-pencil version and a computerized version of an in-basket exercise are compared. A sample of job applicants is used.

Finally, in **chapter 7** the findings of the different chapters are summarized and important theoretical and practical implications are discussed. Furthermore, in this chapter the limitations of the studies presented are discussed and suggestions for future research are made.

Chapter 2

A multimedia situational judgment test with a constructed-response item format: Its relationship with personality, cognitive ability, job experience, and academic performance*

* This chapter is submitted for publication as:
Oostrom, J. K., Born, M. Ph., Serlie, A. W., & Van der Molen, H. T. (submitted). A multimedia situational judgment test with a constructed-response item format: Its relationship with personality, cognitive ability, job experience, and academic performance.
The study in this chapter was also presented at the 6th conference of the International Test Commission (ITC), Liverpool, UK, July 2008.

Abstract

Advances in computer technology have created opportunities for the development of a multimedia situational judgment tests in which responses are filmed with a webcam. This paper examined the relationship of a so-called webcam test with personality, cognitive ability, leadership experience, and academic performance. Data were collected among 153 psychology students. In line with our expectations, scores on a webcam test, intended to measure interpersonally oriented leadership, were related to extraversion, conscientiousness, and emotional stability, but not to cognitive ability. Furthermore, the webcam tests significantly predicted students' learning activities during group meetings over and above a cognitive ability test and a personality questionnaire. Overall, this study demonstrates that webcam tests can be a valid complement to traditional predictors in selection contexts.

Introduction

Over the past decade situational judgment tests (SJTs) have become increasingly popular in research and in practice (McDaniel, Hartman, Whetzel, & Grubb, 2007). In the typical SJT, applicants are presented with a variety of job-related situations and a number of plausible courses of action. The applicants are then asked to evaluate each course of action and rate their effectiveness. SJTs may be used to assess different constructs, both cognitive and non-cognitive (Arthur & Villado, 2008; Chan & Schmitt, 2005). However, most SJTs have been developed to measure interpersonally oriented constructs, such as leadership skills (Salgado & Lado, 2000). Many studies have demonstrated that paper-and-pencil SJTs can be valid predictors of job and academic performance (e.g., Lievens & Sackett, 2006; McDaniel et al., 2007).

Recent advances in computer technology have created opportunities for the development of so-called multimedia or video-based SJTs in which the situations and courses of actions are presented by using video clips (Olson-Buchanan & Drasgow, 2006). Presenting situations in video format instead of written format might enhance the correspondence to the criterion, leading to higher criterion-related validity. In a recent study, Lievens and Sackett (2006) demonstrated that for an interpersonally oriented multimedia SJT, it indeed has a higher criterion-related validity than its written counterpart in predicting students' performance on interpersonally oriented courses. Recently, the use of advanced multimedia computer technology in SJTs has even gone a step further by showing applicants situations through video clips and then asking them to act our their response, which is in turn filmed with a webcam (Lievens et al., 2008). Until now, there has been relatively little research on multimedia SJTs with such constructed-response item formats. This paper will address this shortcoming by examining the construct validity and criterion-related validity of a so-called webcam test. In particular its relationship with personality, cognitive ability, previous job experience, and academic performance is examined. We will first discuss the literature on SJTs and then will propose hypotheses about the construct validity and criterion-related validity of the webcam test.

Situational judgment tests

Situational judgment testing is a measurement method designed to sample behaviors that are assumed to be necessary for job or academic performance (Motowidlo et al., 1990). Samples or simulations are based on the notion of behavioral consistency (Schmitt & Ostroff, 1986). By eliciting a sample of current behaviors, one can predict how someone will behave in the future (Wernimont & Campbell, 1968). Simulations vary in the fidelity with which they present a stimulus and elicit a response (Motowidlo et al., 1990). The highest fidelity simulations use realistic stimuli to present a job-related situation and provide candidates with the opportunity to respond as if they were actually in the job situation. Fidelity decreases if the stimuli have less correspondence with actual work conditions. A paper-and-pencil SJT is an example of a low fidelity simulation, as it presents a verbal description of a job-related situation

and asks candidates to indicate the effectiveness of a number of plausible courses of action.

Although the literature has shown that SJTs can be valid predictors of job and academic performance (e.g., Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; McDaniel et al., 2007), the construct validity of SJTs remains hard to pin down. There is an ongoing discussion among researchers why SJTs predict performance (Lievens et al., 2008). According to Stemler and Sternberg (2006) SJTs measure practical intelligence, which is the ability to adapt to, shape, and select everyday environments. It requires knowledge, both tacit and explicit, about how to deal effectively with situations that occur in the context of everyday experiences. Other researchers argue that SJTs are valid predictors because they reflect constructs that are themselves related to job or academic performance, such as personality and cognitive ability. Meta-analyses have shown that SJTs have an average observed correlation of .31 with cognitive ability (McDaniel et al., 2001). Of the personality dimensions, emotional stability has been found to have the highest observed correlation (r = .31) with SJT performance, followed by conscientiousness (r = .26) and agreeableness (r = .25; McDaniel & Nguyen, 2001). However, the correlations between SJT performance and Big Five personality dimensions vary widely, depending on which behavioral domain is being assessed with the SJT (Chan & Schmitt, 2005).

Multimedia situational judgment tests

Recent advances in multimedia technology have opened the door for an SJT format in which situations are presented through the use of video clips. By utilizing video, it is possible to portray detailed and accurate job-related scenarios, which increases the fidelity of the presented situations (Weekley & Jones, 1997). Studies on the criterion-related validity of multimedia SJTs still are scarce, but the few studies that have been conducted support their predictive validity (Olson-Buchanan et al., 1998; Weekley & Jones, 1997). In a meta-analysis Salgado and Lado (2000) found that multimedia SJTs are good predictors of job performance, with an average observed validity of .25, and add substantial validity over cognitive ability measures (ΔR^2 = .10). There is also evidence that multimedia SJTs are valid predictors of academic performance. For example, Lievens, Buyse, and Sackett (2005) examined the criterion-related validity of a video-based SJT, developed to measure interpersonal and communication skills, for making college admission decisions. They found that the video-based SJT showed incremental validity over cognitively oriented measures for curricula that included interpersonal courses, but not for other curricula. This study demonstrates the importance of differentiating not only among predictor constructs, but also among criterion domains.

Recently, another innovative multimedia SJT has entered applicant selection practices, namely a webcam test. In this multimedia SJT candidates are presented with situations through the use of video clips and are then asked to act out their response, while being filmed by a webcam (Lievens et al., 2008). In line with Arthur and Villado (2008), we view a webcam test as a multimedia SJT with a constructed-response item

format. In this type of multimedia SJT the response format has relatively high fidelity compared to a multimedia SJT with a multiple-choice item format (Funke & Schuler, 1998), as candidates have to create and enact their own answer immediately after the stimulus is presented.

Research on multimedia SJTs with constructed-response formats is scarce (Funke & Schuler, 1998), but promising. We only found two studies on the criterion-related validity of multimedia SJTs with constructed-response item formats within the domain of personnel selection. Funke and Schuler (1998) compared the criterion-related validity of various types of SJTs which were intended to measure social skills. The SJTs systematically differed in the fidelity of the presented situation (either orally or via video) and the fidelity of the responses (multiple-choice, written free, or oral free). Funke and Schuler (1998) found that response fidelity instead of stimulus fidelity moderated the criterion-related validity of situational tests, leading to the highest criterion-related validity for the video test with orally-given responses. Oostrom, Born, Serlie, and Van der Molen (in press) conducted the first field study on a webcam test which was intended to measure effectiveness in the core task of an employment consultant, namely advising job seekers. The results showed that scores on the webcam test incrementally predicted consultants' job performance over and above a job knowledge test.

Present study

Past studies on multimedia SJTs have mainly focused on the criterion-related validity of SJTs. More recently, many authors are calling for a focus towards the processes and constructs underlying SJTs (e.g., Lievens et al., 2008). For this reason, the first aim of our study is to examine the relationships between scores on a webcam test and personality, cognitive ability, and previous job experience.

The webcam test used in this study was intended to measure interpersonally oriented leadership skills. Thus far, no studies have examined the relationships between a webcam test intended to measure interpersonally oriented leadership skills and personality, cognitive ability, and previous job experience. However, two studies have provided construct validity evidence for paper-and-pencil SJTs with a multiple-choice format that were developed to measure leadership skills. Oswald et al. (2004) developed an SJT adapted to 12 content dimensions of academic performance, including a leadership dimension. Schmitt and Chan (2006) analyzed the data of Oswald et al. to provide insight into the construct validity of the SJT. The leadership dimension of the SJT was significantly related to extraversion and conscientiousness. Depending on the scoring strategy, Bergman et al. (2006) found significant correlations in varying degrees between performance on a leadership SJT and cognitive ability, extraversion, openness to experience, and conscientiousness.

The studies of Oswald et al. (2004) and Bergman et al. (2006) provide evidence that paper-and-pencil leadership SJTs are related to the personality traits extraversion and conscientiousness and to cognitive ability. However, it may be expected that a multimedia SJT with a constructed-response format is less strongly related to

cognitive ability than a paper-and-pencil SJT with a multiple-choice format. One of the reasons is that multimedia SJTs have been shown to have a lower cognitive loading than paper-and-pencil SJTs (Lievens & Sackett, 2006), because of the reading component inherent in the latter type of test (Chan & Schmitt, 1997). A second reason is that SJTs with a multiple-choice format have been suggested to measure participants' knowledge of what should be done in the particular job-related situation (Motowidlo et al., 2008). Both knowledge and cognitive ability are cognitive constructs. In contrast, an SJT with a constructed-response format intends to measure participants' actual interpersonally oriented skills, because participants need to create and enact their own answer (Funke & Schuler, 1998). This leads to the following hypothesis:

Hypothesis 1: Scores on a webcam test for interpersonally oriented leadership skills will be more strongly related to the personality traits extraversion and conscientiousness than to cognitive ability.

Job experience has been suggested to positively influence SJT performance, because people with greater job relevant experience are more likely to have encountered the types of job-related situations presented in an SJT and have learned how to respond successfully to these types of situations (e.g., Weekley & Jones, 1999). Several studies have found empirical support for the relationship between job experience and SJTs with a variety of contents (e.g., Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; McDaniel & Nguyen, 2001; Weekley & Jones, 1999). As the webcam test used for this study is developed to measure leadership skills, we will examine the relationship between scores on the webcam test and a specific type of previous job experience, namely leadership experience. We hypothesize the following.

Hypothesis 2: Scores on a webcam test for interpersonally oriented leadership skills will be positively related to leadership experience.

The second aim of our study is to examine the criterion-related validity of a webcam test for interpersonally oriented leadership skills as predictor of grade point average (GPA) and of students' observed learning activities during group meetings. Oswald et al. (2004) re-examined the domain of academic performance and identified 12 dimensions of academic performance that deal with intellectual behaviors, interpersonal behaviors, and intrapersonal behaviors. Our criterion is similar to the leadership dimension of Oswald et al., which was defined as demonstrating skills in a group, such as motivating others and coordinating groups and tasks. Oswald et al. found significant correlations between an SJT with a multiple-choice format, developed to measure the 12 dimensions of academic performance, and self-rating measures of student performance (observed r = .53) and absenteeism (observed r = -.27). However, low correlations were found between the SJT and GPA, implying that SJTs are more predictive of interpersonally oriented criteria than cognitively oriented

criteria. Similar results have been found by Lievens et al. (2005) and Lievens and Sackett (2006). Based on these findings, we hypothesize the following:

Hypothesis 3: Scores on a webcam test for interpersonally oriented leadership skills will have higher validity for predicting students' observed learning activities than for GPA.

We will also examine the incremental validity of a multimedia SJT with a constructed-response item format over and above a cognitive ability test and a personality questionnaire. A large body of research has established measures of cognitive ability and personality as important predictors of academic success (e.g., Lounsbury, Sundstrom, Loveland, & Gibson, 2003; Poropat, 2009). However, the incremental validity of a multimedia SJT with a constructed-response item format over these traditional predictors of academic performance has not yet been examined. There is evidence that SJTs with multiple choice formats have incremental validity over and above traditional predictors, suggesting these SJTs capture a unique part of job and academic performance(e.g., McDaniel et al., 2007). For example, Lievens and Sackett (2006) found that an interpersonally oriented video-based SJT had incremental validity over cognitively oriented predictors in predicting students' scores on interpersonally oriented courses. Similarly, our final hypothesis is:

Hypothesis 4: A webcam test for interpersonally oriented leadership skills will incrementally predict students' observed learning activities over and above a cognitive ability test and a personality questionnaire.

Method

Participants and procedure

The sample consisted of 153 psychology students at a large Dutch University. As part of their educational program, students completed a personality questionnaire, a cognitive ability test, and a webcam test in random order in a proctored setting either at the University or at a HRD consultancy firm. By emphasizing the benefits of practicing with real selection instruments and by providing professional feedback reports on their test scores, the students were stimulated to perform well on the different instruments. To provide a frame of reference, the participants were told that the test battery they were about to complete is generally used in the assessment of managers or supervisors, a profession most students are familiar with. It took the students about 2 hours to complete all the selection instruments. Of the students, 101 were female (66.0%) and 52 were male (34.0%). Their age ranged from 19 to 44 (M = 22.3; SD = 3.17). Most of the students (70.1%) had work experience ranging from less than 1 year to more than 10 years.

Measures

Personality questionnaire. The personality questionnaire was based on the Five Factor Model (FFM) of personality. The FFM personality traits were measured with a 224-item personality questionnaire (Koch, 1998). Each scale consists of 23 to 47 items. Participants have to rate the items on a five-point scale ranging from 1 = *strongly disagree* to 5 = *strongly agree.* The scales of the personality questionnaire show substantial correlations (r = .48 - .72) with scales of the revised NEO-Personality Inventory that were intended to measure the same constructs (Costa & McCrae, 1992). All participants completed the questionnaire within 25 minutes. In our study, coefficient alphas were substantial: α = .92 for extraversion, α = .83 for agreeableness, α = .92 for conscientiousness, α = .88 for emotional Stability, α = .90 for openness to experience. Correlations varying from .10 - .51 were found between the scales.

Cognitive ability test. The cognitive ability test consists of three scales, namely Verbal Reasoning (VR), Number Series (NS) and Abstract Reasoning (AR). The test consists of 81 items (Van Leeuwen, 2004). Together, the three scales aim to measure general cognitive ability. The scales of the cognitive ability test show substantial correlations (r = .44 - .78) with the Dutch intelligence test series of Drenth, a commonly used measure of cognitive ability in The Netherlands (Drenth, 1965). The time limit to complete all items was 51 minutes. Coefficient alphas of the scales, based on a sample of candidates who had completed all items within the time limit, were .87 for the VR scale (N =889), .63 for the NS scale (N = 649), and .68 for the AR scale (N = 757). There were moderate correlations between the three scales (r = .24 – .41). The total amount of correctly answered items represents the participants' scores.

Leadership experience. This variable was measured with the following item: '*How many years of leadership experience do you have?*'. Participants indicated their experience on a five-point scale ranging from 1 = *no experience* to 5 = *more than 10 years.*

Webcam test. The webcam test was designed to measure interpersonally oriented leadership skills. The webcam test consists of ten short videotaped vignettes. Each vignette starts with a narrative description of the situation, followed by a fragment of a conversation between the participant and a subordinate. In this segment a professional actor, playing the subordinate, talks directly to the camera, as if speaking to the participant. After this, the frame freezes and the participant, who plays the role of a supervisor, has to respond as if it were a real situation. These responses are recorded with a webcam. The response time is limited to one minute, which is long enough to react to the situation at hand. The vignettes represented five interpersonally oriented leadership behaviors: Making decisions and solving problems, coordinating the work and activities of others, guiding, directing, and motivating others, developing and building teams, and resolving conflicts and negotiating with others. An example of a

webcam test item is as follows: "*A coworker is misbehaving: He doesn't stick to agreements, his work is below standard or not finished on time. You have talked to him about these problems before. It can no longer go on this way. You asked the coworker to come to your room (introduction)*". *Subordinate:* "*You wanted to talk to me about something?*".

The effectiveness of the responses was judged afterwards by three trained assessors, who gave their ratings independently of one another and worked on the basis of a set of comprehensive scoring instructions. The scoring instructions and the participants' recorded responses were made available via a secure internet site. The assessors rated each response on a five-point scale ranging from (--) *very ineffective* to (++) *very effective.* They had received a frame-of-reference (FOR) training consisting of 1) an introduction on the basics of rating processes and the possible rating errors that can occur, and 2) a workshop on the rating process, in which they were taught what effective and ineffective behaviors were in the specific situations of the webcam test (Bernardin & Buckley, 1981). The total duration of the training was 4 hours. After the first training, as prior practice the assessors had to evaluate the responses of three participants. These ratings were discussed during a second meeting. The second meeting took about 2 hours.

In total there were five assessors (2 female, 3 male) who had a bachelor's degree in work and organizational psychology or sociology. Their age ranged from 23 to 28 years. The inter-rater reliability of the mean of the three assessors, as indexed by a two-way random effects intra-class correlation, was .82. For each response the mean score of the three assessors was calculated. The mean scores were summed resulting in an overall score that could range from 10 to 50. In our study, coefficient alpha of the webcam test equaled .80.

Students' observed learning activities. The psychology curriculum of the participants in this study applies a problem-based learning approach. An important element of the psychology courses are group meetings in which students work on meaningful problems, under the guidance of a tutor (H. G. Schmidt & Moust, 2000). During the group meetings, students discuss problems and their possible explanations or solutions, share their findings from the literature, elaborate on knowledge acquired, and have an opportunity to correct misconceptions. During each meeting one of the students takes on the role of chair and one of the students takes on the role of scribe. At the end of each course the students' tutor fills out a 19-item questionnaire (Loyens, Rikers, & Schmidt, 2007), in which the students are evaluated on a number of criteria, namely how well they have prepared themselves for the group meetings, how active and motivated they were during the group meetings, and how well they fulfilled their roles as chair and scribe. The tutors rate each student on five-point scale-items ranging from 1 = *the student did not show this activity at all* to 5 = *the student showed this activity to a large extent.*

Based on principal component analysis three scales were extracted, which highly corresponded to the three scales described by Loyens et al. (2007). The first scale,

named *Participation*, consisted of 7 items (Eigenvalue 10.71, 29.66% of variance explained). An example of an item is: "*The student actively took part in the discussion of the problem*". Coefficient alpha equaled .93. The second scale, named C*hairmanship*, consisted of 5 items (Eigenvalue 1.34, 25.94% of variance explained). An example of an item is: "*As chair, the student clearly motivated other students to participate in the discussion*". Coefficient alpha equaled .91. The third scale, named *Preparation*, consisted of 5 items (Eigenvalue 1.15, 22.08% of variance explained). An example of an item is "*The student's contributions to the group discussion were of high quality*". Coefficient alpha equaled .89. Two items from the questionnaire were not included in our study, as they showed incoherent factor loadings. Because the three scales demonstrated relatively high intercorrelations (r = .73 - .88), we also included a combined measure of the three scales, which we called *Observed learning activities.* Coefficient alpha of this 17-item scale equaled .96.

GPA. With authorization from the head of the department, the educational office provided the grades of the study participants for all courses. To obtain information on the reliability of this criterion measure, we computed the internal consistency with the grades for each course as items. Coefficient alpha equaled .91.

Results

Means, standard deviations, scale reliabilities and zero-order correlations between the webcam test, the cognitive ability test, the personality questionnaire, and the criterion measures included in this study are presented in Table 1. Before testing the hypotheses, we first looked at significant correlations between demographic characteristics and scores on our predictors and criterion measures. Age was significantly and positively related to emotional stability (r = .24, p < .01), openness to experience (r = .19, p < .05), webcam test scores (r = .19, p < .05), participation during the group meetings (r = .20, p < .01), preparation for the group meetings (r = .18, p < .05), and observed learning activities (r = .17, p < .05). Gender was related to a number of predictors. Differences between male students and female students were found for cognitive ability (r = -.22, p < .01, t = 2.77, p < .01), extraversion (r = -.18, p < .05, t = 2.27, p < .05), emotional stability (r = -.23, p < .01, t = 2.82, p < .01), and openness to experience (r = -.20, p < .01, t = 2.52, p < .05), all in favor of male students. Because of these significant correlations, we controlled for age and gender in the correlation and regression analyses.

Construct validity

Controlled for age and gender, scores on the webcam test showed significant correlations with a number of personality traits, namely extraversion (r = .26, p < .01), conscientiousness (r = .21, p < .05), and emotional stability (r = .19, p < .05). No significant partial correlations were found between scores on the webcam test and agreeableness (r = .05, *ns*), openness to experience (r = .14, *ns*), and cognitive ability (r = .01, *ns*). To test Hypothesis 1, which stated that scores on a webcam test for

interpersonally oriented leadership skills would be more strongly related to the personality factors extraversion and conscientiousness than to cognitive ability, we used Steiger's *z* statistic (Steiger, 1980). The partial correlations between the scores on the webcam test and extraversion and conscientiousness indeed were significantly higher than the partial correlation between the scores on the webcam test and cognitive ability ($z = 2.11$, $p < .05$ and $z = 1.65$, $p < .05$ respectively). Additionally, we compared the partial correlation between the scores on the webcam test and emotional stability with the partial correlation between the scores on the webcam test and cognitive ability. Results showed the correlation between the scores on the webcam test and emotional stability to be significantly higher than the correlation between the scores on the webcam test and cognitive ability ($z = 1.65$, $p < .05$). Based on these results, the first hypothesis could be supported.

Hypothesis 2 was that scores on a webcam test for interpersonally oriented leadership skills would be positively related to leadership experience. After controlling for age and gender, a significant and positive correlation between scores on the webcam test and leadership experience ($r = .28$, $p < .01$) was found, lending support for our second hypothesis. Together, personality, cognitive ability, and leadership experience explained 17% of the variance in webcam test performance ($F = 3.55$, $p < .01$).

Criterion-related validity

Hypothesis 3 stated that scores on a webcam test for interpersonally oriented leadership skills would have higher validity for predicting students' observed learning activities than for GPA. Controlled for age and gender, scores on the webcam test showed significant correlations with observed learning activities ($r = .26$, $p < .01$), and the separate dimensions of participation ($r = .26$, $p < .01$), chairmanship ($r = .25$, $p < .01$), and preparation ($r = .23$, $p < .01$). No significant partial correlation was found between scores on the webcam test and GPA ($r = .03$, *ns*). The partial correlations between webcam test scores and observed learning activities ($z = 2.80$, $p < .01$), participation ($z = 2.70$, $p < .01$), chairmanship ($z = 2.47$, $p < .01$), and preparation ($z = 2.48$, $p < .01$) were significantly higher than the partial correlation between webcam test scores and GPA, lending support for our third hypothesis.

Incremental validity

Hypothesis 4 stated that a webcam test for interpersonally oriented leadership skills would incrementally predict students' observed learning activities over and above a cognitive ability test and a personality questionnaire. To test this hypothesis, a series of hierarchical regression analyses were conducted. In these analyses, age, gender, and leadership experience were entered in the first step, the Big Five personality dimensions in the second step, cognitive ability in the third step, and the webcam test in the final step. The same stepwise regressions were used for each academic performance measure. The results for the regression analyses are presented in Table 2.

Table 1

Means, Standard Deviations, and Zero-Order Correlations between the Predictors and the Academic Performance Measures

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Age	22.33	3.17	(-)														
2. Gender	.66	0.48	-.02	(-)													
3.Leadership experience	0.34	0.80	.12	-.21*	(-)												
Predictors																	
4. Cognitive ability	44.04	11.28	-.04	-.22**	.11	(.73)											
5. Extraversion	3.55	0.52	.13	-.18*	.18*	.02	(.92)										
6. Agreeableness	3.74	0.30	.07	.04	.16	.01	.22**	(.83)									
7. Conscientiousness	3.66	0.38	.15	.12	.14	-.17*	.10	.31**	(.92)								
8. Emotional stability	3.30	0.43	.24**	-.23**	.22**	.15	.34**	.25*	.13	(.88)							
9. Openness to experience	3.79	0.29	.19*	-.20*	.22**	.08	.51**	.37**	.17*	.30**	(.90)						
10. Webcam test	24.33	4.98	.19*	.02	.26**	.02	.26**	.07	.24**	.21**	.10	(.80)					
Academic performance measures																	
11. Participation	3.78	0.40	.20*	.00	.04	.17*	.27**	.03	.23**	.09	.15	.29**	(.93)				
12. Chairmanship	3.86	0.36	.05	.04	-.01	.09	.19*	.15	.24**	.05	.05	.24**	.73**	(.91)			
13. Preparation	3.60	0.42	.18*	.02	.02	.13	.12	.07	.29**	.07	.05	.25**	.88**	.78**	(.89)		
14. Observed learning activities	3.75	0.38	.17*	.02	.03	.15	.20*	.08	.28**	.07	.09	.28**	.96**	.88**	.95**	(.96)	
15. GPA	6.39	0.87	.04	.16	-.08	.23**	-.11	-.05	.18*	-.14	-.09	.04	.47**	.42**	.54**	.52**	(.91)

Note. Coefficient alphas are presented on the diagonal. Personality scales, participation, chairmanship, preparation, and observed learning activities were measured on a 5-point scale, cognitive ability is the number of correct answers with a maximum of 81, the scores on the webcam test had a maximum of 50, and GPA was measured on a scale from 1-10. Gender (0 = male, 1 = female) and leadership experience (0 = no experience, 1 = less than one year, 2 = one to five years, 3 = 6-10 years, and 4 = more than 10 years) were coded. $N = 153$.

* $p < .05$, ** $p < .01$

The webcam test was able to explain 4% ($p < .05$) of the variance in participation during group meetings, 4% ($p < .05$) of the variance in chairmanship during group meetings, 3% ($p < .05$) of the variance in preparation for the group meetings, and 4% ($p < .05$) of the variance in observed learning activities beyond the variance explained by age, gender, personality and cognitive ability. Regarding participation, a significant beta weight was found for webcam test scores (β = .23, $p < .05$). Regarding chairmanship and preparation, significant beta weights were found for conscientiousness (β = .20, $p < .05$ and β = .27, $p < .01$ respectively) and webcam test scores (β = .21, $p < .05$ and β = .20, $p < .01$ respectively). Regarding observed learning activities in general, significant beta weights were found for conscientiousness (β = .24, $p < .05$), cognitive ability (β = .19, $p < .05$), and webcam test scores (β = .22, $p < .05$).

Table 2

Summary of Hierarchical Regression Analyses of Predictors on the Academic Performance Measures

	Participation		Chairmanship		Preparation		Observed learning activities		GPA	
	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2
Step 1										
Age	.14		.04		.18		.14		.11	
Gender	-.02		.06		-.04		.01		.07	
Leadership experience	-.09	.04	-.09	.01	-.09	.05	-.08	.03	-.12	.03
Step 2										
Extraversion	.19		.16		.06		.15		-.12	
Agreeableness	-.05		.07		.01		-.01		-.08	
Conscientiousness	.17		.20*		.27**		.24*		.27**	
Emotional stability	-.14		-.07		-.12		-.13		-.18	
Openness to experience	.05	.08	-.09	.06	-.03	.06	-.03	.07	-.04	.08
Step 3										
Cognitive ability	.16	.02	.17	.03	.17	.03	.19*	.03*	.23*	.05*
Step 4										
Webcam test	.23*	.04*	.21*	.04*	.20*	.03*	.22*	.04*	.08	.01
R^2		.19		.13		.17		.17		.16
F		2.59*		1.71*		2.26*		2.36*		2.15*

Note. Standardized regression weights are for final step. Gender (0 = male, 1 = female) and leadership experience (0 = no experience, 1 = less than one year, 2 = one to five years, 3 = 6-10 years, and 4 = more than 10 years) were coded. $N = 153$

* $p < .05$, ** $p < .01$

The webcam test was not able to explain additional variance in GPA beyond the variance explained by age, gender, personality and cognitive ability. Regarding GPA, significant beta weights were found for conscientiousness ($\beta = .27$, $p < .01$) and cognitive ability ($\beta = .23$, $p < .05$). Together, the predictors explained 19% of the variance in participation ($F = 2.59$, $p < .05$), 13% of the variance in chairmanship ($F = 1.71$, $p < .05$), 17% of the variance in preparation ($F = 2.26$, $p < .05$), 17% of the variance in observed learning activities ($F = 2.36$, $p < .01$), and 16% of the variance in GPA ($F = 2.15$, $p < .05$). Based on these results, our fourth hypothesis was supported.

Discussion

The aim of the present study was to investigate the construct validity and the criterion-related validity of a particular multimedia SJT with a constructed-response item format, namely a webcam test. This was done by examining its relationship with personality, cognitive ability, previous job experience, and academic performance. First of all, the results showed that scores on the webcam test were related to extraversion, conscientiousness, and emotional stability. Previous studies regarding the construct validity of paper-and-pencil leadership SJTs with a multiple-choice format have found similar relationships between SJT performance and personality (Bergman et al., 2006; Oswald et al., 2004). In line with the first hypothesis, scores on the webcam test were more strongly related to extraversion and conscientiousness than to cognitive ability. In addition, it appeared that scores on the webcam test were more strongly related to emotional stability than to cognitive ability. Previous studies have found a relationship between cognitive ability and paper-and-pencil SJTs with a multiple choice format (e.g., Bergman et al., 2006; McDaniel et al., 2001). However, we expected scores on a webcam test to be less strongly related to cognitive ability than paper-and-pencil SJTs with a multiple choice format, because a webcam test has no reading component, and because the constructed-response item format of a webcam test measures the participants' actual interpersonally oriented skills in job-related situations (Motowidlo et al., 2008). The finding that webcam test scores are not related to cognitive ability may have important practical implications. Many organizations believe it is important for business and ethical reasons to create a diverse workforce. As selection instruments with smaller cognitive loading produce smaller subgroup differences (Ployhart & Holtz, 2008), using a webcamtest to measure interpersonally oriented skills, such as leadership skills, may be an effective strategy to reduce adverse impact. Previous studies have already shown that the use of multimedia in SJTs reduces their cognitive loading (Lievens & Sackett, 2006). Perhaps using a constructed-response item format may even further reduce the cognitive loading of an SJT. We recommend future studies to investigate whether the use of a constructed-response item format indeed affects the cognitive loading of an SJT and the subsequent subgroup-differences in test performance.

In line with the second hypothesis, the results showed that webcam test performance was related to leadership experience. This finding seems logical, as people with job relevant experiences are more likely to have encountered the types of job-

related situations presented in the webcam test and have learned how to respond successfully to these types of situations. Although prior studies have also found job experience to be related to performance on SJTs (e.g., Clevenger et al., 2001; McDaniel & Nguyen, 2001), the present study is the first to demonstrate the relationship between leadership experience and performance on a webcam test intended to measure interpersonally oriented leadership. We found a significant relationship between leadership experiences and webcam test performance, while our sample was rather homogeneous regarding age and educational level. Moreover, only 17% of the participants had experience as a leader or supervisor. It is possible that we would have found an even stronger relationship between leadership experience and test performance among actual applicants, because an actual applicant sample would have been more heterogeneous and would have included more participants with relevant job experience. A strong correlation between job experience and webcam test performance would support the assumption of Stemler and Sternberg (2006) that SJTs are measures of practical intelligence, as practical intelligence requires knowledge about how to deal effectively with job-related situations that occur in the context of everyday experiences.

However, in the present study a large part of the variance in webcam test performance is unaccounted for by personality, cognitive ability, and previous experience, namely 83%. Thus, the webcam test seems to measure a construct that is relatively independent of most other individual difference variables that are frequently assessed in personnel selection practices. Whether this construct indeed is interpersonally oriented leadership skills, which is the construct the webcam test intended to measure, or some other situational judgment construct, is not yet entirely clear. It might be helpful to include the webcam test in a multitrait-multimethod matrix (D. T. Campbell & Fiske, 1959) to assess the degree to which there is trait convergence across various measures of interpersonally oriented leadership skills.

With respect to the criterion-related validity of the test, our results support the validity of a webcam test as predictor of academic performance. Scores on the webcam test predicted students' participation during group meetings, how well the students performed their role as a chair during the group meetings, their preparation for these meetings, and the observed learning activities in general. In line with our expectations, the webcam test showed higher validity for predicting students' observed learning activities than for GPA. Similar to the results of Oswald et al. (2004) the webcam test was not related to GPA. These findings suggest that a multimedia SJT with a constructed-response item format can be a valid predictor of academic performance criteria, but as Lievens et al. (2005) suggested, it is important to differentiate within the criterion domain. In the present study, the predictor and criterion domain were carefully specified, as we examined whether a webcam test intended to measure interpersonally oriented leadership was able to predict students' leadership-related behaviors, such as demonstrating skills in a group, motivating others, and coordinating groups and tasks (Oswald et al., 2004).

In line with the fourth hypothesis, the webcam test incrementally predicted students' observed learning activities over and above a cognitive ability test and a personality questionnaire. Similar to previous studies regarding the incremental validity of multimedia SJTs (Lievens et al., 2005; Lievens & Sackett, 2006), the webcam test was able to explain a unique part of variance in academic performance. Our study, thus, demonstrates that a multimedia SJT with a constructed-response item format can be a useful and valid complement to traditional predictors in selection contexts. Together, the predictors explained a substantial part of the variance in academic performance, ranging from 13% of how well the students performed their role as a chair during the group meetings to 19% of students' participation during these meetings.

The present study is one of the first to examine the construct validity and criterion-related validity of a multimedia SJT with a constructed-response item format. Therefore, we believe that this study makes an important contribution to the literature on multimedia SJTs. To summarize, the study demonstrated that scores on an interpersonally oriented leadership webcam test are related to extraversion, conscientiousness, emotional stability, and leadership experience. Furthermore, the study demonstrated that webcam tests can be useful and valuable predictors of academic performance beyond traditional measures as cognitive ability tests and personality questionnaires.

Limitations of this study and suggestions for future research

The current study has some general limitations that should be noted. The first limitation relates to the study setting. Results were obtained in a research setting, which typically lacks the motivational and self-presentational issues inherent in actual high-stakes situations. We attempted to motivate the students to perform well on the different instruments by emphasizing the benefits they could have by practicing with real selection instruments and by giving them a professional report of their scores, but it remains possible that motivational difference between our participants and real applicants exist. Therefore, it is important to replicate our findings in a field study using an actual applicant sample. An applicant sample would also provide the opportunity to assess important issues such as adverse impact or differential prediction.

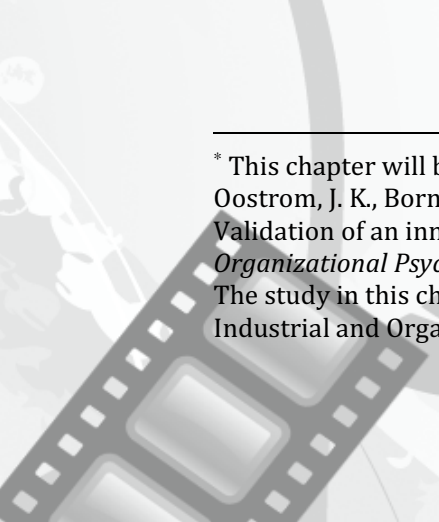
Furthermore, in the webcam test used in this study, the filmed responses had to be rated afterwards. This scoring method is quite costly. As an important advantage of computer technology lies in the automatic scoring, perhaps in the future the costs of scoring could be reduced by taking advantage of the multimedia approach by using voice-recognition software or other automatic scoring possibilities (Powers et al., 2002).

In future studies, it would be interesting to compare the construct validity and criterion-related validity of a multimedia SJT with a multiple choice item format with a multimedia SJT with a constructed-response item format, measuring the same construct with the same situational stimuli. By holding the predictor construct

constant, conclusions can be drawn about the effects of the response format. A number of studies have distinguished between predictor constructs and predictor methods, for example to examine the effects of test medium (video-based versus written SJT) on predictive validity (Chan & Schmitt, 1997; Lievens & Sackett, 2006). However, until now, no studies have examined the effects of the response format of multimedia SJTs on their construct validity and criterion-related validity.

Chapter 3

Webcam testing: Validation of an innovative open-ended multimedia test*



* This chapter will be published as:
Oostrom, J. K., Born, M. Ph., 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.*
The study in this chapter was also presented at the 24th annual conference of the Society of Industrial and Organizational Psychology (SIOP), New Orleans, LA, April 2009.

Abstract

A modern test that takes advantage of the opportunities provided by advancements in computer technology is the multimedia test. The purpose of this study was to investigate the criterion-related validity of a specific open-ended multimedia test, namely a webcam test, by means of a concurrent validity study. In a webcam test a number of work-related situations are presented and participants have to respond as if these were real work situations. The responses are recorded with a webcam. The aim of the webcam test which we investigated is to measure the effectiveness of social work behavior. This first field study on a webcam test was conducted in an employment agency in The Netherlands. The sample consisted of 188 consultants who participated in a certification process. For the webcam test, good inter-rater reliabilities and internal consistencies were found. The results showed the webcam test to be significantly correlated with job placement success. The webcam test scores were also found to be related to job knowledge. Hierarchical regression analysis demonstrated that the webcam test has incremental validity over and above job knowledge in predicting job placement success. The webcam test, therefore, seems a promising type of instrument for personnel selection.

Introduction

The use of advanced technology in personnel selection practices is increasing (Anderson, 2003). More and more psychological tests and questionnaires are administered via computers. The computer has established itself as an efficient tool for administering, scoring and interpreting personnel selection tests (Lievens et al., 2002). Although this development is important for personnel selection practices, the advancements in information technology provide a lot more opportunities (McHenry & Schmitt, 1994). An example of a modern test that takes advantage of the opportunities provided by computer technology is the multimedia test. In multimedia tests realistic work samples are presented via the computer (Funke & Schuler, 1998; Weekley & Jones, 1997). The typical multimedia test consists of a number of video scenarios followed by a series of pre-coded responses an applicant has to choose from (Weekley & Ployhart, 2006). This kind of multimedia test is called a multimedia or video-based situational judgment test (SJT). Another form of multimedia testing is a test with a constructed-response item format, in which applicants are asked to actually respond in their own words to the presented situation. In this kind of multimedia test, not only the situation has become more realistic, but also the manner of responding (Funke & Schuler, 1998). However, there is a lack of studies that critically evaluate the reliability and validity of open-ended multimedia tests (e.g., Lievens et al., 2002). This paper addresses this shortcoming by investigating the criterion-related validity of a specific open-ended multimedia test, the so-called webcam test. We will begin with a discussion of the research on situational tests, followed by a summary of the research on the criterion-related validity of multimedia situational tests and open-ended multimedia tests, and then will propose hypotheses about the criterion-related validity of the webcam test.

Situational tests

Situational tests have become very popular in personnel selection practices (Ployhart & Ehrhart, 2003). These tests are designed to sample behaviors, as opposed to traditional predictors that provide signs of underlying temperament or other traits that are assumed to be necessary for job performance (Motowidlo et al., 1990). Samples or simulations are based on the notion of behavioral consistency. The behavior of applicants in situations similar to those encountered on the job is assumed to provide a good prediction of actual behavior on the job (Schmitt & Ostroff, 1986).

A situational test that recently has garnered serious attention in research and practice, is the SJT (e.g., Chan & Schmitt, 2005; Weekley & Ployhart, 2006). In an SJT, applicants are presented with a variety of situations they are likely to encounter on the job. These situations are usually derived from critical incidents interviews. After each situation a number of possible ways to handle the hypothetical situation is presented. The applicant is asked to judge the effectiveness of the responses in either a forced-choice or Likert-style format.

The psychometric properties of paper-and-pencil SJTs have been evaluated in several studies (e.g., Bergman et al., 2006; Lievens & Sackett, 2006; McDaniel et al., 2007). McDaniel et al. (2007) demonstrated in their meta-analysis that SJTs are valid predictors of job performance (average observed r = .20). SJTs show substantial correlations with other predictors, such as cognitive ability (McDaniel & Nguyen, 2001), and Big Five personality dimensions (e.g., Clevenger et al., 2001; Weekley & Ployhart, 2005). SJTs are also found to be significantly related to job experience (e.g., Weekley & Jones, 1997) and declarative job knowledge (e.g., Clevenger et al., 2001). Even with these significant correlations, several studies have shown that SJTs have incremental validity over and above traditional predictors (Chan & Schmitt, 2005), suggesting SJTs capture a unique part of job performance. For example, Clevenger et al. (2001) demonstrated that an SJT provides incremental validity over cognitive ability, declarative job knowledge, job experience, and conscientiousness. Similarly, McDaniel et al. showed that SJTs have incremental validity over cognitive ability and the Big Five personality dimensions.

Which constructs situational tests capture, is still unclear (McDaniel & Nguyen, 2001). There is a discussion in the literature concerning what situational tests measure. It has been argued that situational tests capture a unique construct. According to Wagner and Sternberg (1985) SJTs measure *tacit knowledge*. Tacit knowledge has been conceptualized as "practical know-how that usually is not openly expressed or stated and must be acquired in the absence of direct instructions" (Wagner, 1987, p. 1236). Other researchers argue that situational tests reflect a number of constructs that are related to job performance (Weekley & Jones, 1999). For example, Chan and Schmitt (1997) have argued that a situational judgment problem is nearly always multidimensional in nature, because solving the problem would involve several abilities and skills. In other words, SJTs according to these researchers mediate the effect of several predictors, such as cognitive ability and job experience (Weekley & Jones, 1999). Finally, F. L. Schmidt (1994) has argued that SJTs measure job knowledge. Job knowledge, in turn, has been consistently found to be related to job performance, cognitive ability, and experience (F. L. Schmidt, Hunter, & Outerbridge, 1986).

Multimedia tests

Recent technological advances have led researchers to explore the possibilities of using multimedia applications in situational tests (Anderson, 2003). The use of multimedia or video provides the opportunity to give a more realistic presentation of work situations (Funke & Schuler, 1998). Multimedia tests have several important advantages compared to traditional selection instruments. By utilizing video and graphics, it is possible to portray detailed and accurate job-related scenarios, which increases the fidelity of the test (Dalessio, 1994). The scenarios provide a realistic job preview to the applicant and are therefore more attractive for applicants in terms of their interest and motivation than traditional paper-and-pencil tests (Stricker, 1982). Richman-Hirsch et al. (2000) demonstrated that compared to a written test, the

multimedia version yielded more positive applicant reactions, even though the linguistic content was identical. The multimedia assessment was perceived as more content valid, more face valid, more enjoyable and led to more satisfaction with the assessment process. Another important advantage is that multimedia tests result in less adverse impact (Goldstein, Braverman, & Chung, 1992). Chan and Schmitt (1997) demonstrated that reading comprehension is uncorrelated with test performance on a multimedia SJT, resulting in less adverse impact compared to the paper-and-pencil version.

The main question in personnel selection is whether a selection instrument is able to predict job performance. Various studies have examined the predictive validity of the multimedia SJT. For example, Dalessio (1994) found a significant relationship between test scores on a multimedia SJT and turnover. Weekley and Jones (2004) developed and validated two multimedia SJTs, one for hourly service workers and one for home care-givers. The SJT scores in both cases provided predictive validity over and above cognitive ability and experience. Olson-Buchanan et al. (1998) developed and validated an interactive video assessment of conflict resolution skills. The video assessment was significantly related to supervisory ratings, collected for research purposes, of how well the assessees dealt with conflict on the job, but it was unrelated to cognitive ability. In a meta-analysis, Salgado and Lado (2000) demonstrated that multimedia tests are good predictors of job performance, with an average observed validity of .25. The gain in validity by adding a multimedia test over other ability measures was .10.

Lievens and Coetsier (2002) described the development of two video-based SJTs as part of an admission exam for medical and dental studies. Four cognitive ability tests and two other situational tests, namely work samples, were also part of this admission exam. Unlike the cognitive ability tests and the other situational tests, the multimedia SJTs in this study did not emerge as significant predictors of first year performance in medical school. According to Lievens and Coetsier the difference in predictive validity of the multimedia SJTs and the other situational tests could be explained by the fidelity of the tests. Simulations vary in the fidelity with which they present a stimulus and elicit a response (Motowidlo et al., 1990). The highest fidelity simulations use very realistic stimuli to represent a task situation and provide applicants with the opportunity to respond as if they were actually in the job situation. Low fidelity simulations simply present a verbal description of a hypothetical work situation, instead of a concrete representation, and ask candidates to describe how they would deal with the situation or to choose a response alternative. In a multimedia SJT the scenarios have an increased fidelity compared to other selection tools. However, the manner of responding has little fidelity, because candidates are not asked to show actual behavior. Instead, they have to choose among a number of response alternatives (Lievens & Thornton, 2005). Therefore, the test may mainly capture the candidates' insight instead of their actual behavior (Lievens et al., 2002; McDaniel & Nguyen, 2001).

Previous studies on multimedia tests have mainly addressed the realism of the stimuli, but Funke and Schuler (1998) demonstrated that response fidelity is also an important aspect. In their study among 75 college students, a comparison was made between various types of multimedia tests. The tests differed in the fidelity of the presented situation (either orally or via video) and the fidelity of the responses (multiple-choice, written free, or oral free). The fidelity of the situation had no impact upon the validity. However, the criterion-related validity of the tests with orally-given free responses was significantly higher than the criterion-related validities of the tests with a multiple choice format and a written response format. In their study, Lievens and Coetsier (2002) also had included situational tests with a high response fidelity, namely work samples. They found that the higher the response fidelity, the higher the predictive validity of the situational tests. In order to maximize the validity of multimedia tests, test developers should, therefore, also focus on response fidelity.

Open-ended multimedia tests

A multimedia test with high response fidelity is one with an open-ended format. In this kind of multimedia tests, job-related situations are presented to the applicants in the same way as in an SJT. After the situation has been presented, the applicant is asked to respond as if it were a real situation. These responses are filmed and judged afterwards by two or more subject matter experts (SME's) on their effectiveness. Because the aim of a situational test is to assess whether or not applicants can behave appropriately and successfully in work-related situations, an open-ended format seems more appropriate than a multiple-choice format, because it allows for a direct and spontaneous expression of a behavioral competency (Funke & Schuler, 1998).

Research on open-ended multimedia tests is relatively scarce (Funke & Schuler, 1998). Next to the study of Funke and Schuler (1998), we were able to trace only the following publications on open-ended multimedia tests that were used for selection purposes. Stricker (1982) developed the first open-ended multimedia test, called the 'Interpersonal Competence Instrument' (ICI), and administered it to 58 female college students. In the ICI, scenes were presented in which a subordinate talks to a superior in a business setting. The inter-rater reliability (r varied from .53 to .90) and internal consistency (α varied from .74 to .82) were substantial and the correlations with other tests supported its construct validity. Based on the findings of Stricker, three open-ended multimedia tests were developed in The Netherlands between 1982 and 1993 to measure the interpersonal competences of managers (Meltzer, 1995). Multiple studies were conducted to shed light on the psychometric properties of these tests, with small samples varying between 5 and 59. General findings were in line with the results reported by Stricker in terms of the internal consistency and the inter-rater reliability.

In their review on multimedia tests, Olson-Buchanan and Drasgow (2006) describe an open-ended multimedia test developed by researchers from the U.S. Customs and Border Protection to assess future border patrol officers (Walker & Goldenberg, 2004, as described in Olson-Buchanan & Drasgow, 2006). Inter-rater reliabilities

ranging from .67 to .78 were found. Olson-Buchanan and Drasgow argue that the open-ended response format is an innovative feature of multimedia situational testing, and research regarding the validity of multimedia tests with this response format should be conducted.

Present study: Webcam testing

In the present study, we investigated the criterion-related validity of an open-ended multimedia test by means of a concurrent validity study. So far, to our knowledge the criterion-related validity of an open-ended multimedia test has not been investigated with measures of actual work performance. Until now, studies on open ended multimedia tests mainly have addressed their internal consistency and inter-rater reliability. The criterion-related validity has only been investigated with samples that largely consisted of college students and actual work performance measures have not yet been used as a criterion (Funke & Schuler, 1998; Stricker, 1982). Consequently, the main goal of this study is to examine the correlation between an open-ended multimedia test and actual measures of work performance, specifically of employment consultants. The criterion measures included in this study are objective job placement success of the consultants' job seeking clients and the manager's appraisal of their work performance.

In the specific open-ended multimedia test used for this study (the webcam test) a number of important work-related situations are presented to the participant, which involved interactions with job seekers. The test was intended to measure effectiveness in the core task of an employment consultant, namely advising job seekers. The webcam test distinguishes itself from other situational tests because of the behavioral response format and by using a small webcam to film the responses of the participants, instead of a video recorder.

The first aim of this study was to investigate the criterion-related validity of the webcam test. Because the webcam test is a high fidelity test, in which realistic stimuli are presented and applicants are provided with the opportunity to respond as if they were actually in the job situation, we expected the webcam test to be positively related to job performance. As noted above, the predictive validity of an open-ended multimedia test has not yet been investigated with measures of actual work performance. However, various studies (Funke & Schuler, 1998; Lievens & Coetsier, 2002) have demonstrated that the fidelity of the responses may positively affect the predictive validity, with relatively high criterion-related validity occurring for a multimedia test with orally-given responses Thus, on the basis of these arguments, our hypothesis is as follows:

Hypothesis 1: There is a positive relation between scores on the webcam test and job performance.

In the present study we also investigated the relation between the webcam test and job knowledge. F. L. Schmidt (1994) has argued that situational tests are nothing

more than tests of job knowledge. If situational tests measure job knowledge, they should strongly relate to a job knowledge test (Weekley & Jones, 1997). McDaniel and Nguyen (2001) demonstrated in their meta-analysis that measures of job knowledge, usually operationalized as measures of job experience, are indeed positively related to situational judgment tests. Based on this finding, McDaniel and Nguyen have argued that situational judgment tests owe some of their criterion-related validity due to their assessment of job knowledge. Therefore, we will examine whether the webcam test is able to explain unique variance in job performance over and above job knowledge. As the webcam test measures actual behavior, it is likely that it will be a unique predictor of job performance. Our two next hypotheses therefore are:

Hypothesis 2: The webcam test is positively related to job knowledge.

Hypothesis 3: The webcam test incrementally predicts job performance over and above job knowledge.

Method

Participants and procedure

We collected data in 2007 among 188 consultants working for a public employment agency in The Netherlands. The consultants' main task is helping people to find a job by giving advice, information, and emotional support. Adequate communication with their clients is a key aspect of their job. Of the participants, 108 were female (57.0%) and 80 were male (43.0%). Their age ranged from 23 to 59 (M = 42.0, SD = 8.51). The participants had worked for 4.7 years on average (SD = 0.89) in the organization and for 31.4 hours on average (SD = 5.77) per week. Their education level ranged from high school to master's degree. Most participants had a higher vocational bachelor's degree (76.1%).

The organization offered its consultants the opportunity to obtain a certificate which demonstrates their competence level. The certification procedure consisted of an assessment through a webcam test, a job knowledge test and a performance rating. Consultants could obtain the certificate after they had passed all three tests. The performance rating consisted of two measures: 1) an objective measure of job success, namely the percentage of the consultant's clients over the last year that had found a job, and 2) a manager's appraisal. The manager's appraisal was provided in the form of a questionnaire filled out by the manager of the consultant by judging the consultants' job performance over the last year. In total, 56 different managers filled out the questionnaire. The objective measure of job success was only available for 90 consultants.

With approval of their manager, consultants voluntarily participated in this certification process. To determine whether participation in the certification process was self-selective, which would mean that the participants were not representative of all the consultants in the organization, we compared their age, years of experience, and the percentage of their clients during the last year that had found a job, to those of the

other consultants (N = 4459). Of these other consultants, 1814 (40.7%) had already obtained a certificate in preceding years. The participants were significantly younger (M = 42.0, SD = 8.57) in comparison to the other consultants (M = 44.4, SD = 9.84, t = 2.91, p < .01, d = 0.23), but this age difference is small. This finding is not surprising because as employees get older, they tend to participate less in training and development activities than younger employees (Maurer, 2001). Years of experience of the participants (M = 4.6, SD = 0.94) did not differ significantly from the other consultants (M = 4.7, SD = 0.78), and also the percentage of the participant's clients of the last year that had found a job (M = 42.5, SD = 9.31) did not differ significantly from the other consultants (M = 42.6, SD = 4.15). Therefore, we concluded that there were no selection effects regarding age, experience and job placement success that could affect our results.

The assessors of the webcam test were 22 senior consultants from the organization itself, who had been trained in evaluating the participants' responses in a course specifically developed for this purpose by an experienced psychologist. This training is explained in more detail in the next paragraph. Of the assessors, 13 were female (59.1%) and 9 were male (40.9%). Their age ranged from 33 to 56 (M = 44.7, SD = 8.00). Their education level ranged from intermediate vocational education to master's degree. Most assessors had a higher vocational bachelor's degree (63.6%).

Measures

Webcam test. The webcam test was developed by a Dutch HRD consultancy firm in close cooperation with the public employment agency. The webcam test aimed to measure effectiveness in the central task of the employment consultant, namely consulting job seekers. Input for the situations came from critical incidents interviews with 10 experienced consultants. Scripts for 12 scenarios were written and videotaped by a production company. Each scene starts with an oral description of the situation, followed by a fragment of a possible conversation between a job seeker and the participant (consultant) in their role of employment consultant. In this fragment a professional actor, playing the job seeker, talks directly to the camera, as if speaking to the participant. After this, the frame freezes and the participant has to respond as if it were a real situation. These responses are filmed with a webcam. The response time is limited to one minute, which is long enough to react to the situation at hand. The total duration of the webcam test is about 45 minutes. An example of a situation in the webcam test is: "*You have an appointment with an elderly client. The client has been looking for a job for several months now, but has not succeeded in finding a job (oral introduction)". Job seeker: "It's obvious why I can't find a job. Who wants to hire someone over his fifties nowadays? There are plenty of young applicants they can choose from who are far less expensive!*". The effectiveness of the responses were judged afterwards by three trained subject matter experts (SME's), with many years of experience as a consultant, who gave their ratings independently of one another and worked on the basis of a set of comprehensive scoring instructions. The scoring instructions and the participants' videotaped responses were available via internet.

The responses were rated on a five-point scale ranging from (--) *very ineffective* to (++) *very effective.* In the example given above, aspects of an effective response are: Showing empathy for the client, explaining the procedures of the employment agency, admitting the fact that it is more difficult to find a job for elderly applicants than for young applicants, and focusing on the positive aspects of being an elderly employee (e.g., years of experience). Aspects of an ineffective response are: Trivializing the problem of the client, not providing information to the client, and focusing on the negative aspects of being an elderly employee. For each response the mean score of the three assessors was calculated. The 12 scores were summed and divided by the maximum obtainable score, resulting in an overall score that could range from 0 – 100.

The assessors received a frame-of-reference (FOR) training consisting of 1) an introduction about the basics of rating processes and the possible rating errors that can occur, and 2) a workshop on the rating process, in which the assessors were taught what effective and ineffective behaviors were in the specific situations of the webcam test (Bernardin & Buckley, 1981). Examples of very effective, average and very ineffective responses were demonstrated for each situation. The assessors rated each response on the five-point scale and submitted their justification for each rating. Then, the trainer informed the assessors what the correct rating for each response was and gave the rationale behind this rating. The assessors had the opportunity to discuss any discrepancies between their ratings and the rationale that was given by the trainer. The total duration of the training was 4 hours. After the first training, as prior practice the assessors had to evaluate the responses of three participants. These ratings were then compared to the ratings of experienced psychologists and discussed during a second meeting. The second meeting took about 2 hours.

Job knowledge test. The job knowledge test measures whether the participant has enough knowledge to perform his or her job effectively. The job knowledge test was very carefully constructed according to the following steps. First, the relevant topics were determined by a group of experienced consultants and managers working at the public employment agency, with the intention to cover all knowledge domains. For the job knowledge test in this study 11 relevant topics were determined, among others the labor market, general service delivery and available training and education programs. The second step was the development of the items. Based on the knowledge domain determined in the first step, critical incidents interviews were conducted by professional text writers and experienced consultants to develop the items. The items were written according to a specific format, namely a multiple choice or multiple select format. In the third step, an expert group independently of one another judged the items on their relevance and realism and estimated the percentage of participants that will answer the item correctly (*p*-value). To retain the items with the highest discriminating power, only the items with an average estimated *p*-value between .40 and .70 were included in the job knowledge test. Items outside this range were removed or re-written. After the job knowledge test was administered to

at least 100 participants, the *p*-value of each item was calculated for a second time. Again, items with a *p*-value below .40 or above .70 were removed or re-written. To prevent circulation of items among participants, each topic was represented by an item pool. From each item pool one to three questions were randomly selected, resulting in a different set of 15 items for each participant. An example of a multiple select item of the job knowledge test is: "*What are the consequences of a tight labor market?*". The answers the participants could choose from are: a) "*The number of vacancies that are difficult to fulfill, will grow*", b) "*Employers become more demanding in their recruitment of new personnel*", c) "*The wages will grow*", d) "*Organizations will increase computerization*", and e) "*Turnover will increase*". The number of correctly answered questions was divided by the total number of questions, resulting in an overall score that could range from 0 – 100.

Job performance. Job performance was measured with job placement success, which is an objective productivity measure, and a manager's appraisal of work performance. Both measures were existing performance data.

Job placement success consisted of two measures, namely the percentage of the participant's (consultant's) clients in 2006 that had found a job before receiving unemployment benefits, and the percentage of the participant's clients that found a job while receiving unemployment benefits. The average of the two measures formed the job placement success scale. Coefficient alpha of this two-item scale was .68. A job seeker becomes a participants' client after he or she registers at one of the departments of the public employment agency, and has been contacted by the participant. Participants therefore could not choose which job seeker to assist. On average, each consultant advises about 150 clients every year.

The manager's appraisal consisted of a questionnaire filled out by the participant's department manager, who judged the participant's individual task performance over the last year. Individual task performance involves learning the tasks and the context in which it is performed as well as being able and motivated to perform the required task (Murphy & Shiarella, 1997). The managers were aware of the fact that their appraisal was part of the certification procedure. This questionnaire consists of five items on a five-point scale ranging from 1 (*never*) to 5 (*always*). Examples of items are: "*The consultant puts a lot of effort in attaining his or her goals*", and "*The consultant has a substantial contribution to the outcomes of the department*". Coefficient alpha of this scale was .82.

Results

Means, standard deviations, reliabilities and correlations between the variables included in this study are presented in Table 1. Before we tested our hypotheses, we first looked at significant correlations between demographic characteristics and all study variables. The unemployment rate of the province the consultant worked in significantly correlated with job placement success (r = -.20, p <.05). Other demographic characteristics showed no significant correlations with our study variables.

Table 1

Means, Standard Deviations and Correlations between all Variables

	M	*SD*	1	2	3	4	5	6	7	8	9
1. Age	41.98	8.51	(-)								
2. Education	2.83	0.62	-.24*	(-)							
3. Gender	1.57	0.50	-.36**	.17*	(-)						
4. Unemployment rate	5.60	0.78	-.09	-.05	-.05	(-)					
5. Job tenure	4.65	0.89	.34**	-.13	-.08	.18*	(-)				
6. Webcam test	64.51	7.98	-.14	.05	.12	-.07	.00	(.82)			
7. Job knowledge test	68.77	10.98	-.10	.04	.01	-.01	-.02	.22**	(-)		
8. Job placement success	42.47	9.31	.15	.04	.19	-.20*	-.13	.26*	.21*	(.68)	
9. Manager's appraisal	4.10	0.51	-.08	.10	.09	-.12	-.09	.13	.13	.25*	(.82)

Note. Reliability coefficients are reported in parentheses on the diagonal. Education (1 = High school, 2 = Intermediate vocational education, 3 = Bachelor, 4 = Master) and gender (1= Male, 2 = Female) were coded. The unemployment rate and the scores on the webcam test, job knowledge test and productivity were on a scale from 0-100. The manager's appraisal was on a five-point sale. N = 188.

$^{*}p < .05$, $^{**}p < .01$

Reliability

The inter-rater reliability of the webcam test was tested with a two-way random intraclass-correlation (*ICC*). Every participant was judged by three SME's out of the larger pool of 22 SME's. The *ICC* per scene ranged from .41 to .81 (M = .65). The overall *ICC* was .71. The internal consistency of the webcam test, estimated by coefficient alpha, was substantial, namely .82.

Criterion-related validity

To test our first hypothesis, namely that there would be a positive relationship between the scores on the webcam test and job performance, we calculated Pearson product moment correlation coefficients. As Table 1 shows, the overall webcam test score manifested a significant and positive correlation with job placement success (r = .26, $p < .05$), but not with the manager's appraisal of job performance (r = .13, *ns*). These findings partly support our first hypothesis.

We tested our second hypothesis by examining the correlation between scores on the webcam test and the job knowledge test. As Table 1 shows, the webcam test scores are significantly related to job knowledge (r = .22, $p < .01$), which supports our hypothesis that the webcam test and job knowledge are positively related.

Moreover, the job knowledge test demonstrated a significant correlation with job placement success (r = .21, $p < .05$). This correlation does not significantly differ from the correlation between the webcam test and job placement success (z = -0.57, *ns*). In other words, the webcam test and job knowledge test do not differ significantly in their ability to predict job placement success. As was the case for the webcam test, the

job knowledge test was not significantly related to the manager's appraisal of job performance (r = .13, *ns*).

Table 2

Hierarchical Regression Analysis

	Job placement success (N = 90)				Manager's appraisal (N = 188)			
	ß	R^2	ΔR^2	F	ß	R^2	ΔR^2	F
Step 1								
Age	.24				-.01			
Gender	.22				.06			
Job tenure	-.10				-.06			
Unemployment rate	-.10	.12	.12	2.73*	-.11	.03	.03	1.20
Step 2								
Job knowledge test	.17	.16	.04	3.82*	.09	.04	.01	1.87
Step 3								
Webcam test	.20	.20	.04	3.68*	.07	.04	.00	.85

Note. Gender (1 = Male, 2 = Female) was coded. F-ratio's are for ΔR^2. Parameter estimates are for final step.

* p < .05, ** p < .01

We tested our third hypothesis, which stated that the webcam test would incrementally predict job performance over and above job knowledge, by examining the relationship between the job knowledge test and the webcamtest on the one hand, and both performance ratings on the other hand by conducting a hierarchical regression analysis, with the job knowledge test and the webcam test as independent variables and job placement success or the manager's appraisal as dependent variable. Age, gender, job tenure, and the unemployment rate of the province the consultant works in were entered as control variables in the first step, followed by the job knowledge test in step 2 and the webcam test in step 3. Table 2 displays the results of the hierarchical regression analyses. Regarding job placement success, after having controlled for age, gender, job tenure, and the unemployment rate, the job knowledge test explained an additional 4% of the variance in job placement success (ß = .17, F = 3.82, p < .05). When the webcam test was added in the next step, it explained an additional 4% of the variance in job placement success (ß = .20, F = 3.68, p < .05). We also conducted a hierarchical regression analysis to examine whether the job knowledge test had incremental validity over and above the webcam test in predicting job placement success. After having controlled for age, gender, job tenure, and the unemployment rate, the webcam test explained an additional 5% of the variance in job placement success (ß = .20, F = 4.85, p < .05). When the job knowledge

test was added in the next step, it explained an additional 3% of the variance in job placement success. However, this R^2 change was not significant (ß = .17, F = 2.67, *ns*).

We next turned to the prediction of the manager's appraisal, conducting the same analyses. As Table 1 already showed, the webcam test and the job knowledge test did not significantly relate to the manager's appraisal. Table 2 displays the results of the hierarchical regression analyses. Controlled for age, gender, job tenure and the unemployment rate of the province the consultant works in, the regression of the job knowledge test and the webcamtest on manager's appraisal demonstrated no significant results. Based on these results, it can be concluded that our third hypothesis is supported for the criterion job placement success, but not for the manager's appraisal.

Discussion

In this study the criterion-related validity of a specific open-ended multimedia test, namely a webcam test, was investigated. As an important prerequisite for attaining predictive validity, results of this first field study on the webcam test showed a substantial inter-rater reliability. This is consistent with previous studies on multimedia tests with an open-ended format (Funke & Schuler, 1998; Meltzer, 1995; Stricker, 1982). The subjective nature of this judgment process could potentially be seen as a disadvantage of the webcam test. However, by rater training, by using a set of comprehensive scoring instructions and by the use of multiple raters, our study shows that a substantial inter-rater agreement can be reached. In line with previous studies (Meltzer, 1995; Stricker, 1982), the internal consistency of the webcam test was high.

For the job placement success criterion, the results supported our hypothesis, which stated that the webcam test would be positively related to job performance. A key issue was whether the webcam test reflects job-specific knowledge, and thus whether this characteristic of the webcam test would be responsible for its predictive validity (e.g., F. L. Schmidt, 1994). If the webcam test measures job knowledge, it should strongly relate to a test developed to measure job knowledge (Weekley & Jones, 1997). Although, we did find a significant correlation between the two tests, this correlation was not very strong. The webcam test incrementally predicted job placement success over and above the job knowledge test, suggesting the webcam test measures more than just job knowledge. The regression analyses also demonstrated that the unemployment rate of the province in which the consultant worked was significantly related to job placement success. Controlled for this effect of unemployment rate, and also age, gender, job tenure and the job knowledge test, the webcam test still was able to explain additional variance in job placement success. For the practice of personnel selection the present findings thus indicate that the webcam test shows incremental validity over job knowledge. Therefore, the findings suggest that the webcam test is a relevant predictor of job performance.

The webcam test and the job knowledge test both nevertheless were not significantly related to the manager's appraisal. The hierarchical regression analysis of the

job knowledge test and the webcam test on the manager's appraisal similarly did not display a significant prediction. There are a number of limitations to the manager's appraisal of job performance that could explain these results. First, the questionnaire was filled out by 56 different managers. Most managers rated only one consultant. Therefore, the comparability of the scores may be questionable to a certain degree. Second, the scores were not normally distributed. There was little variance and a ceiling effect in the manager's appraisal, demonstrated by the overall mean of 4.10 on a five-point scale and the standard deviation of 0.51. These results could be explained by the fact that the managers had to approve participation in the certification process, leading to a select sample of motivated participants. Comparison of years of experience and job placement success of the participants in our study to all other consultants nevertheless yielded no significant differences. However, we were unable to control for other selection effects, such as motivational aspects. If self-selection effects would have occurred in this study, this may have attenuated the validity coefficient. A more random selection of consultants may have produced higher validity coefficients. Another explanation for the ceiling effect could be that the managers were aware of the fact that their appraisal was a part of the certification procedure. This could have led to a leniency in their judgments, which in turn, may have affected the criterion-related validity. Therefore, future studies may additionally want to use managers' appraisals collected for research purposes only, which may lead to less lenient judgments, and thus to larger criterion-related validities.

Motowidlo et al. (2006a) have argued that SJTs are measures of procedural job knowledge. Thus, the fact that the job knowledge test in the present study consisted mainly of questions regarding declarative job knowledge may have its limitations. Certainly the nature of the items in most SJTs suggests that procedural job knowledge might be correlated with SJT scores. However, the participants in our study needed some kind of knowledge of facts, laws, and procedures to give accurate responses in the webcam test, which is supported by the significant correlation we found between the job knowledge test and the webcam test. Another reason to examine the incremental validity of the webcam test over and above a declarative job knowledge test was, that most job knowledge tests used in selection research are measures of declarative job knowledge, not procedural job knowledge (e.g., Borman, White, Pulakos, & Oppler, 1991; Clevenger et al., 2001). There is no reason to interpret webcam tests differently than SJTs. Therefore, based on the assumption of Motowidlo et al. that SJTs are measures of procedural job knowledge, procedural job knowledge also may explain the criterion-related validity coefficients we found for of the webcam test.

The job knowledge test used in the present study consisted of a different set of items for each participant, which prevented circulation of items among participants. Thus, the job knowledge test was not exactly the same for each participant, but we would argue that the participants' scores were comparable to each other. As the job knowledge test was carefully constructed, we believe that the content validity of the test was substantial.

A concurrent design was used to determine the predictive validity of the webcam test. Our sample consisted of experienced consultants with previous knowledge of the job. It is possible that the results from the concurrent validation design used in this study might not be generalizable to applicant samples without prior job experience, because some previous knowledge of the job is needed to address the situations adequately. Yet, job tenure was not significantly related to scores on the webcam test. Furthermore, the motivation of the participants in the present study to perform well on the tests probably was as high as it would have been for applicants, suggesting that it would be unlikely to find a large difference in criterion-related validity if an applicant sample would have been used.

Practical implications and directions for future research

From an applied point of view, a drawback of the webcam test is its development cost. Scripts must be written for the scenarios, the scenarios have to be filmed with professional actors and the recordings have to be edited. Also the evaluation of the responses of each participant by three SME's caused the webcam test to be a relatively expensive selection instrument. Cost estimate per administration of this specific webcam test is approximately 250 euro. Therefore, future research is needed to determine whether the criterion-related validity of the webcam test is superior to that of less expensive selection instruments (e.g., structured behavioral interviews), and to that of the more conventional and documented SJT. Also, future research should examine whether the webcam test shows incremental validity with respect to general cognitive ability. Personnel selection procedures often include measures of cognitive ability due to its high validity for all jobs (F. L. Schmidt & Hunter, 1998). The high production costs of the webcam test may preclude the use of the test as selection instrument if it does not show incremental validity over and above cognitive ability. Past studies had demonstrated that SJTs are correlated with cognitive ability (e.g., Lievens & Sackett, 2006; McDaniel et al., 2007). On the other hand, multimedia SJTs show a lower cognitive component than written SJTs, because of the reading component of the latter type of test (Lievens & Sackett, 2006). Similar to multimedia SJTs, the webcam test does not have a reading component, and the open-ended format allows for a direct and spontaneous expression of a behavioral competency (Funke & Schuler, 1998). However, we still recommend future studies to investigate whether these aspects of the webcam test would form the factors responsible for a potential incremental validity over and above cognitive ability.

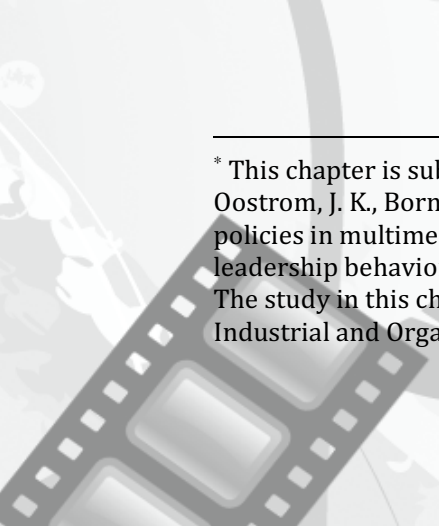
Finally, we recommend studying the acceptability and adverse impact of the webcam test, as these are important aspects of selection tests. Past studies have demonstrated that tests which are more interactive and behaviorally oriented result in more favorable applicant reactions than paper–and-pencil tests and cognitive ability tests (Lievens & Sackett, 2006; Schmitt & Chan, 1999) and generally have less adverse impact (Nguyen, McDaniel, & Whetzel, 2005).

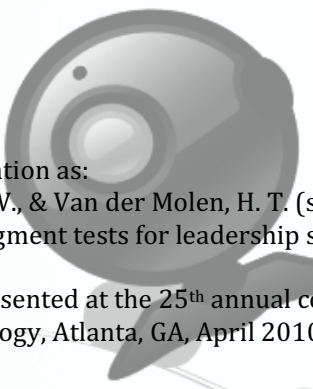
Based on the results of this first field study on the webcam test among employees, we believe that the webcam test is a valuable instrument for personnel selection, and

a promising alternative for traditional selection procedures. The next step is to verify and extend the present findings in an applicant setting using different kinds of predictors.

Chapter 4

Implicit trait policies in multimedia situational judgment tests for leadership skills: Can they predict leadership behavior?*





* This chapter is submitted for publication as:
Oostrom, J. K., Born, M. Ph., Serlie, A. W., & Van der Molen, H. T. (submitted). Implicit trait policies in multimedia situational judgment tests for leadership skills: Can they predict leadership behavior?
The study in this chapter was also presented at the 25th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA, April 2010.

Abstract

To explain why situational judgment tests are often correlated with measures of personality traits, Motowidlo, Hooper, and Jackson (2006b) developed the implicit trait policy theory. Implicit trait policies are inherent beliefs about causal relationships between personality traits and behavioral effectiveness. Among 180 employees, this field study on implicit trait policies examined whether a multimedia situational judgment test that was intended to assess leadership skills can capture individual differences in such policies for extraversion and agreeableness. In addition, it was examined whether these implicit trait policies for extraversion and agreeableness were able to predict leadership behavior. Results confirmed that the situational judgment test was able to capture individual differences in implicit trait policies with respect to extraversion and agreeableness. Furthermore, results showed that implicit trait policies for extraversion can predict leadership behavior over and above leadership experience and the associated personality trait. Implicit trait policies therefore seem a valuable predictor of job performance.

Introduction

Situational judgment tests (SJTs) are a frequently used selection tool, both in the United States and Europe (McDaniel et al., 2001; Salgado, Viswesvaran, & Ones, 2001). SJTs typically present job-related situations followed by a number of alternative response options. Applicants are then asked to evaluate the effectiveness of each response option or indicate the likelihood that they would respond in that way (Whetzel & McDaniel, 2009). Meta-analyses have demonstrated that SJTs have useful levels of validity as predictors of job performance (McDaniel et al., 2007; McDaniel et al., 2001) and that SJTs show substantial correlations with cognitive ability (McDaniel et al., 2001) and with Big Five personality dimensions (McDaniel & Nguyen, 2001). However, there is a paucity of theory regarding the predictive and construct validity of SJTs (Whetzel & McDaniel, 2009). To explain why SJTs are often correlated with measures of personality traits, Motowidlo et al. (2006b) developed the implicit trait policy (ITP) theory.

The ITP theory starts from the assumption that SJTs predict job performance because they measure procedural knowledge, which include a component of general domain knowledge about the costs and benefits of expressing particular personality traits in job-related situations. ITPs are the implicit beliefs of individuals about the effectiveness of different levels of trait expression. For instance, an individual may believe that the expression of agreeableness in SJT response options is generally very effective. When these ITPs are accurate, they represent an individual's general domain knowledge. ITPs are measured by correlating applicants' effectiveness ratings of SJT response options with the level of trait expression of these response options. The central proposition of the ITP theory is that individual differences in personality traits affect judgments of the effectiveness of SJT response options that express those personality traits. Motowidlo et al. (2006a; 2006b) indeed found empirical support for the ITP theory, as they were able to demonstrate that ITPs for agreeableness, conscientiousness, and extraversion are related to individual differences in these personality traits. Recently, Motowidlo and Beier (2010) demonstrated that ITPs are able to predict a composite measure of job performance. To measure ITPs, Motowidlo and Beier used an SJT specifically designed for management and administrative positions in telecommunications industry. Similarly to Motowidlo and Beier, the present study aims to shed light on the predictive validity of ITPs. However, in contrast with the study of Motowidlo and Beier and other studies on the predictive validity of ITPs that have used SJTs designed for specific jobs in specific companies, the present study will use a construct-driven multimedia SJT. A construct-driven SJT has several advantages, namely 1) that the validity of the SJT is expected to generalize across jobs and 2) that it provides the opportunity to conceptually match the predictor and criterion domain (Lievens, 2006).

Specifically, in the present study it will be examined whether a multimedia SJT for leadership skills is able to capture ITPs for targeted traits and whether these ITPs are able to predict leadership behaviors. First, ITP theory will be discussed in more detail followed by an overview of previous research on ITPs. Then, several hypotheses

about the relationships between ITPs, personality traits, leadership experience, and leadership behavior will be proposed.

ITP theory

ITP theory is embedded in social cognition research, which has shown that the judgment of trait-related behaviors of others is determined by the characteristics of the judge him- or herself (e.g., Heider, 1958; Lambert & Wedell, 1991; Markus, Smith, & Moreland, 1985; Prentice, 1990). ITP theory assumes that there are stable differences in individuals' implicit beliefs about the effectiveness of different levels of trait expression (Motowidlo et al., 2006a). An important determinant of how strong one weighs the expression of a particular trait is one's own standing on the trait (Motowidlo et al., 2006b). The reason for this is that individuals tend to believe that their own preferred way to handle a situation is the most effective way. Thus, individual differences in personality traits should affect their judgments of the effectiveness of SJT response options that express those personality traits (Motowidlo et al., 2006a, 2006b). For example, agreeable individuals judge very agreeable response options in an SJT as more effective than disagreeable individuals. Their ITPs for agreeableness would, therefore, be represented by a relatively strong positive correlation between their effectiveness ratings of the response options on the one hand and the degree to which the response options express agreeableness on the other hand.

If the ITP theory is correct, ITPs implicitly measure individual differences in personality traits. When individuals judge the effectiveness of SJT response options that vary in trait expression, they reveal something about their own standing on those traits. This may explain why SJT scores are often correlated with measures of personality traits (McDaniel et al., 2007; McDaniel et al., 2001; McDaniel & Nguyen, 2001). In a meta-analysis, McDaniel and Nguyen (2001) found the following mean observed correlations between SJT scores and Big Five personality traits: r = .25 for agreeableness, r = .26 for conscientiousness, r = .31 for emotional stability, r = .06 for extraversion, r = .09 for openness to experience. Motowidlo et al. (2006a) argued that correlations between SJTs and Big Five personality questionnaires do not have to be particularly strong to support the idea that personality traits have causal effects on ITPs, because personality traits as measured with Big Five personality questionnaires are distinct from their ITP counterparts. This distinctiveness is caused by the fact that a Big Five personality questionnaire is an explicit measure of personality while ITPs might be considered an implicit measure of personality. Implicit measures are less affected by social desirability (De Houwer, 2006), faking, and self-presentation biases (Bornstein, 2002). Therefore, implicit and explicit measures of the same construct are usually only modestly correlated to one another (e.g., Bornstein, 2002; De Houwer, 2006; Fazio & Olson, 2003).

Motowidlo et al. (2006b) argued that ITPs may also be affected by prior experience. Individuals develop implicit beliefs about effective ways to behave by experience in relevant situations and through a learning effect. Through experiences, individuals learn that the expression of certain personality traits is generally more effective than

the expression of other personality traits (Motowidlo & Beier, 2010). Individuals will develop ITPs accordingly, independently of their own standing on those traits. SJTs scores indeed have been found to be related to prior job experience (e.g., Clevenger et al., 2001; McDaniel & Nguyen, 2001; Weekley & Jones, 1999).

Previous research on ITPs

Motowidlo et al. (2006a; 2006b) conducted a number of studies among undergraduates that tested relationships between ITPs as measured with paper-and-pencil SJTs and associated personality traits as measured with the NEO-FFI. Significant correlations were found between ITPs for agreeableness, conscientiousness, and extraversion and the associated personality traits. In a study among 99 undergraduates, Motowidlo et al. (2006a) tested the hypothesis that ITPs for agreeableness and extraversion as measured with a paper-and-pencil SJT could predict behavioral expressions of these traits in a role play exercise. The results partly supported this hypothesis, as they found a significant correlation between ITPs for agreeableness and role play agreeableness scores, even when they partialed out the agreeableness scores on the NEO-FFI. However, ITPs for extraversion did not predict behavioral expressions of extraversion. The explanation of Motowidlo et al. (2006a) for this latter finding is that the specific facets of extraversion represented in the ITP measure, for example taking charge in social situations and standing up for one's own interest, might be different from the specific facets of extraversion expressed in the role play exercise, for example being enthusiastic, unreserved and talkative.

Recently, in a study among 115 employees, Motowidlo and Beier (2010) demonstrated that ITPs for agreeableness and conscientiousness were significantly related to a performance composite score build up by supervisor ratings of ten different workplace behaviors. To measure these ITPs they used an SJT that was designed to predict a number of competencies in managerial and administrative jobs in the telecommunications industry (e.g., leadership, flexibility, sensitivity, communication),

Among 71 undergraduates, D. Miller, Smith-Jentsch, and Afek (2008) examined the relationships between ITPs for agreeableness and conscientiousness, the personality scale scores of agreeableness and conscientiousness, job experience, and peer ratings of typical agreeableness and conscientiousness. To measure ITPs, D. Miller et al. (2008) used an SJT that was designed for a state welfare to work readiness program. Significant correlations were found between ITPs and the associated personality scale scores. Furthermore, it was found that ITPs for agreeableness explained unique variance in peer ratings of agreeableness over and above the agreeableness scores on the personality questionnaire. D. Miller et al. also found support for the hypothesis that ITPs for agreeableness are affected by undergraduate's prior customer service experience.

Present Study

Previous studies on ITPs have shown that it is possible to use SJTs to assess individual differences in ITPs and that ITPs can predict peer trait ratings, behavioral

expressions in a role play exercise and a composite measure of job performance (D. Miller et al., 2008; Motowidlo & Beier, 2010; Motowidlo et al., 2006a, 2006b). In the present study, we sought to both replicate and extend these findings by examining the relationships between ITPs as measured with a multimedia SJT for leadership skills, the associated personality scale scores, leadership experience, and observed leadership behavior. As far as we know no studies have actually examined the relationship between ITPs as measured with a construct-driven SJT (in our case leadership skills) and job behavior in the relevant domain (in our case leadership behavior). Furthermore, the present study will examine whether ITPs can be captured by a multimedia SJT. In a multimedia SJT the situations and response options are presented through the use of video clips. Multimedia SJTs therefore create a richer and more realistic assessment environment, as they present voice intonations, facial expressions, and other nonverbal behaviors that would also be displayed in actual job situations (Olson-Buchanan & Drasgow, 2006).

First, it will be examined whether a multimedia SJT for leadership skills is able to capture individual differences in ITPs by examining the relationship between ITPs and the associated personality scale scores. We believe ITPs to be an important determinant of participants' effectiveness ratings of SJT response options representing leadership behaviors, as current leadership research has emphasized the role of implicit theories in leadership perceptions (e.g., Epitropaki & Martin, 2004; Lord, De Vader, & Alliger, 1986; Lord, Foti, & De Vader, 1984). These implicit leadership theories represent cognitive schemas of traits and behaviors that followers expect from leaders, such as being enthusiastic and supportive (Epitropaki & Martin, 2004), and have been found to be affected by an individual's own personality traits (e.g., Keller, 1999; Lord et al., 1986). SJT response options were constructed to express either high or low levels of the two personality traits that are the most related to interpersonal interactions, namely extraversion and agreeableness (McCrae & Costa, 1989; Wiggins, 1995). The first hypothesis therefore is:

Hypothesis 1: ITPs for extraversion (H1a) and ITPs for agreeableness (H1b) as measured with a multimedia SJT for leadership skills will be positively related to personality scale scores of extraversion and agreeableness respectively.

According to ITP theory, two major factors have a causal impact on ITPs, namely personality and experience (Motowidlo et al., 2006a). Through experience, individuals will learn that certain personality traits are generally more effective than other personality traits, regardless of their own standing on those traits. D. Miller et al. (2008) demonstrated that ITPs for agreeableness are indeed affected by prior job experience (r = .21). According to Motowidlo and Beier (2010), people can acquire general knowledge about trait effectiveness through general life experiences. However, specific domain knowledge can be learned only through job experience in that particular domain. As the multimedia SJT in the current study measures knowledge

about trait effectiveness in leadership behaviors only specific leadership experience is expected to influence participants' ITPs. Therefore, the second hypothesis is:

Hypothesis 2: Leadership experience will be positively related to ITPs for extraversion (H2a) and ITPs for agreeableness (H2b) as measured with a multimedia SJT for leadership skills.

Furthermore, the predictive validity of ITPs for extraversion and agreeableness as measured with a multimedia SJT of leadership skills will be examined. J. Hogan and Holland (2003) emphasized the need to align predictor and criterion domains by using the same underlying construct, as this will enhance the validity of the predictor. Given that the multimedia SJT intends to measure leadership skills, it should only measure participants' ITPs about the effectiveness of extraversion and agreeableness in leadership behaviors. To clearly examine the predictive validity of ITPs as measured with the leadership SJT, one should therefore use a criterion that measures participants' leadership behavior. In the present study, a differentiated criterion measurement was used deduced from the competency framework of Bartram (2005b). In this framework, competencies are defined as observable workplace behaviors. To obtain a complete picture of participants' observable workplace behaviors, multiple raters were included in the study. Specifically, peer ratings and supervisor ratings on one aligned competency (leading and deciding) and two non-aligned competencies (supporting and cooperating and analyzing and interpreting) were used. Along these lines, the following hypothesis was formulated:

Hypothesis 3: ITPs for extraversion (H3a) and ITPs for agreeableness (H3b) as measured with a multimedia SJT for leadership skills will be more strongly related to leadership behaviors than to non-leadership behaviors observed in the workplace.

Both personality traits, including extraversion and agreeableness, and leadership experience have been found to be positively related to leadership behavior (e.g., Judge, Bono, Ilies, & Gerhardt, 2002; Thomas & Cheese, 2005). For example, Judge et al. (2002) have meta-analytically demonstrated that leadership is significantly related to the personality traits extraversion (r = .22) and agreeableness (r = .06). Therefore, it is interesting to investigate whether the relationship between ITPs and observed leadership behavior can be solely attributed to the causal effects of personality and leadership experience on ITPs, or whether ITPs explain unique variance in observed leadership behavior beyond the variance explained by personality traits and leadership experience. To our knowledge, this study is the first to examine whether ITPs have incremental validity over and above personality and prior job experience in predicting job performance ratings. Based on the findings of Motowidlo et al. (2006a) and D. Miller et al. (2008) that ITPs for agreeableness explain a significant part of

variance in role play agreeableness and peer ratings of agreeableness beyond the variance explained by explicitly measured agreeableness, it can be hypothesized that:

Hypothesis 4: ITPs for extraversion (H4a) and ITPs for agreeableness (H4b) as measured with a multimedia SJT for leadership skills will explain a significant part of variance in observed leadership behavior beyond the variance explained by personality scale scores of extraversion and agreeableness and leadership experience.

Method

Participants and procedure

This study was conducted among assessment candidates of GITP, a large HRD consultancy firm in the Netherlands. With the invitation for an assessment, an information brochure and an invitation to participate in the study were sent to all candidates of GITP in 2008 and 2009. Subsequently in total, 450 candidates registered themselves voluntarily to participate. Next, they received an e-mail invitation to complete a multimedia SJT and a job performance scale. The response rate was 40.0% (180 participants). The age of the participants varied between 22 and 57 (M = 38.8, SD = 8.44). One hundred and ten participants were male (61.1%) and 70 participants were female (38.9%). Educational levels ranged from high school to master's degree. A large part of participants worked in commercial (34.7%) or social (9.7%) sectors. To check whether participation was in any way selective, we compared the age, gender, educational level, and assessment outcome of participants to all candidates of 2008 and 2009 (N = 13701). None of these comparisons yielded significant results.

Measures

Personality questionnaire. As part of their assessment program at GITP, participants completed a 224-item personality questionnaire (Koch, 1998), based on the Five Factor Model of personality (Goldberg, 1990). Participants had to provide their answers to the items on a five-point scale ranging from 1 = *strongly disagree* to 5 = *strongly agree.* All participants completed the questionnaire within 25 minutes. We only used participants' scores on the extraversion (27 items) and the agreeableness (28 items) scale. An example of an item of the extraversion scale is as follows: "*Rate yourself on the following statement: Enjoys meeting new people*". In a study among 261 GITP candidates (GITP, 2010), the extraversion and the agreeableness scale of the personality questionnaire were found to correlate with the extraversion (r = .70) and the agreeableness (r = .49) scale of the revised NEO-Personality Inventory (Costa & McCrae, 1992). Coefficient alphas were substantial: α = .92 for extraversion and α = .85 for agreeableness.

Multimedia SJT. The multimedia SJT was designed to predict leadership skills. The SJT consisted of 17 short videotaped vignettes of interpersonal situations that leaders are likely to encounter on the job. The vignettes were introduced by a narrator. Each

situation froze at an important point and four possible ways for a leader to handle the situation were presented. Participants were asked to judge the four response options on a five-point scale ranging from (--) *very ineffective* to (++) *very effective*. Participants were instructed to complete the multimedia SJT in a calm environment at home or at work. All participants completed the multimedia SJT within 45 minutes.

To check whether the multimedia SJT indeed measured procedural knowledge in leadership situations, four experts on situational judgment testing (2 female, 2 male) independently categorized the vignettes of the multimedia SJT. The experts worked in the field of personnel selection and had experience with constructing SJTs. Their age ranged from 30 to 57 (M = 39.3, SD = 12.74). The categorization of the vignettes was based on the classification of leadership behaviors provided by O*Net, that consisted of five categories, namely 1) making decisions and solving problems, 2) coordinating the work and activities of others, 3) guiding, directing, and motivating others, 4) developing and building teams, and 5) resolving conflicts and negotiating with others. The experts were asked to indicate to which category each vignette belonged by writing down the category number (1 to 5) after each vignette. When they believed that a vignette did not belong to one of the categories of leadership behaviors they were instructed to write down a 6 after the vignette. One expert indicated that two of the vignettes did not belong to one of the categories of leadership behaviors, the other three experts indicated that each vignette belonged to one of the categories of leadership behaviors. The one-way random effects intra-class correlation (*ICC*) for absolute agreement was .86, indicating that there was substantial agreement in the categorization of the vignettes among the experts. The experts indicated that two vignettes belonged to the category of making decisions and solving problems, two vignettes to coordinating the work and activities of others, nine vignettes to guiding, directing, and motivating others, two vignettes to developing and building teams, and two vignettes belonged to the category of resolving problems and negotiating with others.

The response options were based on the Interpersonal Adjective Scale of Wiggins (1995), and therefore were intended to express low or high levels of extraversion and agreeableness. We computed the participants' ITPs for extraversion and agreeableness by calculating the correlation between the participants' effectiveness ratings of the 68 response options in the SJT and the a priori intended level of the trait (coded as 0 = *low* and 1 = *high*). Fisher's z-transformation was used to normalize the correlation coefficients. To check whether the response options indeed expressed the intended low or high levels of extraversion and agreeableness, four subject matter experts (2 male, 2 female) who worked as a consultant at GITP, rated the SJT response options according to the level of extraversion and agreeableness each expressed. They used a seven-point scale ranging from 1 = *very introverted / very disagreeable* to 7 = *very extraverted / very agreeable*. Mean trait ratings were computed for each response option. The correlation between the mean trait rating of the four subject matter experts and the a priori trait level was .52 (p < .01) for extraversion and .84 (p < .01) for agreeableness. There was substantial agreement between the

experts about the level of extraversion and agreeableness the response options expressed (*ICC* = .66 for extraversion and *ICC* = .90 for agreeableness). In total, 20 response options expressed both a high level of extraversion and a high level of agreeableness, 19 response options expressed a high level of extraversion and a low level of agreeableness, 16 response options expressed a low level of extraversion and a high level of agreeableness, and 12 response options expressed both a low level of extraversion and a low level of agreeableness.

Leadership experience. This variable was measured with the following item: 'How many years of leadership experience do you have?'. Participants indicated their experience on a five-point scale ranging from 1 = *no experience* to 5 = *more than 10 years*.

Criterion measure. A differentiated criterion measurement was used based on the Great Eight competency framework of Bartram (2005b). Three competencies were chosen from this competency framework. One competency was aligned with the predictor, namely the competency leading and deciding. Two competencies were non-aligned. The competency supporting and cooperating was chosen as a first non-aligned competency, relating to contextual behaviors. The competency analyzing and interpreting was chosen as the second non-aligned competency, relating to task behaviors. The competency leading and deciding includes behaviors such as taking control and exercising leadership, initiating actions, giving directions and taking responsibilities. This competency was measured with 10 items. An example of an item is: "*Knows how to motivate employees to achieve their goals*". The competency supporting and cooperating includes behaviors such as showing respect and positive regard for others in social situations, and working effectively with individuals and teams, clients and staff. This competency was measured with 10 items. An example of an item is: "*Shows respect for others*". The competency analyzing and interpreting includes behaviors such as showing evidence of clear and analytical thinking, applying one's expertise effectively, quickly taking on new technology, and communicating well in writing. This competency was measured with 15 items. An example of an item is: "*Applies new techniques and procedures effectively*". All items of the criterion measurement had to be rated on a five-point scale ranging from 1 = *strongly disagree* to 5 = *strongly agree*

Participants were instructed to ask at least one individual in their direct work environment (preferably their supervisor) to evaluate their job behaviors by filling out the questionnaire. In total, 97 peer ratings and 71 supervisor ratings were obtained. Coefficient alphas were substantial, varying from .76 for the peer ratings of supporting and cooperating to .89 for the peer ratings of analyzing and interpreting.

Results

Preliminary analyses

Means, standard deviations, scale reliabilities and correlations between all variables are presented in Table 1. Before testing the hypotheses, we first looked at significant correlations between demographic characteristics and leadership experience, scores on the personality scales, ITPs, and the criterion measures. Gender was significantly related to leadership experience (r = -.21, p < .01) and to supervisor ratings of supporting and cooperating (r = -.33, p < .01). Males (M = 3.12, SD = 1.45) had more leadership experience than females (M = 2.52, SD = 1.31, t = 2.75, p < .01) and supervisors indicated that males (M = 3.88, SD = 0.42) showed more supporting and cooperating behavior than females (M = 3.57, SD = 0.49, t = 2.86, p < .01). Age was significantly related to leadership experience (r = .42, p < .01) and to peer ratings of all three workplace behaviors (r between .22, p < .05 and .33, p < .01). Because of these significant correlations, gender and age were controlled for in the regression analyses. Similar to the findings of Motowidlo et al. (2006a), a substantial significant negative correlation was found between ITPs for extraversion and ITPs for agreeableness (r = -.47, p < .01).

As reported in previous studies (e.g., Harris & Schaubroeck, 1998; Mount, 1984), peer ratings and supervisor ratings were modestly related (r = .49, p < .01 for leading and deciding and r = .32, p < .05 for supporting and cooperating), except for analyzing and interpreting (r = .22, *ns*). Mean peer ratings were significantly higher than mean supervisor ratings (t = 2.62, p < .05). Previous studies have found similar mean peer-supervisor rating differences (e.g., Thornton, 1980; Harris & Schaubroeck, 1998).

Hypotheses testing

Table 1 shows that ITPs for extraversion and ITPs for agreeableness positively correlated with the personality scale scores of extraversion and agreeableness (r = .20, p < .05 and r = .21, p < .01 respectively). This lends support for Hypotheses 1a and 1b, which stated that ITPs for extraversion and agreeableness as measured with a multimedia SJT for leadership skills would be positively related to the personality scale scores of extraversion and agreeableness respectively.

Hypothesis 2a, stating that leadership experience would be positively related to ITPs for extraversion, was supported. Leadership experience and ITPs for extraversion correlated significantly (r = .18, p < .05). However, Hypothesis 2b, stating that leadership experience would be positively related to ITPs for agreeableness, was not supported by the data. Leadership experience and ITPs for agreeableness were not significantly correlated (r = -.06, *ns*).

Regression analyses showed that, taken together, leadership experience (β = .15, p < .05) and extraversion (β = .15, p < .05) explained 5% of the variance (F = 4.82, p < .01) in ITPs for extraversion. Leadership experience and agreeableness (β = .20, p < .01) explained 4% of the variance (F = 3.95, p < .05) in ITPs for agreeableness. However, in the prediction of ITPs for agreeableness no significant beta weight was found for leadership experience (β = -.09, *ns*).

Table 1

Means, Standard Deviations, Scale Reliabilities, and Correlations between Personality Traits, ITPs, and Criterion Measures

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Gender	.39	.49	(-)												
2. Age	38.83	8.44	-.08	(-)											
3. Leadership experience	2.89	1.42	-.21**	.42**	(-)										
4. Personality trait - Extraversion	3.67	0.51	.03	.03	.18*	(.92)									
5. Personality trait - Agreeableness	3.91	0.32	-.02	.07	.17*	.25**	(.85)								
6. ITPs Extraversion	.39	.13	-.15	.02	.18*	.20*	.12	(-)							
7. ITPs Agreeableness	.72	.19	.09	-.06	-.06	.13	.21**	-.47**	(-)						
8. Peer ratings leading and deciding	3.63	0.57	-.07	.23**	.38**	.13	.09	.36**	-.09	(.87)					
9. Supervisor ratings leading and deciding	3.41	0.52	-.07	.04	.20	.10	.07	.31**	-.10	.49**	(.83)				
10. Peer ratings supporting and cooperating	3.82	0.44	-.09	.33**	.06	.03	.20*	.17	.09	.53**	.11	(.76)			
11. Supervisor ratings supporting and cooperating	3.75	0.47	-.33*	.08	.01	.01	.06	.03	.08	.27*	.53**	.32*	(.82)		
12. Peer ratings analyzing and interpreting	3.74	0.48	.07	.22*	-.01	-.04	-.07	.14	.05	.57**	.12	.70**	.18	(.89)	
13. Supervisor ratings analyzing and interpreting	3.59	0.43	-.08	.04	-.11	-.02	-.04	.07	.08	.25	.58**	.12	.61**	.22	(.85)

Note. Scale reliabilities are presented on the diagonal. Personality scales and ratings of workplace behaviors were measured on a 5-point scale. Gender (0 = male, 1 = female) and leadership experience (0 = no experience, 1 = less than one year, 2 = one to five years, 3 = 6-10 years, and 4 = more than 10 years) were coded. $N = 180$ for demographic characteristics, personality scales, and ITPs. $N = 97$ for peer ratings of workplace behavior. $N = 71$ for supervisor ratings of workplace behavior.

$p < .05$, ** $p < .01$

To test Hypothesis 3a, which stated that ITPs for extraversion would be more strongly related to leadership behaviors than to non-leadership behaviors observed in the workplace, Steiger's *z* statistic was used. Results supported this hypothesis. As shown in Table 1, ITPs for extraversion significantly correlated with peer ratings (r = .36, p < .01) and supervisor ratings (r = .31, p < .01) of leading and deciding. As predicted, ITPs for extraversion were not significantly related to any ratings of supporting and cooperating or to any ratings of analyzing and interpreting. The correlation between ITPs for extraversion and peer ratings of leading and deciding was significantly higher than the correlation between ITPs for extraversion and peer ratings of supporting and cooperating (z = 1.98, p < .05) and than the correlation between ITPs for extraversion and peer ratings of analyzing and interpreting (z = 2.39, p < .01). The correlation between ITPs for extraversion and supervisor ratings of leading and deciding also was significantly higher than the correlation between ITPs for extraversion and supervisor ratings of supporting and cooperating (z = 2.43, p < .01) and than the correlation between ITPs for extraversion and supervisor ratings of analyzing and interpreting (z = 2.21, p < .05).

Table 2

Incremental Validity of ITPs for Extraversion in Predicting Leading and Deciding

	Peer ratings (N = 97)		Supervisor ratings (N = 71)	
	β	ΔR^2	β	ΔR^2
Step 1				
Gender	.06		-.05	
Age	.11	.05	-.05	.01
Step 2				
Leadership experience	.27*	.09**	.10	.04
Step 3 - Personality trait				
Extraversion	.01	.00	.05	.00
Step 4 - ITPs				
ITPs Extraversion	.30**	.09**	.28*	.07*
R^2	.23		.12	
F	5.45**		1.83*	

Note. Standardized regression weights are for final step. Gender (0 = male, 1 = female) and leadership experience (0 = no experience, 1 = less than one year, 2 = one to five years, 3 = 6-10 years, and 4 = more than 10 years) were coded.

* p < .05, ** p < .01

Regarding agreeableness, Table 1 shows that ITPs for agreeableness were not significantly correlated with peer ratings of leading and deciding (r = -.09, *ns*), nor with supervisor ratings of leading and deciding (r = -.10, *ns*). These correlations were

not significantly higher than the correlations between ITPs for agreeableness and ratings of the other observed workplace behaviors (*z* varied from -.1.79, *ns* to -1.24, *ns*). Thus, ITPs for agreeableness were not more strongly related to observed leadership behaviors than to other observed workplace behaviors. Based on these findings, Hypothesis 3b could not be supported.

A series of stepwise hierarchical regression analyses were conducted to test Hypothesis 4, stating that ITPs for extraversion and ITPs for agreeableness as measured with a multimedia SJT for leadership skills would explain a significant part of variance in leadership behavior, beyond the variance explained by personality scale scores of extraversion and agreeableness and leadership experience. Gender and age were entered in the first step, leadership experience in the second step, extraversion or agreeableness as measured with the personality questionnaire in the third step, and ITPs for extraversion or ITPs for agreeableness were entered in the final step. Tables 2 and 3 present the results for the ITPs for extraversion and the ITPs for agreeableness respectively.

Table 3

Incremental Validity of ITPs for Agreeableness in Predicting Leading and Deciding

	Peer ratings (*N* = 97)		Supervisor ratings (*N* = 71)	
	β	ΔR^2	β	ΔR^2
Step 1				
Gender	.06		-.04	
Age	.09	.05	-.08	.01
Step 2				
Leadership experience	.33**	.09**	.20	.04
Step 3 - Personality trait				
Agreeableness	.05	.00	.04	.00
Step 4 - ITPs				
ITPs Agreeableness	-.08	.01	-.09	.01
R^2	.15		.05	
F	3.19**		0.70	

Note. Standardized regression weights are for final step. Gender (0 = male, 1 = female) and leadership experience (0 = no experience, 1 = less than one year, 2 = one to five years, 3 = 6-10 years, and 4 = more than 10 years) were coded.

* $p < .05$, ** $p < .01$

ITPs for extraversion were able to explain additional variance in peer ratings of leading and deciding (β = .30, $p < .01$, $\Delta R^2 = .09$, $p < .01$) and in supervisor ratings of leading and deciding (β = .28, $p < .05$, $\Delta R^2 = .07$, $p < .05$) over and above leadership experience and the personality trait extraversion. Taken together, the control

variables and the predictors explained 23% of the variance (F = 5.45, p < .01) in peer ratings of leading and deciding and 12% of the variance (F = 1.83, p < .05) in supervisor ratings of leading and deciding. These results as a whole support Hypothesis 4a, as ITPs for extraversion explained a significant part of variance in peer ratings and supervisor ratings of leading and deciding beyond the variance explained by the personality scale scores and by leadership experience.

ITPs for agreeableness were not able to explain additional variance in peer ratings of leading and deciding (β = -.08, *ns*, ΔR^2 = .01, *ns*) and supervisor ratings of leading and deciding (β = -.09, *ns*, ΔR^2 = .01, *ns*) over and above leadership experience and the personality trait agreeableness. The control variables and the predictors did explain a significant part of the variance in peer ratings of leading and deciding (R^2 = .15, F = 3.19, p < .01). Regarding peer ratings, a significant beta weight was found for leadership experience (β = .33, p < .01). Based on these findings, Hypothesis 4b could not be supported.

Discussion

The aim of the present study was to examine the relationships between ITPs as measured with a multimedia SJT for leadership skills, personality scale scores, leadership experience, and leadership behavior. The present study examined whether ITPs as measured with a construct-driven multimedia SJT can predict job behavior in the relevant domain. Furthermore, the present study examined whether ITPs have incremental validity over and above personality and prior job experience in predicting job behavior in the relevant domain. Results confirmed that a multimedia SJT for leadership skills can be used to measure individual differences in ITPs and that ITPs for extraversion are able to predict leadership behavior over and above leadership experience and personality. Each of the findings will be discussed below.

The results demonstrated that a multimedia SJT for leadership skills indeed can be used to capture ITPs for targeted personality traits. The first hypothesis, which stated that ITPs for extraversion and agreeableness would be positively related to the personality scale scores of extraversion and agreeableness respectively, was supported. The present results are in line with Motowidlo et al. (2006a; 2006b) and D. Miller et al. (2008) who showed that it is possible to use SJTs to capture individual differences in ITPs for agreeableness, conscientiousness, and extraversion. These findings may have important implications, as researchers have long sought personality measures that are less affected by social desirability, faking, and self-presentation biases than explicit personality questionnaires (Fazio & Olson, 2003; Vaillant, 1998; Frost et al., 2007).

Support was also found for the hypothesis that ITPs are affected by job experience. Our results demonstrated that leadership experience explained a significant part of variance in ITPs for extraversion. In other words, employees who had more experience as a leader held stronger beliefs about the effectiveness of extraversion in the leadership behaviors that were demonstrated in the SJT. D. Miller et al. (2008) already demonstrated that prior customer service experience was related to ITPs for

agreeableness. These findings demonstrate that individuals develop implicit beliefs about effective ways to behave on the job by experience in job relevant situations.

As the multimedia SJT was developed to predict leadership skills, it was argued that it should particularly measure participants' ITPs about the effectiveness of extraversion and agreeableness in leadership behavior. The results demonstrated that ITPs for extraversion can predict both peer ratings and supervisor ratings of leadership behavior, and that they indeed showed more validity for predicting leadership behavior than for other non-leadership workplace behavior. Thus, ITPs for extraversion indeed showed good convergent and discriminant validity. These results demonstrate the importance of aligning predictor and criterion domains. Because of the good convergent and discriminant validity found in this study, we recommend future studies on ITPs to also carefully conceptually match the predictor and criterion domain.

Furthermore, our results demonstrated that ITPs for extraversion explained unique variance in peer ratings of leadership behavior and supervisor ratings of leadership behavior over and above leadership experience and personality scale score. Thus, the relationship between ITPs for extraversion and leadership behavior is not attributable solely to the causal effects of personality and leadership experience on ITPs. Although peers and supervisors differ in their opportunity to observe participants' leadership behavior, both ratings were predicted by participants' ITPs for extraversion. This finding emphasizes the important role of ITPs for extraversion in leadership behavior.

In sum, our results confirm that our SJT captures individual differences in ITPs with respect to extraversion and agreeableness, that leadership experience affects ITPs for extraversion, and that ITPs for extraversion are predictive of leadership behavior over and above leadership experience and the associated personality scale score. However, the results were not unequivocal, as expectations regarding the relationship between ITPs for agreeableness and leadership experience and leadership behavior were not supported. This is in contrast with previous studies on ITPs (D. Miller et al., 2008; Motowidlo & Beier, 2010; Motowidlo et al., 2006a, 2006b). For example, Motowidlo and Beier (2010) did find a significant positive correlations between ITPs for agreeableness and performance. Our findings are most probably due to a less clear relationship between agreeableness and the criterion measure, namely leadership behaviors. In their meta-analyses on personality and leadership, Judge et al. (2002) found that extraversion emerged as the most important trait of effective leaders, followed by conscientiousness and openness to experience. Agreeableness was also related to leadership behavior, but seemed less relevant than other personality traits. On the one hand altruism and interpersonal sensitivity seem important traits for leaders. On the other hand, agreeable persons are likely to be more modest (Goldberg, 1990) and have more need for affiliation (Yukl, 1998). These facets of agreeableness are negatively related to leadership. It is possible that because of this multifaceted relationship between agreeableness and leadership, ITPs for

agreeableness were not related to leadership experience and ratings of leadership behavior.

A remarkable finding was the strong negative correlation between ITPs for extraversion and ITPs for agreeableness (r = -.47, p < .01). The correlational scores for the ITPs were computed in the same way as in the study of Motowidlo et al. (2006a), who reported correlations between ITPs for extraversion and agreeableness ranging from r = -.11 to r = -.30. As each response option in the SJT varied in both the level of extraversion and the level of agreeableness it expressed, the ITPs for the two traits were not measured independently. There were slightly more response options in which there was an incongruent level of trait expression, that is a high level expression of one trait and a low level expression of the other trait. When participants judge such a response option as very effective, this would positively affect their ITPs for the trait that is expressed in a high level and at the same negatively affect their ITPs for the trait that is expressed in a low level. This could explain the strong negative correlations between ITPs for extraversion and ITPs for agreeableness.

Limitations of this study and suggestions for future research

There are a number of limitations to the study that should be noted. The first limitation involves the subjective performance rating. Although, assessments of workplace behaviors most commonly consist of ratings made by the participants' supervisor, peer, or subordinate, these ratings are potentially biased by selective recall or halo effects. Hogan, Curphy, and Hogan (1994) therefore argued that leadership behavior should be objectively measured in terms of team, group, or organizational effectiveness. Despite this obvious limitation, we believe that there are also several strengths concerning the performance ratings used in the present study, such as the inclusion of multiple raters and the fact that the peer ratings and supervisor ratings support each other with respect to the validity of ITPs for extraversion. Nevertheless, we advise future research to include objectively measured performance outcomes.

The second limitation involves the voluntary nature of study participation. Comparison of age, gender, educational level, and assessment outcome of candidates that actually participated in the study to all other candidates yielded no significant differences, but it remains plausible that motivational difference between the participants and other candidates exists. These motivational differences could have affected performance on the predictors but also on the criterion measure. The employees that participated were actively seeking feedback on their skills and competencies. As in most studies in which performance ratings from multiple raters are obtained (e.g., Murphy & Shiarella, 1997; Scullen, Mount, & Goff, 2000), participants in this study selected the peers and supervisors who rated them. We do not know how these raters perceived the feedback-seeking behavior of the participants, nor do we know on which basis the peers and supervisors were selected. These issues appear to be worthwhile topics for future research.

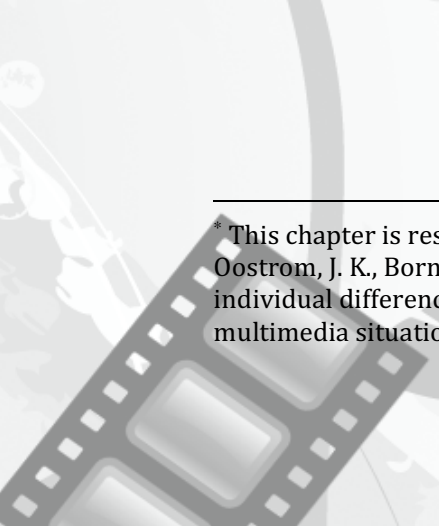
An important question is whether the present results are generalizable to other SJTs. The present SJT had a multimedia format, which creates a more rich and realistic assessment environment than a paper-and-pencil SJT (Olson-Buchanan & Drasgow, 2006). Furthermore, the SJT in the present study was intended to assess leadership skills. Most SJTs are intended to assess interpersonally oriented constructs, such as communication skills or leadership skills (Weekley & Ployhart, 2006). These types of SJTs therefore might only be able to capture individual difference in ITPs for extraversion and agreeableness, as these traits are most related to interpersonal interactions. Future studies should therefore examine to which degree the type of SJT, in terms of format and construct, influences the measurement of ITPs and their predictive validity.

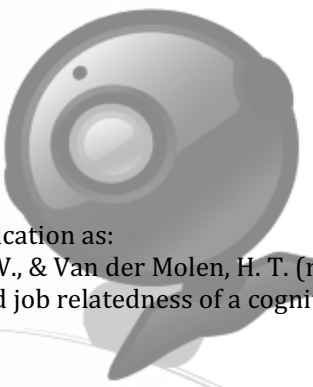
Conclusion

The present study is the first field study that tested the relationship between ITPs as measured with a construct-driven multimedia SJT and job behavior related to that construct. Therefore, we believe our study has made a valuable contribution to the ITP theory of Motowidlo et al. (2006a; 2006b). The present study confirms that multimedia SJTs for leadership can be used to capture individual differences in ITPs. It was also demonstrated that ITPs as measured with a multimedia SJT for leadership are predictive of both peer ratings and supervisor ratings of leadership behavior over and above leadership experience and personality scale scores. ITPs, as implicit measure of personality traits, therefore seem a valuable predictor of job performance, as they are less affected by social desirability and self-presentation biases than explicit personality measures.

Chapter 5

The role of individual differences in the perceived job relatedness of a cognitive ability test and a multimedia situational judgment test*





* This chapter is resubmitted for publication as:
Oostrom, J. K., Born, M. Ph., Serlie, A. W., & Van der Molen, H. T. (resubmitted). The role of individual differences in the perceived job relatedness of a cognitive ability test and a multimedia situational judgment test.

Abstract

Although there is a growing number of publications concerning applicant reactions to different selection instruments, the relationships between individual differences and applicant reactions have largely remained unexplored. The aim of the present study was to examine the effects of several testing-related and general individual differences (anxiety, self-evaluations, and personality) on the most commonly studied dimension of applicant reactions, namely the perceived job relatedness of selection instruments. Participants were 153 psychology students, who completed a cognitive ability test and a multimedia SJT as part of their educational program. Our results indicated that computer anxiety negatively affected perceived job relatedness and core self-evaluations, subjective well-being, agreeableness, emotional stability, and openness to experience positively affected perceived job relatedness. Openness to experience was the most consistent predictor of perceived job relatedness. The results of our study suggest that certain individuals may be more predisposed to react positively to selection instruments. Therefore, we concluded that the nature of the applicant pool should be carefully considered when designing interventions to improve applicant reactions.

Introduction

There has been a vast amount of research on the validity and utility of selection instruments that have demonstrated how an organization can benefit from using valid selection instruments (e.g., Barrick & Mount, 1991; McDaniel et al., 2007; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Salgado et al., 2001; F. L. Schmidt & Hunter, 1998). As a result, researchers have started to develop an interest in examining personnel selection from the applicant's perspective (e.g., Anderson, 2003; Ryan & Ployhart, 2000; Rynes & Connerley, 1993; Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993). Measuring how applicants react to selection instruments has been found to be not only relevant for applicants themselves, but also for the organization. Previous studies have demonstrated that applicant reactions are related to intentions to accept the job, intentions to recommend the organization to others, the likelihood of litigation against the outcome of the selection procedure, and perceived organizational attractiveness (Anderson et al., 2004; Chan & Schmitt, 2005; Gilliland, 1993; Ryan & Ployhart, 2000).

Much of the research on applicant reactions has focused on descriptive questions, such as the comparison of favorability reactions across procedures and instruments (e.g., Hausknecht, Day, & Thomas, 2004; Kluger & Rothstein, 1993; Rynes & Connerley, 1993). However, theory is lacking on why applicants evaluate different selection instruments in a different manner (Anderson, 2003). Recent research has, therefore, moved beyond the comparison of applicant reactions across different instruments to the assessment of how test-related factors, such as test content or test method, affect those reactions (e.g., Bauer, Truxillo, Paronto, Weekley, & Campion, 2004; Chan & Schmitt, 1997; Kanning et al., 2006). For example, Chan and Schmitt (1997) demonstrated that the way in which a situational judgment test (SJT) is administered affects its face validity. Participants rated the face validity of a video-based SJT significantly more positive than the face validity of a paper-and-pencil SJT. Yet, one domain of antecedents has remained largely unexplored, namely individual differences between applicants. Differences in test anxiety, computer anxiety or openness to experience are likely to influence applicant reactions, yet have only been included in a few studies (Bernerth, Feild, Giles, & Cole, 2006; Ryan, Greguras, & Ployhart, 1996; Wiechmann & Ryan, 2003).

The aim of the present study is to examine the relationship of a number of testing-related and general individual differences with the most frequently studied dimension of applicant reactions, namely perceived job relatedness (Chan & Schmitt, 2004). Gilliland (1993) defined job relatedness as the extent to which a test appears to measure content relevant to the job (face validity) and at the same time appears to be predictively valid (perceived predictive validity). Smither et al. (1993) provide evidence that these aspects are two related, but distinguishable, dimensions of job relatedness. However, in most studies job relatedness, face validity, and perceived predictive validity are used as interchangeable terms. Because personnel selection instruments are increasingly administered via computers (e.g., Lievens et al., 2002), we examined the effects of individual differences on the perceived job relatedness of

two often used computer-based selection instruments, namely a cognitive ability test and a multimedia situational judgment test (SJT) intended to measure managerial skills.

The perceived job relatedness of cognitive ability tests and multimedia SJTs

The perceived job relatedness of selection instruments has been found to influence several valued organizational outcomes. Bauer, Maertz, Dolen, and Campion (1998) examined the effects of five justice dimensions (information known about the test, chance to perform, treatment at the test site, consistency of the test administration, and job relatedness) on organizational attractiveness, intentions to accept a position, intentions to encourage others to apply, perceptions of testing fairness, and test-taking self-efficacy. Of these justice dimensions, job relatedness appeared to be the most consistently and significantly related to the organizational outcomes. Furthermore, researchers have argued that low job relatedness may result in biased or inaccurate test scores, and therefore reduces the operational validity of selection instruments (e.g., Cascio, 1987; Robertson & Kandola, 1982; Smither et al., 1993).

Some selection instruments are perceived as more job-related than others. In general, applicants perceive work samples or other high fidelity assessments to be more job-related than cognitive ability tests (Hausknecht et al., 2004; Macan, Avedon, Pease, & Smith, 1994; Ployhart & Ryan, 1998; Rynes & Connerley, 1993; Smither et al., 1993). Hausknecht et al. (2004) meta-analytically demonstrated that selection instruments with a transparent relationship with job tasks, such as interviews or works samples, are perceived as more favorable than selection instruments with a less transparent relationship with job tasks, such as cognitive ability tests and personality questionnaires. However, none of the reported studies surveyed participants that actually completed the selection instruments they were evaluating. Kluger and Rothstein (1993) argue that differences in the amount of cognitive effort required to respond to test items and ego involvement may also produce differences in applicant reactions. Ego involvement reflects the degree of concern with one's level of performance relative to others (Koestner, Zuckerman, & Koestner, 1987). Cognitive ability tests generally yield the most cognitive effort and ego-involvement, and are, thus, less favorably perceived than other selection instruments.

A number of studies have specifically evaluated applicants' perceived job relatedness concerning multimedia SJTs (e.g., Chan & Schmitt, 1997; Kanning et al., 2006; Lievens & Sackett, 2006). Most of these studies have examined the effects of specific test characteristics on applicants' perceived job relatedness of the particular SJTs. For example, Kanning et al. (2006) examined reactions to SJT items that differed with regard to interactivity (non-interactive versus interactive) and medium (video versus paper-and-pencil). Video-based SJT items, in which the response of the participants determines the further course of the item, were perceived as the most favorable in terms of enjoyment, acceptance, and job relatedness.

Individual differences and perceived job relatedness

To attract applicants and retain them in the selection process, organizations have to understand applicant's preferences towards selection instruments (Macan et al., 1994). The literature on applicant reactions until now lacks a clear consensus regarding potential causes of applicants' perceived job relatedness (Chan & Schmitt, 2004; Ryan & Ployhart, 2000). Research has shown that test content and test characteristics affect the perceived job relatedness of selection instruments, but there is still substantial variance in these perceptions that remains unexplained. Brutus (1995) proposed that the perceived job relatedness of selection instruments is affected by test characteristics, but also may be affected by individual differences. Individual differences include applicants' pretest feelings and attitudes that may reflect previous experiences or attitudes about tests, such as anxiety and self-efficacy, and also applicants' more general characteristics, such as core self-evaluations and personality (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997). Examining the effects of individual differences on the perceived job relatedness of selection instruments seems important for two reasons. Conceptually, it would further increase our understanding of the nature of applicant reactions. Practically, it would help test developers to identify specific sources of differences in applicant reactions. If negative applicant reactions are due to individual differences instead of test content, modifying the test content or test administration medium will have little effect (Schmitt & Chan, 1999). Interestingly, despite several calls for the inclusion of individual differences in the applicant reaction literature (Anderson, 2003; Bauer et al., 2004; Chan & Schmitt, 2004; Ryan & Ployhart, 2000), the relationships between individual differences and applicant reactions have remained largely unexplored. This paper will address this shortcoming by examining the effects of individual differences on the perceived job relatedness of a cognitive ability test and a multimedia SJT. There are several individual differences that we expect or that have been previously shown to affect applicant reactions. These can be clustered into three categories: Anxiety, Self-evaluations, and Personality.

Anxiety. Test anxiety is composed of individuals' cognitive and affective reactions to evaluative situations, in the times prior to, during, and after evaluative tasks (Cassady & Johnson, 2002). Test anxiety consists of two dimensions, namely physiological responses experienced during evaluative situations and excessive worrying (Hembree, 1988). Individuals with test anxiety are often concerned with subsequent confrontations with similar evaluative tasks and with loss of self-worth (Depreeuw, 1984). Test anxiety has been found to be related to withdrawal from the selection process (Schmit & Ryan, 1997).

As the cognitive ability test and the multimedia SJT are administered via the computer, computer anxiety may also affect applicant reactions. Computer anxiety is an affective response where people are worried about damaging the computer, looking stupid or losing control over their work (Bloom & Hautaluoma, 1990). A number of studies found that the lack of experience with computers is a major determinant of

computer anxiety (e.g., Beckers & Schmidt, 2003; Heinssen, Glass, & Knight, 1987). Wiechman and Ryan (2003) demonstrated that computer anxiety explained significant variance in process fairness, face validity, perceived difficulty, enjoyment, and self-assessed performance regarding a computer-based in-basket exercise. Therefore, our first hypothesis is:

Hypothesis 1: Anxiety (test anxiety and computer anxiety) will be negatively related to the perceived job relatedness of a cognitive ability test and a multimedia SJT.

Self-evaluations. In our study, the category self-evaluations contains three dimensions, namely test-taking self-efficacy, core self-evaluations, and subjective well-being. Test-taking self-efficacy is the belief that one can perform effectively (Bandura, 1997), that is in this case to perform well on the selection instrument. According to Bandura (1997), self-efficacy determines how much effort people will expend on an activity and how long they will persevere when confronting obstacles. Of the self-evaluation constructs, to our knowledge only test-taking self-efficacy has been studied in relation to applicant reactions. Horvath, Ryan, and Stierwalt (2000) demonstrated that individuals who believe that they will perform well will see the test as fairer and more predictively valid. Test-taking self-efficacy has also been found to be positively related to the perceived job relatedness of several selection instruments (Gilliland, 1994; Ryan et al., 1996; Wiechmann & Ryan, 2003), enjoyment, perceived test ease, and self-assessed test performance (Wiechmann & Ryan, 2003). Core self-evaluations and subjective well-being have not yet been examined with respect to applicant reactions. According to Judge, Locke, and Durham (1997), core self-evaluations is a broad dispositional trait that is indicated by four more specific traits, namely self-esteem, generalized self-efficacy, locus of control, and emotional stability. Core self-evaluations was found to be positively related to job and life satisfaction (Judge, Locke, Durham, & Kluger, 1998), and higher initial levels of work success and steeper work success trajectories (Judge & Hurst, 2008). Subjective well-being comprises people's long-term levels of pleasant affect, lack of unpleasant affect, and life satisfaction (Diener, 1994). Characteristics related to subjective well-being include confidence, optimism, self-efficacy, likeability, effective coping with challenge and stress, originality, and flexibility (Lyubomirsky, King, & Diener, 2005). We believe that individuals with positive dispositions will have more positive emotions and cognitions in evaluative situations, and therefore will react more positively concerning the perceived job relatedness of a cognitive ability test and a multimedia SJT. Therefore, we hypothesize the following:

Hypothesis 2: Self-evaluations (test-taking self-efficacy, core self-evaluations, and subjective well-being) will be positively related to the perceived job relatedness of a cognitive ability test and a multimedia SJT.

Personality. Extensive research has documented the relationship between personality traits and job performance (e.g., Barrick & Mount, 1991) and employee attitudes (e.g., Judge, Heller, & Mount, 2002; Organ, 1994). However, the relationship between personality traits and applicant reactions has been examined in only a limited number of studies (Bernerth et al., 2006; Truxillo, Bauer, Campion, & Paronto, 2006; Viswesvaran & Ones, 2004; Wiechmann & Ryan, 2003). Among these there is a study by Wiechmann and Ryan (2003), who examined the relationship between openness to experience and a number of applicant reactions towards a computer-based in-basket exercise. They found a positive relationship between openness to experience and face validity. Truxillo, Bauer, Campion, and Paronto (2006) found that neuroticism was consistently negatively related and agreeableness was consistently positively related to police recruit applicants' perceived fairness of a paper-and-pencil multiple choice test, to self-assessed performance, and to perceptions of the hiring organization. Regarding a paper-and-pencil organizational leadership test, Bernerth et al. (2006) found that agreeableness and openness to experience were positively related to the perceived procedural justice about the use of a leadership test as selection instrument and also to the perceived distributive justice about the selection decision. Furthermore, neuroticism was negatively related to the perceived distributive justice about the selection decision.

Agreeableness focuses on interpersonal relations. Specifically, it is related to individual differences in the motivation to maintain positive relations with others (Graziano & Eisenberg, 1997). Highly agreeable individuals are trusting, sympathetic, and cooperative (Costa & McCrae, 1992). Individuals who score low on agreeableness tend to be temperamental, argumentative, emotional, and difficult to calm when distressed (Skarlicki, Folger, & Tesluk, 1999). Therefore, individuals low on agreeableness might have a tendency to react more negatively to selection instruments.

Emotional stability represents an individual's tendency to experience psychological distress (Costa & McCrae, 1992). Individuals with low scores of emotional stability tend to be fearful of novel situations and susceptible to feelings of helpfulness and dependence (Wiggins, 1995). Emotional stability also refers to the subjective ability to respond to external stimuli while keeping emotions and impulses under control (Marcati, Gianluigi, & Peluso, 2008). As evaluative situations are generally experienced as stressful, individuals who score low on emotional stability will be inclined to project their negative emotions on their perceived job relatedness of the selection instruments.

Individuals high in openness to experience tend to be intellectually curious and behaviorally flexible in their attitudes and values (Costa & McCrae, 1992). Individuals low in openness to experience fear the unknown and ambiguity involved in evaluative situations (Bernerth et al., 2006). Therefore, it is likely that there will be some resistance to modern computer-based selection instruments. Individuals who are less resistant to new experiences may react more positively to computer-based selection instruments than individuals who are resistant to new experiences (Wiechmann & Ryan, 2003).

Based on the results of Wiechmann and Ryan (2003), Truxillo et al. (2006), and Bernerth et al. (2006) we expect agreeableness, emotional stability and openness to experience to be positively related to the perceived job relatedness of a cognitive ability test and a multimedia SJT. Therefore, our last hypothesis is as follows:

Hypothesis 3: Agreeableness, emotional stability, and openness to experience will be positively related to the perceived job relatedness of a cognitive ability test and a multimedia SJT.

Method

Participants and procedure

This study was conducted among 153 psychology students at a large Dutch University. Of the students, 85 were master students (55.6%) and 68 were bachelor students (44.4%), 101 were female (66.0%) and 52 were male (34.0%). Their age ranged from 19 to 44 (M = 22.3; SD = 3.17). Of the students, 106 (69.3%) had experience with cognitive ability tests and 41 (26.8%) had experience with multimedia SJTs. Most of them had some kind of work experience (70.1%).

As part of their educational program, students completed a cognitive ability test and a multimedia SJT intended to measure managerial skills. We attempted to motivate the students to perform well on the selection instruments by emphasizing the benefits they could have in the future when they would really apply for a job, by practicing with genuine selection instruments, and by giving them a professional report of their scores. To provide a frame of reference, the participants were told that the tests they were about to complete are generally used in the assessment of managers, a profession most students are familiar with. Before completing the actual cognitive ability test and multimedia SJT participants had to fill out a computer-based personality questionnaire and a paper-and-pencil questionnaire containing items on test anxiety, computer anxiety, core self-evaluations, and subjective well-being. After the introduction of the cognitive ability test and the multimedia SJT, participants had to fill out a questionnaire containing items on test-taking self-efficacy. Immediately after completing each selection test participants had to fill out a questionnaire containing items on face validity, perceived predictive validity, and self-assessed test performance. It took the students about two and a half hour to complete all tests and questionnaires.

Measures

Individual differences. Personality, test anxiety, computer anxiety, core self-evaluations, and subjective well-being were measured before participants started the tests. Participants rated the items on a scale ranging from 1 = *strongly disagree* to 5 = *strongly agree*.

The personality traits were measured with a 224-item computer-based personality questionnaire developed by GITP (Koch, 1998), a large Dutch HR-consultancy firm. Each scale consists of 23 to 47 items. An example of an item for Extraversion is as

follows: '*Rate yourself on the following statement: Enjoys meeting new people*'. The scales of the personality questionnaire show substantial correlations (r = .48 - .72) with scales of the revised NEO-Personality Inventory that were intended to measure the same constructs (Costa & McCrae, 1992). Coefficient alphas are substantial: α = .92 for Extraversion, α = .83 for Agreeableness, α = .92 for Conscientiousness, α = .88 for Emotional Stability, α = .90 for Openness to experience. Correlations from .09 - .51 were found between the scales.

Test anxiety was defined as the individuals' cognitive and affective reactions to evaluative situations, in the times prior to, during, and after evaluative tasks (Cassady & Johnson, 2002). This construct was measured with seven items, adopted from Cassady and Johnson (2002). An example of an item is: '*At the beginning of a test, I am so nervous that I often can't think straight*'. In this study, coefficient alpha equals .85.

Computer anxiety is an affective response where people are worried about damaging the computer, looking stupid or losing control over their work (Bloom & Hautaluoma, 1990). This construct was measured with five items, adopted from Heinssen, Glass, and Knight (1987). An example of an item is: '*I hesitate to use a computer for fear of making mistakes that I can not correct*'. In this study, coefficient alpha equals .81.

Core self-evaluations was defined as basic conclusions or bottom-line evaluations that individuals hold about themselves (Judge et al., 1997), and was measured with the 12-item Core Self Evaluation Scale of Judge, Erez, Bono, and Thoreson (2003). An example of an item is: '*I am confident I get the success I deserve in life*'. In this study, coefficient alpha equals .86.

Subjective well-being was measured with the Satisfaction With Life Scale (Diener, Emmons, Larsen, & Griffin, 1985), a five-item scale designed to measure global cognitive judgments of satisfaction with one's life. An example of an item is: '*In most ways my life is close to ideal*'. In this study, coefficient alpha equals .70.

Test-taking self-efficacy was measured after a short introduction of the test. Participants rated the items on a scale ranging from 1 = *strongly disagree* to 5 = *strongly agree*. Test-taking self-efficacy was measured with three items, adopted from Pintrich and De Groot's (1990) Motivated Strategies for Learning Questionnaire (MSLQ). An example of an item is: '*I think I will do very well on this test*'. In this study, coefficient alpha equals .83 for the cognitive ability test and .81 for the multimedia SJT.

Post-test measures. *Face validity, perceived predictive validity*, and *self-assessed test performance* were measured after each test, but before participants received feedback on their test scores. Participants rated the items on a scale ranging from 1 = *strongly disagree* to 5 = *strongly agree*.

Face validity was measured with three items adopted from Smither et al. (1993). Face validity is defined as the extent to which test takers perceive the content of the selection procedure to be related to the job. Unlike content validity, face validity is assessed by test takers who do not have the expertise of test developers or other subject matter experts. To provide a frame of reference, participants were asked to

give ratings on the items concerning relationships between the test and the job of a manager. An example of an item is: '*It would be obvious to anyone that the test is related to a managerial job*'. In this study, coefficient alpha equals .74 for the cognitive ability test and .69 for the multimedia SJT.

Perceived predictive validity was measured with three items adopted from Smither et al. (1993). Perceived predictive validity is defined as the perception of how well the selection procedure predicts future job performance, regardless of how the selection procedure looks like (Smither et al., 1993). To provide a frame of reference, participants were asked to give ratings on the items concerning relationships between the test and the job of a manager. An example of an item is: '*I am confident that the test can predict how well an applicant will perform in a managerial job*'. In this study, coefficient alpha equals .81 for the cognitive ability test and .73 for the multimedia SJT. A series of confirmatory factor analyses was conducted to test whether face validity and perceived predictive validity are distinguishable dimensions of job relatedness. The second-order structure, with job relatedness as the higher level factor and face validity and perceived predictive validity as the first-order factors, showed good fit (Hu & Bentler, 1999) for both the cognitive ability test ($\chi^2 = 9.03$, $df = 6$, $p = .17$, CFI = .99, SRMR = .03, RMSEA = .06) and the multimedia SJT ($\chi^2 = 10.67$, $df = 6$, $p = .10$, CFI = .98, SRMR = .04, RMSEA = .07). Moreover, the fit of the second-order structure was significantly better for both the cognitive ability test ($\Delta\chi^2 = 27.52$, $df = 3$, $p < .01$) and the multimedia SJT ($\Delta\chi^2 = 41.96$, $df = 3$, $p < .01$) than the fit of the model with job relatedness as single factor. These results confirm that face validity and perceived predictive validity are two related, but distinguishable, dimensions of job relatedness.

Self-assessed test performance was measured with four items, based on the scale of Wiechmann and Ryan (2003). An example of an item is: '*I think I have performed well on the test*'. In this study, coefficient alpha equals .83 for the cognitive ability test and .78 for the multimedia SJT.

Cognitive ability test. The computer-based cognitive ability test is developed by GITP (Van Leeuwen, 2004), a large Dutch HRD consultancy firm, and consists of three scales, namely Verbal Reasoning (VR), Number Series (NS) and Abstract Reasoning (AR). Together, the three scales aim to measure general cognitive ability. The test consists of 81 items. An example of an item of the NS scale is as follows: '*Complete the following series of numbers: 10 11 13 16 20 25?*' The scales of the cognitive ability test show substantial correlations ($r = .44$ - .78) with the Dutch intelligence test series of Drenth, a frequently used measure of cognitive ability in The Netherlands (Drenth, 1965). The time limit to complete all items was 51 minutes. Coefficient alphas of the scales, based on a sample of candidates who had completed all items within the time limit, were .87 for the VR scale (N =889), .63 for the NS scale (N = 649), and .68 for the AR scale (N = 757). There were moderate correlations between the three scales (r = .24 – .41). The total amount of correctly answered items represents the participants' scores, which could range from 0 – 81.

Multimedia SJT. The SJT consists of 17 short video clips, representing a wide range of work-related situations managers are likely to encounter on their job. Each situation depicts a manager and a subordinate interacting on the job and describes an interpersonal or job-related problem. After each situation, four possible ways to handle the situations are presented via video clips. Participants are asked to judge these response alternatives on a five-point scale ranging from (--) *very ineffective* to (++) *very effective*. An expert-based scoring method was used to score the participants' effectiveness ratings of the response alternatives (Bergman et al., 2006). Ten experts individually watched the videotaped vignettes and rated the four response alternatives on the same five-point scale. The absolute distance between the mean effectiveness ratings of the experts and the participants' effectiveness ratings was calculated for each response alternative. The absolute distances of all responses were summed and extracted from 100, so participants receive a higher score if they tend to agree with the experts. All participants completed the multimedia SJT within 45 minutes. In this study, coefficient alpha equals .91.

Results

Descriptive statistics

Means, standard deviations, reliabilities and correlations between all study variables are presented in Table 1 for the cognitive ability test and in Table 2 for the multimedia SJT. Before we tested the hypotheses, we first looked at significant correlations between demographic characteristics and the other study variables. Age was significantly related to emotional stability (r = .24, p < .01), openness to experience (r = .19, p < .05), and the perceived predictive validity of the cognitive ability test (r = .36, p < .01). Gender was related to a number of study variables. The largest difference between male students and female students was found for self-efficacy regarding the cognitive ability test (r = -.38, p < .01, t = 4.51, p < .01) and core self-evaluations (r = -.25, p < .05, t = 3.45, p < .01) in favor of the male students. Job experience was significantly related to the perceived predictive validity of both the cognitive ability test (r = .18, p < .05) and the multimedia SJT (r = .19, p < .05). Experience with a cognitive ability test was significantly related to test-taking self-efficacy (r = .23, p < .01), core self-evaluations (r = .17, p < .05), and emotional stability (r = .24, p < .01). Experience with the multimedia SJT was significantly related to test-taking self-efficacy (r = .23, p < .01) and conscientiousness (r = .27, p < .01). Because of these significant correlations, we controlled for age, gender, job experience and test experience in the regression analyses.

We conducted paired-sample t-tests to examine whether the perceived job relatedness of the cognitive ability test differed from the perceived job relatedness of the multimedia SJT. Participants rated the face validity (M = 4.41, SD = 0.51) and the predictive validity (M = 3.60, SD = 0.61) of the multimedia SJT significantly higher than the face validity (M = 3.76, SD = 0.81, t = -8.92, p < .01) and the predictive validity (M = 2.91, SD = 0.77, t = -9.95, p < .01) of the cognitive ability test.

The role of individual differences in job relatedness perceptions

Research has shown that test performance has an influence on applicant reactions (Chan, Schmitt, Sacco et al., 1998). Thus, to provide a stringent test of the effects of individual differences on the perceived job relatedness of the cognitive ability test and the multimedia SJT, we controlled for self-assessed test performance in the analyses. In this study self-assessed test performance is a more appropriate control variable than actual test performance, because participants were not yet notified of their test scores when they reported the perceived job relatedness of the selection instruments.

The results for Hypotheses 1 - 3, regarding the effects of individual differences on job relatedness, are given in Table 1 for the cognitive ability test and in Table 2 for the multimedia SJT. Hypothesis 1, which stated that test anxiety and computer anxiety would be negatively related to perceived job relatedness, received only weak support. No significant correlations were found between test anxiety and the perceived job relatedness of the cognitive ability test and the multimedia SJT. However, computer anxiety was negatively related to the face validity of the multimedia SJT (r = -.20, p < .05). It was unrelated to the face validity of the cognitive ability test (r = -.08, *ns*), and also unrelated to the perceived predictive validity of the cognitive ability test (r = -.13, *ns*) and the multimedia SJT (r = .01, *ns*). No significant correlations were found between test anxiety and the perceived job relatedness of the cognitive ability test and the multimedia SJT.

Hypothesis 2 stated that test-taking self-efficacy, core self-evaluations and subjective well-being would be positively related to perceived job relatedness. This hypothesis was partly supported as the dimension core self-evaluations was positively related to the perceived predictive validity of the cognitive ability test (r = .19, p < .05) and the face validity of the multimedia SJT (r = .20, p < .05), and subjective well-being was positively related to the face validity of the multimedia SJT (r = .17, p < .05) and the perceived predictive validity of the multimedia SJT (r = .17, p < .05). No significant correlations were found between test-taking self-efficacy and the face validity and the perceived predictive validity of the cognitive ability test and the multimedia SJT.

Hypothesis 3, which stated that agreeableness, emotional stability, and openness to experience would be positively related to perceived job relatedness, was supported regarding the perceived job relatedness of the cognitive ability test. Agreeableness was positively related to its face validity (r = .20, p < .05) and its perceived predictive validity (r = .22, p < .05), emotional stability was positively related to its face validity (r = .27, p < .01) and its perceived predictive validity (r = .26, p < .01), and openness to experience was positively related to its face validity (r = .27, p < .01) and its perceived predictive validity (r = .29, p < .01). Openness to experience was also significantly related to the face validity of the multimedia SJT (r = .19, p < .05). We did not find other significant correlations between the personality dimensions and the perceived job relatedness of the multimedia SJT. Therefore, Hypothesis 3 was not supported regarding the multimedia SJT.

Table 1

Descriptive Statistics and Correlations between Individual Differences and the Perceived Job Relatedness of the Cognitive Ability Test

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Age	22.33	3.17	(-)															
2. Gender	.66	0.48	-.02	(-)														
3. Job experience	1.51	1.12	.34**	-.05	(-)													
4. Test experience	3.81	0.84	-.17*	-.08	-.02	(.82)												
Individual differences																		
5. Test anxiety	2.41	0.67	.01	-.07	.03	-.14	(.85)											
6. Computer anxiety	1.35	0.44	-.04	-.08	.04	-.16	.19*	(.81)										
7. Test-taking self-efficacy	3.33	0.56	.09	-.38**	.01	.23**	-.25**	-.19*	(.83)									
8. Core self-evaluations	3.75	0.48	.12	-.25**	.09	.17*	-.55**	-.39**	.31**	(.86)								
9. Subjective well-being	3.79	0.53	-.05	.01	-.02	.15	-.29**	-.32**	.30**	.57**	(.70)							
10. Extraversion	3.55	0.52	.13	-.18*	.11	.08	-.08	-.16	.04	.34**	.19*	(.92)						
11. Agreeableness	3.74	0.30	.07	.04	.03	.05	.01	-.17	.00	.15	.12	.22**	(.83)					
12. Conscientiousness	3.66	0.38	.15	.12	.10	.02	-.10	-.05	.-.01	.16*	.16	.10	.31**	(.92)				
13. Emotional stability	3.30	0.43	.24**	-.23**	.06	.24**	-.41**	-.34**	.24**	.62**	.39**	.34**	.25*	.13	(.88)			
14. Openness to experience	3.79	0.29	.19*	-.20*	.09	.16	-.12	-.27**	.20*	.35**	.11	.51**	.37**	.17*	.30**	(.90)		
Post-test measures																		
15. Face validity	3.76	0.81	.14	.02	.02	.05	-.04	-.08	-.03	.14	-.01	.12	.20*	.14	.27**	.27**	(.74)	
16. Perceived predictive validity	2.91	0.77	.36**	-.20*	.18*	-.01	-.01	-.13	.14	.19*	.05	.16	.22*	.14	.26**	.29**	.54**	(.81)

Note. Coefficient alphas are presented on the diagonal, between parentheses. Gender is coded as follows: 0 = male, 1 = female. Job experience is coded as follows: 0 = no experience, 1 = less than one year, 2 = one to five years, 3 = 6-10 years, and 4 = more than 10 years. All scales range from 1-5. The correlations with face validity and perceived predictive validity are controlled for self-assessed test performance. $N = 153$.

* $p < .05$, ** $p < .01$

Table 2

Descriptive Statistics and Correlations between Individual Differences and the Perceived Job Relatedness of the Multimedia SJT

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Age	22.33	3.17	(-)															
2. Gender	.66	0.48	-.02	(-)														
3. Job experience	1.51	1.12	.34**	-.05	(-)													
4. Test experience	2.80	1.16	.19*	-.08	.13	(.84)												
Individual differences																		
5. Test anxiety	2.41	0.67	.01	.20*	.03	-.09	(.85)											
6. Computer anxiety	1.35	0.44	-.04	-.07	.04	-.05	..19**	(.81)										
7. Test-taking self-efficacy	3.39	0.51	-.01	-.19*	.09	.23**	-.10	-.08	(.81)									
8. Core self-evaluations	3.75	0.48	.12	-.25**	-.02	.07	-.55**	-.39**	.31**	(.86)								
9. Subjective well-being	3.79	0.53	-.05	.01	.13	-.02	-.29**	-.32**	.35**	.57**	(.70)							
10. Extraversion	3.55	0.52	.13	-.18*	.11	.11	-.08	-.16	.27**	.34**	.19*	(.92)						
11. Agreeableness	3.74	0.30	.07	.04	.03	.00	.01	-.17*	-.05	.15	.12	.22**	(.83)					
12. Conscientiousness	3.66	0.38	.15	.12	.10	.27**	-.10	-.05	.14	.16*	.16	.10	.31**	(.92)				
13. Emotional stability	3.30	0.43	.24**	-.23**	.06	.14	-.41**	-.34**	.24**	.62**	.39**	.34**	.25*	.13	(.88)			
14. Openness to experience	3.79	0.29	.19*	-.20*	.09	.04	-.12	-.27**	.24**	.35**	.11	.51**	.37**	.17*	.30**	(.90)		
Post-test measures																		
15. Face validity	4.41	0.51	.-.05	.13	-.07	-.05	-.06	-.20*	.08	.20*	.17*	.15	.08	-.13	.12	.19*	(.69)	
16. Perceived predictive validity	3.60	0.61	-.04	.09	.19*	.15	.03	.01	.12	.04	.17*	.05	.08	.10	.01	-.09	.39**	(.73)

Note. Coefficient alphas are presented on the diagonal, between parentheses. Gender is coded as follows: 0 = male, 1 = female. Job experience is coded as follows: 0 = no experience, 1 = less than one year, 2 = one to five years, 3 = 6-10 years, and 4 = more than 10 years. All scales range from 1-5. The correlations with face validity and perceived predictive validity are controlled for self-assessed test performance. $N = 153$.

* $p < .05$, ** $p < .01$

Table 3

Hierarchical Regression Model Testing for the Association of Individual Differences and Face Validity of the Cognitive Ability Test

	β	*t*	R^2	ΔR^2	ΔF
Step 1 – Control variables					
Age	.10	0.99			
Gender	-.02	-0.20			
Job experience	-.11	-1.18			
Test experience	-10	1.08			
Self-assessed test performance	.08	0.98	.07	.07	1.77
Step 2					
Openness to experience	.20	2.18*	.12	.06	7.84**
Step 3					
Emotional stability	.19	1.99*	.15	.03	3.95**
F (7,147) = 3.08**					

Note. Gender is coded as follows: 0 = male, 1 = female. Job experience is coded as follows: 0 = no experience, 1 = less than one year, 2 = one to five years, 3 = 6-10 years, and 4 = more than 10 years. β coefficients in the overall model are presented. R^2 and ΔR^2 may appear inconsistent due to rounding.

* $p < .05$, ** $p < .01$

Table 4

Hierarchical Regression Model Testing for the Association of Individual Differences and Perceived Predictive Validity of the Cognitive Ability Test

	β	*t*	R^2	ΔR^2	ΔF
Step 1 – Control variables					
Age	.28	3.00**			
Gender	-.15	-1.73			
Job experience	.03	0.34			
Test experience	.03	0.37			
Self-assessed test performance	.10	1.20	.16	.16	4.62**
Step 2					
Openness to experience	.19	2.11*	.19	.03	4.46**
F (6, 144) = 4.71**					

Note. Gender is coded as follows: 0 = male, 1 = female. Job experience is coded as follows: 0 = no experience, 1 = less than one year, 2 = one to five years, 3 = 6-10 years, and 4 = more than 10 years. β coefficients in the overall model are presented.

$p < .05$, ** $p < .01$

Table 5

Hierarchical Regression Model Testing for the Association of Individual Differences and Face Validity of the Multimedia SJT

	β	*t*	R^2	ΔR^2	ΔF
Step 1 – Control variables					
Age	-.08	-0.87			
Gender	.20	2.23*			
Job experience	-.04	-0.43			
Test experience	-.05	-0.52			
Self-assessed test performance	.02	0.23	.03	.03	0.65
Step 2					
Openness to experience	.19	2.10*	.08	.06	7.77**
Step 3					
CSE	.19	2.03*	.11	.03	4.12*
F (7, 149) = 2.23*					

Note. Gender is coded as follows: 0 = male, 1 = female. Job experience is coded as follows: 0 = no experience, 1 = less than one year, 2 = one to five years, 3 = 6-10 years, and 4 = more than 10 years. β coefficients in the overall model are presented. R^2 and ΔR^2 may appear inconsistent due to rounding.

* $p < .05$, ** $p < .01$

Table 6

Hierarchical Regression Model Testing for the Association of Individual Differences and Perceived Predictive Validity of the Multimedia SJT

	β	*t*	R^2	ΔR^2	ΔF
Step 1 – Control variables					
Age	-.11	-1.19			
Gender	.12	1.41			
Job experience	.22	2.40			
Test experience	.14	1.64*			
Self-assessed test performance	-.07	-0.77	.08	.08	2.07*
Step 2					
Subjective well-being	.19	2.18*	.11	.03	4.74*
F (6, 143) = 2.56*					

Note. Gender is coded as follows: 0 = male, 1 = female. Job experience is coded as follows: 0 = no experience, 1 = less than one year, 2 = one to five years, 3 = 6-10 years, and 4 = more than 10 years. β coefficients in the overall model are presented.

* $p < .05$, ** $p < .01$

In addition, we conducted a series of stepwise multiple regression analyses, to examine which individual difference explains most of the variance in job relatedness perceptions. Step 1 included the control variables: Age, gender, job experience, test experience, and self-assessed test performance. Step 2 included the individual differences which we expected to affect perceived job relatedness (see Table 3 – 6). Regarding the face validity of the cognitive ability test, openness to experience (β = .20, t = 2.18, p < .05) and emotional stability (β = .19, t = 1.99, p < .05) survived the stepwise procedure. Regarding the perceived predictive validity of the cognitive ability test, only openness to experience (β = .19, t = 2.11, p < .05) explained additional variance up to and beyond the control variables. Regarding the face validity of the multimedia SJT, openness to experience (β = .19, t = 2.10, p < .05) and core self-evaluations (β = .19, t = 2.03, p < .05) explained additional variance up to and beyond the control variables. Regarding the perceived predictive validity of the multimedia SJT, only subjective well-being (β = .19, t = 2.18, p < .05) explained additional variance up to and beyond the control variables.

Discussion

The aim of this study was to examine the relationship between individual differences and perceived job relatedness, which consisted of two related, but distinguishable dimensions, namely face validity and perceived predictive validity. The results indicated that computer anxiety, core self-evaluations, subjective well-being, agreeableness, emotional stability, and openness to experience affected the perceived job relatedness of a cognitive ability test and a multimedia SJT, but not systematically. Openness to experience was the most consistent predictor of job relatedness perceptions. Given that perceived job relatedness is related to several important organizational outcomes (e.g., Bauer et al., 1998), and considering that the organization's selection procedure is the first contact moment between an employee and an organization, the results reported in this study may have practical implications. We will discuss each of our findings in turn.

First, we expected that test anxiety and computer anxiety would be negatively related to the perceived job relatedness of a cognitive ability test and a multimedia SJT (Hypothesis 1). We found weak support for this hypothesis, as only computer anxiety was significantly related to face validity of the multimedia SJT. The non-significant effects of test anxiety and computer anxiety are surprising, as these individual differences have previously been found to be related to a variety of applicant reactions (Schmit & Ryan, 1997; Wiechmann & Ryan, 2003). These results could possibly be explained by the homogeneous sample, regarding age, cultural background and educational level. Students are frequently exposed to test situations. In our sample nearly 70% of the students had experience with cognitive ability tests, and nearly 30% had experience with multimedia SJTs. Furthermore, students work with computers on a daily basis, demonstrated by the low mean of 1.35 for computer anxiety on a five-point scale. Therefore, it is important to verify and extend our findings in a more heterogeneous sample.

Our second hypothesis stated that test-taking self-efficacy, core self-evaluations, and subjective well-being would be positively related to the perceived job relatedness of a cognitive ability test and a multimedia SJT. This hypothesis was partly supported as the dimension core self-evaluations was positively related to the perceived predictive validity of the cognitive ability test and the face validity of the multimedia SJT, and subjective well-being was positively related to the face validity of the multimedia SJT and the perceived predictive validity of the multimedia SJT. Moreover, in the prediction of the perceived job relatedness of the multimedia SJT, core self-evaluations and subjective well-being were able to explain additional variance over and above age, gender, job experience, test experience, and self-assessed test performance. To our knowledge, core self-evaluations and subjective well-being until now have not yet been examined with respect to applicant reactions. Our findings suggest that self-evaluations should be considered when assessing applicant reactions.

Test-taking self-efficacy has previously been found to be positively related to the perceived job relatedness of selection instruments (Gilliland, 1994; Ryan et al., 1996; Wiechmann & Ryan, 2003). However, our study did not indicate any relationship between test-taking self-efficacy and perceived job relatedness. The setting of our study could possibly explain the non-significant relationship between test-taking self-efficacy and job relatedness perceptions. Self-efficacy is related to how much effort an individual will expend on an activity and how long they will persevere when confronting obstacles (Bandura, 1997). Our results were obtained in a research setting, which typically lacks the motivational and self-presentational issues inherent in actual high-stakes situations. It is possible that applicants would have exerted more effort and gave up les quickly when confronted with difficult items than our participants did. Therefore, differences in test-taking self-efficacy may have more influence on perceptions in a real applicant sample.

Furthermore, we expected that agreeableness, emotional stability, and openness to experience would be positively related to job relatedness perceptions (Hypothesis 3). Despite previous calls for investigating the role of personality traits in explaining differences in applicant reactions (e.g., Chan & Schmitt, 2004; Ryan & Ployhart, 2000), there has been only limited research on the effects of personality on applicant reactions (Bernerth et al., 2006; Wiechmann & Ryan, 2003). The hypothesized relationships between personality and perceived job relatedness were generally supported at the correlational level. Our results indicated that agreeableness, emotional stability, and openness to experience were indeed positively related to the face validity and the perceived predictive validity of the cognitive ability test. Openness to experience was also significantly related to the face validity of the multimedia SJT. These findings are consistent with past findings regarding the relationship between openness to experience and applicant reactions. For example, Bernerth et al. (2006) found that agreeableness, emotional stability, and openness to experience were positively related to distributive justice perceptions about the selection decision. Our findings, coupled with the findings of Bernerth et al., suggest that certain individuals may be more predisposed to react positively to selection instruments.

While, the relationships between individual personality dimensions and perceived relatedness were less consistent in the regression analyses, openness to experience still accounted for additional variance over and above age, gender, job experience, test experience, and self-assessed test performance in the face validity of the cognitive ability test and the multimedia SJT, and the perceived predictive validity of the cognitive ability test. Thus, individuals who are more amenable to new experiences seem to react more positively to computer-based selection instruments than individuals who are resistant to new experiences. Wiechmann and Ryan (2003) also found a positive relationship between openness to experience and the face validity of a computer-based in-basket exercise. Like Wiechmann and Ryan, we measured the perceived job relatedness of modern computer-based selection instruments. Therefore, we can not generalize our findings to selection instruments in general. It is quite plausible that openness to experience is less important when using traditional paper-and-pencil tests. Therefore, we recommend future studies to examine the relationships between personality and the perceived job relatedness of other selection instruments as well.

The importance of examining the relationship between individual differences and job relatedness perceptions using other selection instruments is also emphasized by the different correlations we found for the perceived job relatedness of the cognitive ability test and the perceived job relatedness of the multimedia SJT. For example, the face validity of the cognitive ability test was related to agreeableness, emotional stability, and openness to experience, while the face validity of the multimedia SJT was related to computer anxiety, core self-evaluations, subjective well-being, and openness to experience. This implies that relationships between individual differences and the perceived job relatedness of one selection instrument can not be generalized to other selection instruments. This conclusion is relevant for future research, because most studies on the effects of individual differences on applicant reactions have included only one selection instrument (Bernerth et al., 2006; Truxillo et al., 2006; Wiechmann & Ryan, 2003). The correlates of perceived job relatedness could possibly be determined by the type of construct the test measures. Kluger and Rothstein (1993) argue that differences in the amount of cognitive effort required to respond to test items may produce differences in applicant reactions. Recently, Yeo and Neal (2008) demonstrated that subjective cognitive effort is, in turn, related to personality. Thus, personality might explain more variance in the perceived job relatedness of selection instruments that require relatively more cognitive effort. To assess whether the construct a selection instrument measures indeed affects the correlates of the perceived job relatedness of that particular selection instrument, we recommend future studies to include multiple selection instruments when examining relationships between individual differences and applicant reactions.

We believe that the present study contributed to the knowledge of applicant reactions. Traditionally, researchers have focused on descriptive questions, such as the comparison of favorability reactions across procedures and instruments (e.g., Hausknecht et al., 2004; Kluger & Rothstein, 1993; Rynes & Connerley, 1993). Other

researchers have assessed how test-related factors, such as test content or test method, affect applicant reactions (e.g., Bauer et al., 2004; Chan & Schmitt, 1997; Kanning et al., 2006). For example, Chan and Schmitt (1997) found the face validity of a multimedia SJT to be significantly more positive than the face validity of a paper-and-pencil SJT. However, our findings revealed that stable individual differences may also account for a portion of variance in applicant reactions, thus, suggesting there may be a stable component to applicant reactions in addition to test-related factors. Future applicant reaction research should, therefore, consider individual differences to obtain a more complete understanding of the factors affecting applicant reactions.

Limitations of this study and suggestions for future research

The current study has some general limitations that should be noted. First, we only measured the perceived job relatedness of the selection instruments before the participants received feedback on their test scores. These perceptions of job relatedness may relate to behaviors exhibited by applicants during later stages of the selection process prior to the organization's decision (e.g., intentions to accept the job). However, because test feedback can influence applicant reactions (Bauer et al., 1998), we recommend future studies to also measure the perceived job relatedness of selection instruments after participants receive feedback on their test scores, as these perceptions may be related to more long-term behaviors (Ryan & Ployhart, 2000).

Secondly, as in most studies on applicant reactions (e.g., Bernerth et al., 2006; Chan et al., 1997; Hausknecht et al., 2004; Kluger & Rothstein, 1993; Wiechmann & Ryan, 2003), results were obtained in a research setting, using a population that only consisted of students. The research setting allowed us to assess more individual differences and reactions prior and after each selection instrument than would have been possible in a field setting. Several researchers have noted that the nature of procedural justice perceptions justifies the use of both student and field samples (e.g., Bernerth et al., 2006; Ryan & Ployhart, 2000). Moreover, we attempted to motivate the students to perform well on the selection instruments, by emphasizing the benefits they could have by practicing with genuine selection instruments, and by giving them a professional report of their scores. We believe that the present study provides a contribution to the current literature on applicant reactions, but care should be taken when generalizing the results to an applicant sample.

The use of an applicant sample will also provide the opportunity to assess ethnicity differences in antecedents of the perceived job relatedness of selection instruments. For example, Viswesvaran and Ones (2004) found differences across ethnic groups in the importance they placed on different aspects of selection system characteristics that relate to fairness perceptions. Future research could examine whether these ethnicity difference also apply to the perceived job relatedness of selection instruments. Furthermore, the use of an applicant sample will also provide the opportunity to assess relationships between applicant reactions and important consequences for organizations, such as applicant retention, withdrawal from the hiring process, and subsequent job performance (Hausknecht et al., 2004).

Previous research has shown that job relatedness perceptions of instruments are influenced by the context in which the instrument is being used (e.g., Elkins & Phillips, 2000; Murphy, Thornton, & Prue, 1991). For example, Elkins and Philips (2000) demonstrated that a biodata instrument is more positively perceived in terms of job relatedness when the instrument is used for the selection of entry-level international managerial jobs than for the selection of non-specified managerial jobs. In the present study participants were told that the cognitive ability test and the multimedia SJT they were about to complete were generally used in the assessment of managers. Because both selection tests are used in the assessment of a variety of managerial jobs in a variety of companies, we intended to make the findings generalizable to this wide range of managerial jobs. Therefore, the job context was not specified in the present study. Yet, in future studies it would be worth examining whether the type of managerial job to which applicants are applying for affects the relationship between individual differences and the perceived job relatedness of the selection instruments.

In the present study we examined the effects of individual differences on the perceived job relatedness of two often used selection instruments. Although, perceived job relatedness is the most studied dimension of applicant reactions to different selection instruments (e.g., Chan & Schmitt, 1997; Lievens & Sackett, 2006; Ryan & Ployhart, 2000), other reactions, for example fairness perceptions, have also been found to affect organizational outcomes (e.g., Bauer et al., 1998; Ryan & Ployhart, 2000). Therefore, we would recommend studying the effects of individual differences on a broader range of applicant reactions.

The results of our study suggest that certain individuals may be more predisposed to react positively to selection instruments. Applicant reactions are, thus, not only influenced by the selection instrument or medium itself, but also by factors outside the organization's control. Interventions to improve applicant reactions are, therefore, less likely to be effective for all applicants. The nature of the applicant pool should be carefully considered when designing interventions to improve applicant reactions. We encourage further research on the effect of individual differences on applicant reactions using additional measures, samples, and selection instruments.

Chapter 6

Pretest and posttest reactions to a paper-and-pencil and a computerized in-basket exercise*

* This chapter is submitted for publication as:
Oostrom, J. K., Bos-Broekema, L., Serlie, A. W., Born, M. Ph., & Van der Molen, H. T. (submitted). Pretest and posttest reactions to a paper-and-pencil and a computerized in-basket exercise.

Abstract

The present study compared pretest and posttest face validity perceptions, predictive validity perceptions, and fairness perceptions regarding a paper-and-pencil version and a computerized version of an in-basket exercise. Furthermore, the nature of these reactions and their relationship with test performance were examined. Data were collected among 205 applicants. Contrary to our expectations, results showed that the paper-and-pencil in-basket was more positively perceived in terms of predictive validity than the computerized in-basket exercise. The comparison of the other applicant reactions yielded no significant differences between the two versions of the in-basket exercise. Results from structural equation modeling showed that applicants' general beliefs in tests affected pretest reactions. Applicants' test performance influenced posttest reactions via self-assessed test performance. Theoretical and practical implications of these results are discussed.

Introduction

In the past decades, new technology has influenced personnel selection practices (Anderson et al., 2004; Bartram, 2005a). For example, traditional paper-and-pencil tests are more and more replaced by computerized tests. The use of computerized tests has several economic and practical benefits, such as reduced costs, increased standardization, a positive image of the organization, and the possibility to provide immediate feedback to applicants (Chan & Schmitt, 2004; Thibodeaux & Kudisch, 2003; Wiechmann & Ryan, 2003). However, in order to realize the benefits of the use of computerized tests in selection contexts, it has to be accepted by applicants. There are several important practical implications associated with positive applicant reactions to selection instruments and procedures, namely stronger intentions to accept job offers, intentions to recommend the organization to others, a decreased likelihood of litigation, and perceived organizational attractiveness (Anderson et al., 2004; Gilliland, 1993; Ryan & Ployhart, 2000; Smither et al., 1993).

The increased use of computerized tests in selection practices has led to a rapid growth in research comparing applicant reactions to traditional paper-and-pencil tests and computerized tests (Anderson et al., 2004; Bartram, 2005a). These studies generally yield that computerized tests are equally or better perceived than their paper-and-pencil counterparts (e.g., Mead, 2001; Potosky & Bobko, 2004; Reynolds et al., 2000; Salgado & Moscoso, 2003). Although this research has increased our understanding of how applicants react to new technology in selection contexts, there are a number of limitations to these studies affecting their practical and theoretical value. First, past research has often used student samples rather than actual applicant samples (e.g., Chan & Schmitt, 1997; Horvath et al., 2000; Potosky & Bobko, 2004; Rynes & Connerley, 1993; Wiechmann & Ryan, 2003). A student sample limits the ecological validity of studies on applicant reactions. Second, in the majority of research applicant reactions are measured on a single occasion, either pretest (e.g., Rynes & Connerley, 1993; Schmit & Ryan, 1997) or posttest (e.g., Lievens & Sackett, 2006; Richman-Hirsch et al., 2000; Salgado & Moscoso, 2003). Several researchers have argued that pretest and posttest reactions cannot be considered as interchangeable (e.g., Chan & Schmitt, 2004; Chan et al., 1997; Hausknecht et al., 2004). Inferences drawn from studies in which applicant reactions are measured on a single occasion may therefore vary based on when the reactions were measured. Third, the nature of the studies on applicant reactions to computerized selection instruments has been rather descriptive and comparative, rather than explanatory (e.g., Kanning et al., 2006; Reynolds et al., 2000; Richman-Hirsch et al., 2000).

The present study aims to fill these three voids by conducting a field study that draws upon the model of Chan, Schmitt, Sacco et al. (1998) which has been developed to explain the nature of the three most commonly studied dimensions of applicant reactions, namely face validity perceptions, predictive validity perceptions, and fairness perceptions (Chan & Schmitt, 2004), and their relationship with test performance. Face validity refers to the extent to which the content of the selection procedure seems to be related to the job (Smither et al., 1993). Perceived predictive

validity refers to the perception of how well the selection instrument predicts future job performance, regardless of what the selection instrument looks like (Smither et al., 1993). Fairness refers to the extent to which a test seems to rule out biases and provide applicants with the same opportunity to perform well (Gilliland, 1993). The objectives of the present study are twofold. Our fist aim is to compare pretest and posttest reactions to a paper-and-pencil version with pretest and posttest reactions to a computerized version of one of the most widely used assessment center exercises, namely an in-basket exercise. The second aim is to examine the nature of pretest and posttest reactions to an in-basket exercise, by drawing upon the model of Chan, Schmitt, Sacco et al. (1998). We will first provide an overview of earlier research on applicant reactions to computerized testing. Then, we will explain the hypothesized structural model regarding the nature of pretest and posttest reactions.

Applicant reactions to computerized testing

Several studies on applicant reactions to advanced technology have compared applicant reactions to paper-and-pencil instruments and computerized instruments with identical contents (Anderson et al., 2004; Bartram, 2005a). Most of these studies have examined personality questionnaires (e.g., Mead, 2001; Salgado & Moscoso, 2003) or situational tests (e.g., Chan & Schmitt, 1997; Kanning et al., 2006; Lievens & Sackett, 2006; Potosky & Bobko, 2004; Richman-Hirsch et al., 2000). For example, Salgado and Moscoso (2003) reported that undergraduate students and managers perceived a computerized personality questionnaire as more comfortable, more pleasant, more preferable, better for the organization, less fatiguing, and less intimidating than a paper-and-pencil personality questionnaire. No differences were found in terms of perceived scientific value, quality of examination, fairness, respect to personal intimacy, accuracy, effectiveness, probability to fake, and confidentiality. Potosky and Bobko (2004) found that graduate students enjoyed taking a computerized cognitive ability test and a situational judgment test (SJT) more than a paper-and-pencil cognitive ability test and SJT. Richman-Hirsch et al. (2000) investigated a paper-and-pencil version, a written version administered by computer (computerized version), and a full-motion video version (multimedia version) of a conflict resolution skill assessment. Managers who completed the multimedia version perceived the assessment as more content valid, more job-related and more enjoyable than the computerized version and the paper-and-pencil version. The authors argued that computerization per se was not enough to affect test reactions; it was the multimedia nature of the assessment that caused more positive reactions. We are aware of only one study in which applicant reactions to a computerized in-basket exercise were examined. Wiechmann and Ryan (2003) examined applicant reactions to a paper-and-pencil and a computerized in-basket exercise that varied in difficulty according to the technical level of the job. Their sample consisted of undergraduate students. In contrast to their expectations, posttest reactions in terms of process fairness and liking did not differ between the two test versions.

In the present study, we will compare pretest and posttest reactions to a paper-and-pencil in-basket exercise with pretest and posttest reactions to a computerized in-basket exercise with identical contents. As described above, researchers have demonstrated that computerized tests are equally or better perceived than their paper-and-pencil counterparts (e.g., Mead, 2001; Potosky & Bobko, 2004; Reynolds et al., 2000; Salgado & Moscoso, 2003). These findings have been attributed to the novelty of computerized testing (Wiechmann & Ryan, 2003), and to the more realistic presentation of items in situational tests and in-basket exercises (Chan & Schmitt, 1997). Since most individuals spent a substantial part of their day working on their computer, we expect the computerized in-basket exercise to show more resemblance to daily work processes, and therefore to be perceived as more face valid and predictively valid. According to Wiechmann and Ryan (2003) individuals perceive computers as more objective, accurate, and less prone to biases than traditional paper-and-pencil tests. As these factors are related to fairness (Chan & Schmitt, 2004), we expect the computerized in-basket exercise also to be perceived as more fair than the paper-and-pencil in-basket exercise. This leads to the following hypothesis: Applicants taking the computerized in-basket exercise perceive the test as more face valid, more predictively valid, and more fair than applicants taking the paper-and-pencil in-basket exercise (Hypothesis 1).

The nature of applicant reactions

Although previous studies on applicant reactions to computerized selection instruments have enhanced our knowledge of the effects of computerization of traditional paper-and-pencil tests on applicant reactions, they have been rather descriptive and comparative in nature. There are a number of conceptualizations, that have provided an explanatory framework of applicant reactions that have not yet been applied to computerized testing (e.g., Bauer et al., 1998; Chan & Schmitt, 1997; Gilliland, 1993; Macan et al., 1994). However, a number of researchers (e.g., Ryan & Huth, 2008; Van Vianen, Taris, Scholten, & Schinkel, 2004) argue that these conceptualizations are insufficient for providing strong psychological explanations regarding the underlying processes of applicant reactions.

To better understand the nature of applicant reactions and their relationship with test performance, Chan, Schmitt, Sacco et al. (1998) developed a structural model for pretest and posttest reactions to cognitive ability tests. They tested this model using a sample of undergraduate students and found empirical support for the model. In the present study we will examine to what extent this model can be generalized to other selection procedures in an actual field study. We will test the model separately for both a paper-and-pencil in-basket exercise and a computerized in-basket exercise. In Figure 1 the hypothesized structural model is presented. The following sections describe the different paths of the model and provide empirical support where possible.

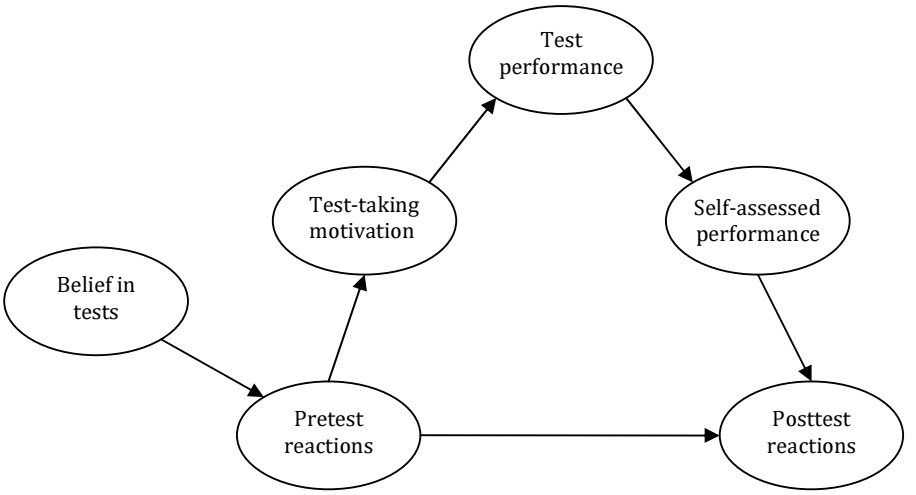


Figure 1. Hypothesized model of determinants and outcomes of pretest and posttest reactions.

The effects of belief in tests on pretest reactions. Pretest reactions are measured by asking participants to report their perceptions of a test after reading a written description of the test or sample test items, but before they start the actual test. According to Chan, Schmitt, Sacco et al. (1998) pretest reactions are important because in addition to test characteristics, they may reflect applicant's prior experiences or general beliefs about tests. As asserted by Arvey et al. (1990), general beliefs in tests refer to whether an applicant believes that tests are a good way of selecting people into jobs, that they are a good reflection of what a person can do on the job, that they are valid and that they should not be eliminated from the selection process. Chan, Schmitt, Sacco et al. found that belief in tests positively influenced pretest face validity perceptions, predictive validity perceptions, and fairness perceptions. Lievens, DeCorte, and Brysse (2003) demonstrated that belief in tests affects overall fairness perceptions, job relatedness perceptions, and scientific value perceptions of different selection procedures, as measured prior to the selection process. Similarly, we hypothesize that belief in tests will be positively related to applicant's pretest reactions (Hypothesis 2).

The effects of pretest reactions on test performance. Several researchers have suggested that applicants who hold negative reactions to selection tests have lower test-taking motivation and, therefore, perform poorer than participants who hold positive reactions to selection tests (Arvey et al., 1990; Chan et al., 1997; Chan, Schmitt, Sacco et al., 1998). Past studies have provided empirical evidence for the relationship between test-taking motivation and test performance (Arvey et al., 1990; Chan et al., 1997; Sanchez, Truxillo, & Bauer, 2000). Chan et al. (1997) demonstrated

that face validity perceptions influence test-taking motivation, which in turn affects test performance. The effects of face validity perceptions on test performance were fully mediated by test-taking motivation. However, Chan, Schmitt, Sacco et al. (1998) did not include test-taking motivation in their model, neither did they examine this variable. Instead, they tested the direct effects of pretest reactions on test performance and assumed test-taking motivation to explain the relationship between pretest reactions and test performance. In the present study, we will examine this assumption and assess the indirect effects of pretest reactions on test performance via test-taking motivation. We hypothesize that test-taking motivation will fully mediate the relationship between pretest reactions and test performance (Hypothesis 3).

The effects of test performance on posttest reactions. When applicants develop perceptions after completing a selection test, it is possible that their (perceived) test performance influences these perceptions (Ployhart & Harold, 2004). Posttest reactions, thus, in addition to test characteristics may reflect the performance of the applicant on the test as a result of a self-serving bias (Chan, Schmitt, Sacco et al., 1998). A self-serving bias is an individual's tendency to attribute success to one's own personal dispositions and failure to external factors (D. T. Miller, 1978). According to Fiske and Taylor (1991) surprising, stressful, novel, unfavorable and important events trigger an individual to use attributional heuristics, such as the self-serving bias. As these conditions are inherent in selection contexts (Ployhart & Ryan, 1997), attributional heuristics are likely to play a large role in selection contexts. Several researchers support the importance of attributions in selection contexts (e.g., Arvey et al., 1990; Chan, Schmitt, Jennings, Clause, & Delbridge, 1998; Chan, Schmitt, Sacco et al., 1998; Kluger & Rothstein, 1993; Ployhart & Harold, 2004; Ployhart & Ryan, 1997).

Chan, Schmitt, Sacco et al. (1998) examined the self-serving bias effect by testing the association between actual test performance and posttest reactions. However, applicants are often unable to assess their own performance on selection tests accurately (e.g., Macan et al., 1994; Ryan & Ployhart, 2000; Truxillo, Seitz, & Bauer, 2008). For this reason, we believe that actual test performance can only indirectly influence posttest reactions, namely via self-assessed test performance. Therefore, in line with the self-serving bias perspective, we hypothesize that self-assessed test performance will fully mediate the relationship between actual test performance and posttest reactions (Hypothesis 4).

Method

Participants and procedure

Data were collected among applicants for different career moves in various organizations, who went through a psychological assessment at a Dutch HRD consultancy firm. The content of the psychological assessment procedure depends on the type of job or promotion the applicant applies for, although every psychological assessment

contains a cognitive ability test and a personality questionnaire. When applicants' psychological assessment contained an in-basket exercise, they were asked to participate in the present study. These applicants were informed that the HRD consultancy firm wishes to improve their tests and services and therefore needed help from applicants by filling out a questionnaire before and after the completion of the in-basket exercise. Applicants were guaranteed that study participation was voluntary and that their responses on the questionnaires would be treated confidentially and would not influence the course of their psychological assessment. No applicants refused to participate in the study.

After the instruction for the in-basket exercise, participants filled out a questionnaire containing items regarding demographic characteristics, control variables (experience with e-mail software programs and computer skills), belief in tests, test-taking motivation, face validity perceptions, predictive validity perceptions, and fairness perceptions. The choice between the paper-and-pencil version and the computerized version of the in-basket exercise was based on the office location where the assessment was conducted. Random assignment was not possible, because applicants who were tested at the same location could be competitors for the same job or promotion. As the equivalence of both versions of the in-basket exercise was not yet determined, random assignment would have been unfair. Immediately after participants completed the in-basket exercise, they filled out a questionnaire containing items regarding self-assessed test performance, face validity perceptions, predictive validity perceptions, and fairness perceptions.

The total sample consisted of 205 participants, of whom 106 participants (67 male, 39 female) completed the paper-and-pencil version of the in-basket exercise and 99 participants (50 male, 49 female) completed the computerized in-basket exercise. Ages ranged from 23 to 57 years (M = 38.0, SD = 8.40). Most participants had a higher vocational bachelor's degree (45.9%) and had more than 10 years of work experience (54.1%). Seventy participants (34.1%) indicated that they had experience with psychological assessments. Thirty-seven (18.0%) participants had experience with paper-and-pencil in-basket exercises and seven (3.4%) participants had experience with a computerized in-basket exercise. The participants had an average experience with the use of e-mail software programs of 9.3 years (SD = 4.3) and 94% of the participants used their e-mail software program more than once on a daily basis. Participants rated their computer skills with an average of 3.94 (SD = 0.78) on a five-point scale, ranging from 1 = *very poor* to 5 = *very well.*

Independent sample t-tests demonstrated that the participants who completed the paper-and-pencil version of the in-basket exercise did not significantly differ from the participants who completed the computerized in-basket exercise in terms of gender, age, educational level, years of working experience, experience with psychological assessments, experience with the use of e-mail software programs, frequency with which they use e-mail software programs, computer skills, cognitive ability, and Big Five personality dimensions.

Measures

In-basket exercise. The in-basket exercise was a simulation exercise of managerial daily work activities designed by the HRD consultancy firm. Within 70 minutes, participants have to read and respond to 14 memos or e-mails that were addressed to a general manager of a cleaning service company. The memos or e-mails cover a broad range of problems, such as unsatisfied customers regarding the provided services, invoices, or problematic behavior of employees. Participants had access to information about employees and clients, organizational charts, policy guidelines, and a calendar. Participants received a carefully constructed set of instructions containing information about their role as general manager and how to respond to the memos or e-mails. In the paper-and-pencil in-basket exercise participants had to respond to each memo by writing their responses on the documents. The computerized in-basket exercise resembles an e-mail software program. The e-mail software program contains a calendar, an inbox with the 14 e-mails, and an intranet environment where the information about the organization can be found. Participants had to respond to each e-mail message by sending a reply or by forwarding the e-mail message.

The second author of this manuscript rated the participants' responses to each letter or e-mail using detailed scoring sheets. The responses were scored on three managerial competencies (prioritizing tasks, analyzing and evaluating information, and making judgments) using a five-point scale, ranging from 1 = *very poor demonstration of this skill* to 5 = *very good demonstration of this skill*. The competency scores were summed and divided by the total number of responses that assessed the particular competency, resulting in three competency scores, ranging from 1 to 5. The mean score of these three competency scores represented a participant's total test score. To provide an estimate of the reliability of the ratings for this single judge, two (female) subject matter experts, who worked at the HRD consultancy firm, independently scored the responses of five randomly chosen paper-and-pencil in-basket exercises and five randomly chosen computerized in-basket exercises using the same detailed scoring sheets. Inter-rater reliabilities, as indexed by a two-way random effects intra-class correlation (*ICC*), were fair according to the classification of Fleiss (1986). For the paper-and-pencil in-basket exercise *ICCs* varied between .62 for prioritizing tasks and .67 for analyzing information. For the computerized in-basket exercise *ICCs* varied between .64 for making judgments and .72 for analyzing information.

Applicant reactions measure. Participants filled out the pretest questionnaire after the test instructions, but before they started the actual in-basket exercise. The pretest questionnaire contained items regarding control variables, belief in tests, test-taking motivation, face validity perceptions, predictive validity perceptions, and fairness perceptions. Participants filled out the posttest questionnaire immediately after they had completed the actual in-basket exercise. The posttest questionnaires contained items regarding self-assessed test performance, face validity perceptions, predictive validity perceptions, and fairness perceptions. Identical items were used in

the pretest questionnaire and in the posttest questionnaire to measure face validity perceptions, predictive validity perceptions, and fairness perceptions. Participants rated all items on a five-point scale, ranging from 1 = *strongly disagree* to 5 = *strongly agree*.

The same item as Chan, Schmitt, Sacco et al. (1998) was used to assess participants' belief in tests. The item is the following: '*I think that employment selection tests are a good way of selecting people into jobs*'. Chan, Schmitt, Sacco et al. adopted this item from the 3-item Belief in Tests Scale developed by Arvey et al. (1990). Participants' motivation to do well on the test, was measured with three items adopted from the Motivation sub-scale of Arvey et al.'s (1990) Test Attitude Survey (TAS). An example of an item is: '*I will try to do the very best I can on this in-basket exercise*'. Face validity was measured with three items adopted from Smither et al. (1993). An example of an item is: '*I think that the actual content of the test is related to the job*'. Perceived predictive validity was measured with three items adopted from Smither et al. (1993). An example of an item is: '*I think that the test can predict how well an applicant will perform on the job*'. Participants' fairness perceptions were measured with three items adopted from Gilliland (1992). An example of an item is: '*I feel that using this test to select applicants is fair'*. Self-assessed test performance was measured with three items, based on the scale of Wiechmann and Ryan (2003). An example of an item is: '*I think I have performed well on the test*'.

Results

Preliminary results

Table 1 presents the means, standard deviations, scale reliabilities, and intercorrelations of the pretest reactions, posttest reactions, and test scores for both versions of the in-basket exercise. Before we tested our hypotheses, we first looked at significant differences in test performance on the paper-and-pencil in-basket exercise and the computerized in-basket exercise. Independent sample t-tests showed that the performance of the participants who completed the paper-and-pencil in-basket exercise was significantly higher than the performance of the participants that completed the computerized in-basket exercise on the prioritizing tasks scale (M = 3.20, SD = 0.38 and M = 3.01, SD = 0.45 respectively, t = 3.25, p < .01), the analyzing information scale (M = 3.27, SD = 0.44 and M = 2.99, SD = 0.43 respectively, t = 4.70, p < .01), and the total score (M = 3.24, SD = 0.34 and M = 3.06, SD = 0.39 respectively, t = 3.54, p < .01). The self-assessed performance of the participants who completed the paper-and-pencil version (M = 3.49, SD = 0.68) was also significantly higher than the self-assessed performance of the participants who completed the computerized version of the in-basket exercise (M = 3.18, SD = 0.82, t = 3.78, p < .01).

Main results

Our first hypothesis, which stated that applicants taking the computerized in-basket exercise would perceive the test as more face valid, more predictively valid, and more fair than applicants taking the paper-and-pencil in-basket exercise, was

tested with independent sample t-tests. Results showed that pretest face validity perceptions (M = 3.69, SD = 0.69), pretest predictive validity perceptions (M = 2.86, SD = 0.63), and pretest fairness perceptions (M = 3.19, SD = 0.62) regarding the paper-and-pencil in-basket exercise did not differ from pretest face validity perceptions (M = 3.69, SD = 0.66, t = -0.05, *ns*), pretest predictive validity perceptions (M = 2.72, SD = 0.61, t = 1.61, *ns*), and pretest fairness perceptions (M = 3.12, SD = 0.57, t = 0.78, *ns*) regarding the computerized in-basket exercise. Also, posttest face validity perceptions (M = 3.47, SD = 0.74) and posttest fairness perceptions (M = 3.05, SD = 0.58) regarding the paper-and-pencil in-basket exercise did not significantly differ from posttest face validity perceptions (M = 3.51, SD = 0.65, t = 1.61, *ns*) and posttest fairness perceptions (M = 2.93, SD = 0.67, t = 1.61, *ns*) regarding the computerized in-basket exercise. We did find a significant difference between posttest predictive validity perceptions (M = 2.75, SD = 0.66) regarding the paper-and-pencil in-basket exercise and posttest predictive validity perceptions (M = 2.54, SD = 0.67, t = 2.30, p < .01) regarding the computerized in-basket exercise. However, this difference was in the opposite directions of our expectations. Based on these results, our first hypothesis could not be supported.

Model testing

The relationships in our hypothesized model were tested separately for the two in-basket exercises using structural equation modeling (AMOS 16.0, Arbuckle, 2007). We used several indices to judge the fit of the model to our data, including the Chi-square test. Although the Chi-square test is the most widely used measure of model fit in organizational research (e.g., Kelloway, 1996), it is also highly sensitive to sample size (Jöreskog & Sörbom, 1993). Hence, we used a number of alternative fit indices, namely the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the standardized root mean square residual (SRMR), and the Root Mean Square Error of Approximation (RMSEA). CFI and TLI values of .95 or higher, SRMR values of .08 or less, and RMSEA values of .06 or less indicate a relatively good fit between the hypothesized model and the observed data (Hu & Bentler, 1999), whereas CFI and TLI values of .90 or higher (Marsh, Hau, & Wen, 2004), SRMR values of .10 or less (Hu & Bentler, 1995), and RMSEA values of .08 or less (MacCallum, Browne, & Sugawara, 1996) indicate an acceptable fit.

The hypothesized model provided an acceptable fit to the data regarding the paper-and-pencil in-basket exercise, χ^2 = 139.82, df = 96, p < .01, CFI = .93, TLI = .91, SRMR = .09, RMSEA = .07. Regarding the computerized in-basket exercise, the hypothesized model provided a good fit to the data, χ^2 = 121.26, df = 96, p = .04, CFI = .96, TLI = .95, SRMR = .07, RMSEA = .07. Figures 2 and 3 present the full models associated with the paper-and-pencil in-basket exercise and the computerized in-basket exercise with the standardized parameter estimates and structural parameter estimates, respectively. In both models, all estimated path coefficients and factor loadings were significant (p < .05).

Table 1

Means, Standard Deviations, Scale Reliabilities, and Intercorrelations of Pretest Measures, Posttest Measures, and Test Performance for the Paper-and-Pencil and the Computerized In-Basket Exercise

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13
Pretest measures															
1. Belief in tests	3.80/3.72	0.70/0.57	(-)	.14	.20*	.34**	.33**	.23*	.22*	.24*	.06	.02	.07	.05	.05
2. Test-taking motivation	4.57/4.55	0.43/0.45	.15	(.67/.73)	.24*	.17	.08	.13	-.02	.06	-.03	.26**	.15	.22*	.26**
3. Face validity	3.69/3.69	0.69/0.66	.41**	.47**	(.78/.81)	.38**	.24*	.61**	.32**	.25*	.07	.02	-.09	.02	-.02
4. Predictive validity	2.86/2.71	0.63/0.61	.21*	.11	.49**	(.69/.68)	.60**	.42**	.57**	.31**	.16	.08	.14	.02	.09
5. Fairness	3.19/3.12	0.62/0.57	.20*	.09	.32**	.59**	(.62/.52)	.32**	.46**	.47**	.09	.01	-.08	-.10	-.07
Posttest measures															
6. Face validity	3.47/3.51	0.73/0.65	.20*	.30**	.58**	.36**	.34**	(.85/.74)	.52**	.44**	.18	.20*	.06	.04	.12
7. Predictive validity	2.75/2.54	0.66/0.67	.19*	.08	.27**	.61**	.59**	.42**	(.72/.74)	.57**	.29**	.19*	.15	.11	.19*
8. Fairness	3.05/2.93	0.58/0.67	.18	.03	.25*	.49**	.67**	.40**	.62**	(.49/.66)	.18	.10	.07	-.03	.05
Test performance															
9. Self-ass. performance	3.49/3.18	0.68/0.82	-.07	.15	.14	-.04	-.01	.28**	.17	.15	(.86/.89)	.23*	.24*	.30**	.31**
10. Prioritizing tasks	3.20/3.01	0.38/0.45	.01	.15	.00	-.07	-.04	.08	.08	-.01	.29**	(.71/.74)	.61**	.57**	.87**
11. Analyzing information	3.27/2.99	0.44/0.43	.05	.03	.09	-.13	-.04	.15	.03	-.01	.26**	.57**	(.72/.65)	.42**	.79**
12. Making judgments	3.26/3.19	0.40/0.52	.05	.28**	.14	-.04	-.12	.19*	.03	-.02	.23*	.58**	.53**	(.62/.71)	.82**
13. Total test score	3.24/3.06	0.34/0.39	.04	.17	.09	-.10	-.08	.17	.05	-.02	.31**	.84**	.85**	.83**	(.79/.77)

Note. Left from the diagonal are the correlations for the paper-and-pencil measure, right from the diagonal are the correlations for the computerized measure. Scale reliabilities (coefficient alpha) are presented on the diagonal, between parentheses. When two means, standard deviations, or scale reliabilities are presented, the first is for the paper-and-pencil measure (N = 106) and the second for the computerized measure (N = 99). All scales range from 1-5.

* $p < .05$, ** $p < .01$

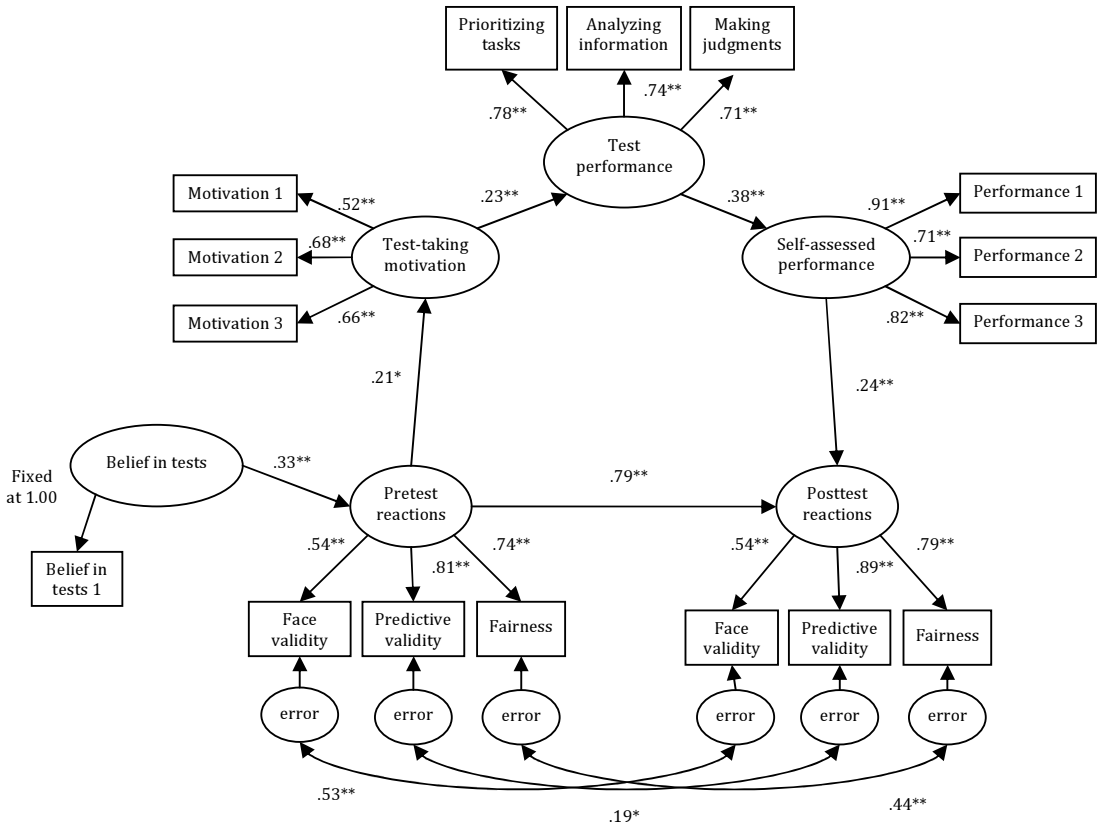


Figure 2. Full structural model with standardized measurement and structural parameter estimates for applicant reactions regarding the paper-and-pencil in-basket exercise.

Consistent with Hypothesis 2, belief in tests had a significant direct effect on pretest reactions regarding the paper-and-pencil in-basket exercise ($\gamma = .33$, $p < .01$) and pretest reactions regarding the computerized in-basket exercise ($\gamma = .43$, $p < .01$). Based on these results, our second hypothesis could be supported.

To test whether test-taking motivation would fully mediate the relationship between pretest reactions and test performance (Hypothesis 3), bootstrapping procedures in AMOS 16.0 were used. Bootstrapping procedures have been recommended to assess mediation effects with small to moderate samples (Preacher & Hayes, 2004; Shrout & Bolger, 2002). By extracting 1000 bootstrapped samples from the dataset based on random sampling with replacement, 90% confidence intervals (CIs) were calculated. Significant indirect effects of pretest reactions on both the performance on the paper-and-pencil in-basket exercise (*estimate* = .04, *SE* = .03, *lower CI* = .01, *higher CI* = .22, $p < .05$) and the performance on the computerized in-basket exercise (*estimate* = .07, *SE* = .06, *lower CI* = .01, *higher CI* = .21, $p < .05$) were found. No significant direct effects of pretest reactions on both the performance on the paper-and-pencil in-basket exercise (*estimate* = -.09, *SE* = .12, *lower CI* = -.28, *higher CI* = .06, $p < .05$) and the performance on the computerized in-basket exercise (*estimate* = .01,

SE = .16, *lower CI* = -.22, *higher CI* = .23, *ns*) were found, after controlling for test-taking motivation. Taken together, these findings are consistent with the full mediation effect of Hypothesis 3.

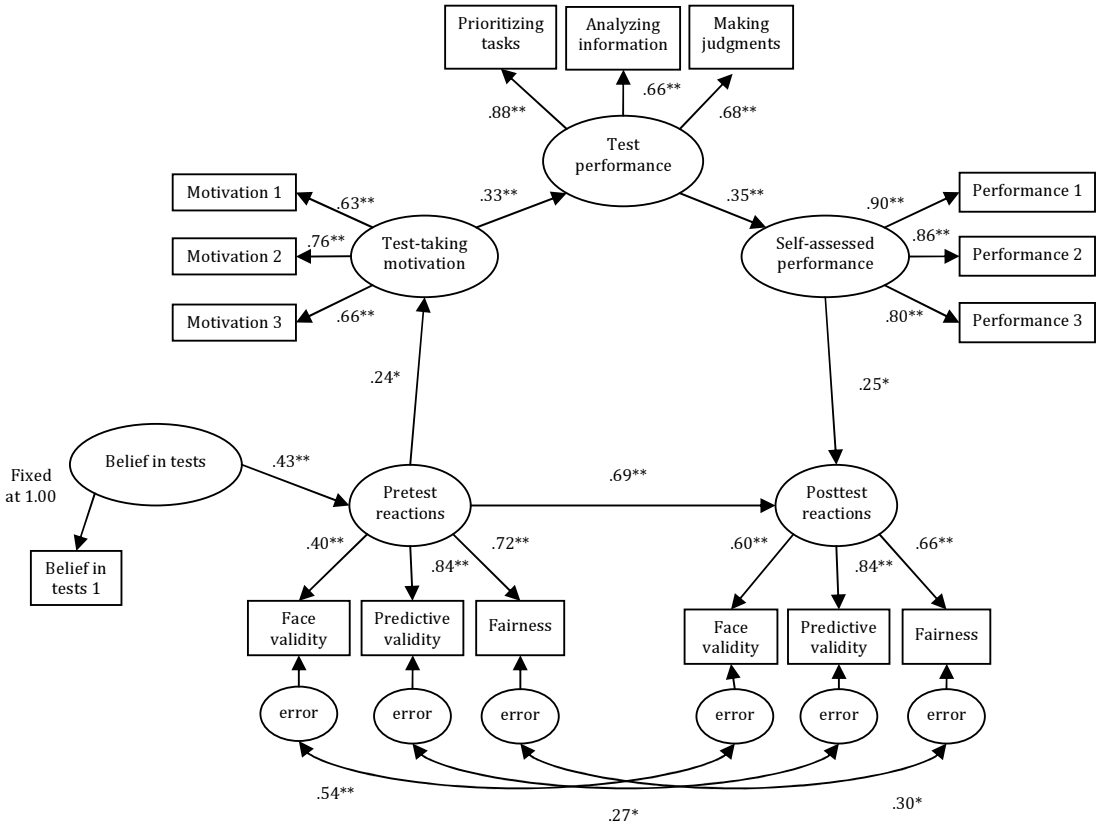


Figure 3. Full structural model with standardized measurement and structural parameter estimates for applicant reactions regarding the computerized in-basket exercise.

Our last hypothesis stated that that self-assessed test performance would fully mediate the relationship between actual test performance and posttest reactions (Hypothesis 4). As before, 90% CIs were calculated by extracting 1000 bootstrapped samples from the dataset. A significant indirect effect of test performance on both the posttest reactions regarding the paper-and-pencil in-basket exercise (*estimate* = .14, *SE* = .08, *lower CI* = .04, *higher CI* = .29, $p < .01$) and the posttest reactions regarding the computerized in-basket exercise (*estimate* = .07, *SE* = .05, *lower CI* = .01, *higher CI* = .18, $p < .05$) was found. After including self-assessed test performance as mediator in the relationship between test performance and posttest reactions, the direct effect of test performance on both the posttest reactions regarding the paper-and-pencil in-basket exercise (*estimate* = .03, *SE* = .16, *lower CI* = -.22, *higher CI* = .30, *ns*) and the posttest reactions regarding the computerized in-basket exercise (*estimate* = .13, *SE* =

.13, *lower CI* = -.08, *higher CI* = .36, *ns*) becomes non-significant, implying that self-assessed test performance fully mediates the relationship between actual test performance and posttest reactions. Based on these results, Hypothesis 4 could be supported.

Discussion

The first aim of this study was to compare pretest and posttest reactions regarding a paper-and-pencil version to pretest and posttest reactions regarding a computerized version of one of the most widely used assessment center exercises, namely an in-basket exercise. The second aim of this study was to examine the nature of pretest and posttest reactions to an in-basket exercise, by drawing upon the model of Chan, Schmitt, Sacco et al. (1998). The results of the present study indicated that, contrary to our expectations, the paper-and-pencil version was equally or better perceived than the computerized version of the in-basket exercise. The hypothesized model regarding the nature of applicant reactions could be confirmed. That is, pretest reactions partly reflected applicant's general beliefs about tests and posttest reactions partly reflected the performance of the applicant on the test. Each of these findings will now be discussed in more detail.

First, the expectation was that applicants taking the computerized in-basket exercise would perceive the test as more face valid, more predictively valid, and more fair than applicants taking the paper-and-pencil in-basket exercise (Hypothesis 1). Contrary to this expectation, results showed no significant differences between reactions to the paper-and-pencil in-basket exercise and the computerized in-basket exercise. However, posttest predictive validity perceptions did differ between the two test versions, but in the opposite direction of the hypothesis. Participants who completed the paper-and-pencil in-basket exercise perceived the test as more predictively valid than the participants that completed the computerized in-basket exercise. These results are not in line with previous studies that have demonstrated that computerized tests are equally or better perceived than their paper-and-pencil counterparts (e.g., Mead, 2001; Potosky & Bobko, 2004; Reynolds et al., 2000; Salgado & Moscoso, 2003).

It is possible that a difference in difficulty of our in-basket exercises has affected our results. Participants who completed the computerized version scored significantly lower on two of the three scales of the in-basket exercise, indicating that the computerized version was more difficult than the paper-and-pencil version. It seemed that the computerization of the in-basket exercise negatively affected participants' test performance, because a number of activities required more time in the computerized version than in the paper-and-pencil version, such as switching between different computer screens to examine the various resource materials (charts, diagrams, calendars, etc.), or learning the different functions of the e-mail software program. On forehand, we did not expect that computerization would adversely affect performance, because a number of reviews concerning the equivalence of computerized versions and paper-and-pencil versions of the same test has

shown that there was no problem associated with computerization of non-cognitive tests (e.g., Mead & Drasgow, 1993; King & Miles, Bartram, 1994). Because for almost all applicants the time limit of 70 minutes was long enough to address each memo in the paper-and-pencil in-basket exercise, we presumed that this time limit would also be sufficient to address each e-mail in the computerized in-basket exercise.

As our results demonstrated, actual test performance affects self-assessed performance. Therefore, a self-serving bias could explain our finding that predictive validity perceptions regarding the computerized version were significantly lower than the predictively valid perceptions regarding the paper-and-pencil version of the in-basket exercise. Participants could have attributed their self-assessed performance to the predictive validity of the exercise. More and more traditional paper-and-pencil tests are being computerized, because of the assumed economic and practical benefits (Chan & Schmitt, 2004; Wiechmann & Ryan, 2003). However, when computerization adversely affects test performance and subsequently adversely affects applicant reactions, these assumed benefits are not realized. We therefore recommend examining the equivalence of computerized tests and paper-and-pencil tests in terms of test performance, before studying the effects of computerization on applicant reactions.

Furthermore, we expected that belief in tests would be positively related to pretest reactions (Hypothesis 2). In line with previous studies (Chan, Schmitt, Sacco et al., 1998; Lievens et al., 2003; Van Vianen et al., 2004), our results demonstrated a strong direct effect of belief in tests on pretest reactions regarding both the paper-and-pencil in basket exercise and the computerized in-basket exercise. Thus, in addition to test characteristics, pretest reactions seem to reflect applicant's general beliefs in tests. Future research is needed to explore why general test beliefs so strongly affect pretest reactions. Factors that have been suggested to affect belief in tests are prior test experience (Wiechmann & Ryan, 2003) and an applicant's performance history (Ryan & Ployhart, 2000).

Results further showed that test-taking motivation fully mediated the relationship between pretest reactions and test performance on both the paper-and-pencil in-basket exercise and the computerized in-basket exercise (Hypothesis 3). In the model of Chan, Schmitt, Sacco et al. (1998), test-taking motivation was assumed to fully explain the relationship between pretest reactions and test performance. The present study provided empirical support for this assumption. The findings are in line with Chan et al. (1997), who demonstrated that face validity perceptions influence test-taking motivation, which in turn affects test performance. As an effect of test-taking motivation on test performance affects the construct validity of a test, the results of the present study may have important practical implications. If face validity perceptions, predictive validity perceptions, and fairness perceptions affect test-taking motivation, then test constructors have a means of controlling individual difference in test-taking motivation. For example, by constructing more realistic test items, as the realism of test items has been suggested to influence applicants' test-taking motivation (Bauer & Truxillo, 2006).

Although we found a high correlation between pretest and posttest reactions for both the paper-and-pencil in-basket exercise and the computerized in-basket exercise, pretest and posttest reactions seem to have different causes and effects. As mentioned above, pretest reactions seem to be affected by applicants' belief in tests, whereas posttest reactions seem to be affected by test performance. Our results demonstrated that performance indirectly affected posttest reactions via self-assessed test performance for both the paper-and-pencil in-basket exercise and the computerized in-basket exercise (Hypothesis 4). In addition to test characteristics, posttest reactions thus seem to reflect the performance of applicants on the test as a result of a self-serving bias. These findings are in line with the literature, as several researchers demonstrated the importance of attributions, such as the self-serving bias in selection contexts (e.g., Arvey et al., 1990; Chan, Schmitt, Jennings et al., 1998; Chan, Schmitt, Sacco et al., 1998; Kluger & Rothstein, 1993; Ployhart & Harold, 2004; Ployhart & Ryan, 1997). The present study demonstrated that the relevant test for the self-serving bias is the relationship between self-assessed test performance and posttest reactions, as self-assessed test performance fully mediated the relationship between actual test performance and posttest reactions. We therefore recommend future studies on applicant reactions and attributional heuristics in selection contexts to include a measure of self-assessed test performance.

The hypothesized model regarding the nature of the three most commonly studied dimensions of applicant pretest and posttest reactions and their relationship with test performance applied equally regardless of test medium (paper-and-pencil versus computerized test). Differences in test performance and perceived predictive validity do not seem to have affected the pattern of relationships between pretest reactions, (self-assessed) test performance, and posttest reactions. We believe this is a positive finding, regarding the generalizability of our results.

Limitations of this study and suggestions for future research

The present study has some general limitations that should be noted. First, we only measured posttest reactions before the participants received feedback on their test scores. These posttest reactions have been found to be related to important outcomes, such as organizational attractiveness, general perceptions of testing fairness, and applicants' general test-taking self-efficacy (Bauer et al., 1998). However, feedback has been found to also influence applicant reactions about the selection process (Bauer et al., 1998; Truxillo, Bauer, Campion, & Paronto, 2002). Therefore, we recommend future studies to also measure applicant reactions after participants have received feedback on their test scores. These reactions are important, as they are related to long-term outcomes, such as applicant withdrawal from the selection process (Ryan & Ployhart, 2000).

A second limitation is that participants could not be fully randomly assigned to the two conditions. Participants were assessed by means of either the paper-and-pencil version or the computerized version of the in-basket exercise, depending on the location of the HRD consultancy firm's office. Unfortunately random assignment was

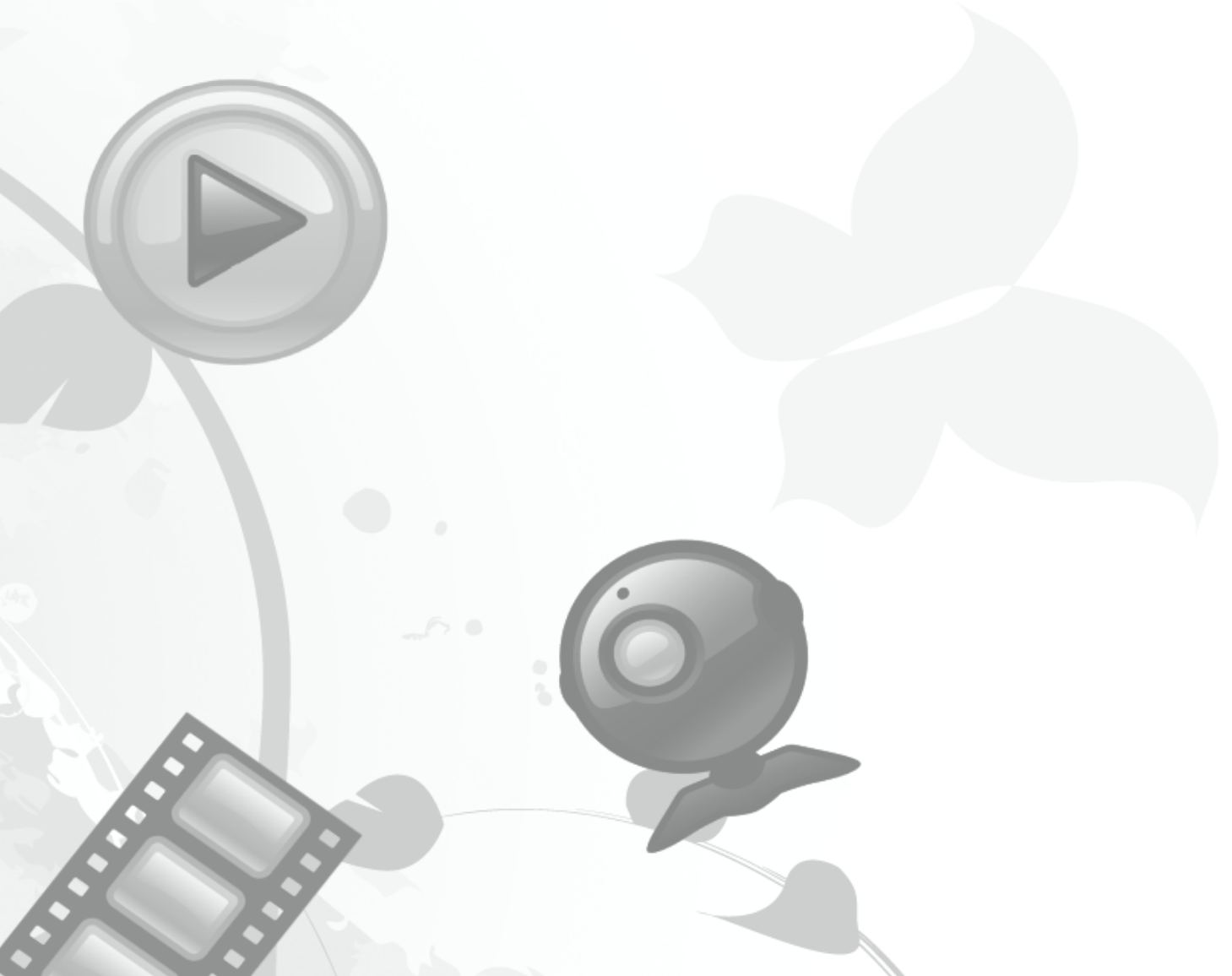
not possible, as it would have been unfair to assess competitors for the same job or promotion with different versions on the in-basket exercise. Although comparisons of demographic characteristics, experience with the use of e-mail software programs, computer skills, cognitive ability, and personality yielded no significant differences between participants in the two conditions, it would have been preferable to have randomly assigned equivalent groups, because this would have allowed stronger inferences about the equivalence of the applicant reactions, and would have reduced alternative explanations for the results (Peterson, Kolen, & Hoover, 1989).

Another potential limitation to the study was the measurement of belief in tests and fairness perceptions. Belief in tests was measured with a single item, adopted from the study of Chan, Schmitt, Sacco et al. (1998). Chan, Schmitt, Sacco et al. formulated this item in such a way that it captured the central issue regarding applicants' general beliefs in employment tests. Nevertheless, using more items to measure belief in tests could provide a broader representation of the construct and a possibility to analyze the internal consistency. The internal consistency of the fairness perceptions scale was relatively low (scale reliabilities varied between .49 and .66), which could have attenuated the effects of the pretest and posttest reactions on other study variables, such as test-taking motivation. It is possible that the items of our fairness measure were too generic. For example, the item '*I feel that using this test to select applicants is fair*' could refer to different domains. More domain-specific level items, such as '*I feel that using this test to select applicants is fair to both males and females*' may results in a more reliable measure. Future studies should therefore use a more domain-specific measure of fairness perceptions.

Despite these potential limitations, we believe the current study contributes to our knowledge on applicant reactions in several ways. We demonstrated that pretest reactions and posttest reactions are influenced by different external variables. Pretest reactions were influenced by applicants' general beliefs in tests, whereas posttest reactions were influenced by (perceived) performance on the test. Furthermore, the current study contributes to our knowledge on the nature of applicant reactions, because an established model, comprising the determinants of the three most widely studies applicant reactions, namely face validity perceptions, predictive validity perceptions, and fairness perceptions, was tested in an actual field study. Moreover, we demonstrated that the relationships between belief in tests, pretest reactions, (self-assessed) test performance, and posttest reactions apply equally regardless of the test medium of an in-basket exercise. Whether the model is generalizable to other selection instruments should be investigated.

Chapter 7

Summary and discussion



More and more organizations make use of new technology, such as multimedia tests, in the recruitment and selection of personnel (Lievens et al., 2002). In a multimedia test applicants are usually presented with a number of challenging job-related situations. The situation then freezes at an important moment and applicants are asked to evaluate a number of courses of action by indicating how they would act in this particular situation (Weekley & Ployhart, 2006). This type of multimedia test is called a multimedia situational judgment test (SJT). Recently, another innovative multimedia test has entered personnel selection practices, namely a webcam test. A webcam test can be conceptualized as a multimedia SJT with a constructed-response item format (Arthur & Villado, 2008). In a webcam test applicants are presented with job-related situations through the use of video clips and are then asked to act out their response, while being filmed by a webcam (Lievens et al., 2008).

Although organizations have rushed to incorporate multimedia SJTs and webcam tests into their selection systems (Anderson, 2003), research regarding these type of tests still is scarce. This dissertation aimed to address this shortcoming by presenting five empirical studies on the validity and acceptability of multimedia tests. An overview of the main findings in these five chapters will be presented in the next paragraph.

Summary of Main Findings

Personnel selection tests need to be assessed against a number of criteria, including criterion-related validity, incremental validity, construct validity, and acceptability (Cook, 2009). Guided by these four criteria an overview of the main findings in this dissertation will be provided.

Criterion-related validity

This type of validity refers to the degree to which a test estimates an external criterion, such as academic or job performance (Nunnally, 1978). Chapters 2, 3, and 4 addressed the criterion-related validity of multimedia tests. In general, results were supportive of the validity of multimedia tests both as a predictor of academic performance and as a predictor of job performance.

In **chapter 2** we examined whether a webcam test for interpersonally oriented leadership skills was able to predict academic performance, which was conceptualized as students' grade point average (GPA) and students' observed learning activities, such as how well they perform their role as a chair during group meetings. As previous studies had showed that SJTs are more predictive of interpersonally oriented criteria than of cognitively oriented criteria (e.g., Lievens & Sackett, 2006; Oswald et al., 2004), it was hypothesized that scores on a webcam test for interpersonally oriented leadership skills would have higher validity for predicting students' observed learning activities than for GPA. Data were collected among 153 psychology students. Results supported the validity of the webcam test as predictor of academic performance. In particular, scores on the webcam test predicted students' participa-

tion during group meetings, how well they performed their role as a chair during the group meetings, their preparation for these meetings, and the observed learning activities in general. As expected, the webcam test showed higher validity for predicting students' observed learning activities than for GPA. These findings suggest that a multimedia SJT with a constructed-response item format can be a valid predictor of academic performance criteria.

Chapter 3 concerned the criterion-related validity of a webcam test that was intended to measure effectiveness in the core task of an employment consultant, namely advising job seekers. This first field study on a webcam test was conducted in an employment agency. The sample consisted of 188 consultants who participated in a certification process, which consisted of an assessment through a webcam test, a job knowledge test, a measure of objective job placement success of the consultants' clients, and a manager's appraisal of the consultants' job performance. It was hypothesized that scores on the webcam test would be positively related to job placement success and to the manager's appraisal. Results partly supported the validity of the webcam test as predictor of job performance. Scores on the webcam test predicted the job placement success criterion, but not the manager's appraisal. This latter finding may be explained by the fact that the managers were aware of the fact that their appraisal was a part of the certification procedure. This may have led to a leniency in their judgments, which in turn, may have affected the criterion-related validity.

In **chapter 4** we examined whether implicit trait policies (ITPs) as measured with a multimedia SJT for leadership were able to predict employees' observable workplace behaviors. ITPs are implicit beliefs of individuals about the effectiveness of different levels of trait expression (Motowidlo et al., 2006b). For instance, an individual may believe that the expression of agreeableness in SJT response options is generally very effective. SJTs capture ITPs by correlating applicants' effectiveness ratings of the different response options with the level of trait expression of these response options. Motowidlo and Beier (2010) demonstrated that ITPs as measured with an SJT are able to predict a composite measure of job performance. Similarly to Motowidlo and Beier, the aim in chapter 4 was to shed light on the predictive validity of ITPs. However, in contrast to the study of Motowidlo and Beier and other studies on the predictive validity of ITPs, a construct-driven multimedia SJT was used. A construct-driven SJT has several advantages, namely that the validity of the SJT is expected to generalize across jobs and that it provides the opportunity to conceptually align the predictor and criterion domain (Lievens, 2006). The multimedia SJT was developed to predict leadership skills and the response options expressed either high or low levels of extraversion and agreeableness. It was hypothesized that ITPs for extraversion and ITPs for agreeableness would be more strongly related to participants' leadership behaviors, which is a conceptually aligned criterion, than to non-leadership behaviors. Data were collected among 180 assessment candidates. Results demonstrated that ITPs for extraversion predicted peer ratings and supervisor ratings of leadership behaviors, and that they indeed showed more validity for

predicting leadership behaviors than for non-leadership behaviors. However, no significant correlation was found between ITPs for agreeableness and leadership behaviors. The multifaceted relationship between agreeableness and leadership possibly explains why ITPs for agreeableness were unable to predict ratings of leadership behavior.

Incremental validity

This form of criterion-related validity refers to whether a selection test adds to the prediction of a criterion above what is predicted by other selection tests (Hunsley & Meyer, 2003). Chapters 2, 3, and 4 addressed the incremental validity of multimedia tests. In general, results were supportive of the incremental validity of multimedia tests in both the prediction of academic performance and job performance.

In **chapter 2** it was examined whether a webcam test for interpersonally oriented leadership would incrementally predict students' observed learning activities over and above a cognitive ability test and a personality questionnaire. A large body of research has established measures of cognitive ability and personality to be important predictors of academic success (e.g., Lounsbury et al., 2003; Poropat, 2009). The value of the webcam test would therefore increase if it showed incremental validity over these traditional predictors. Results showed that the webcam test was able to explain a unique part of variance in academic performance, which demonstrated that a webcam test can be a useful and valid complement to traditional predictors in selection contexts.

Several authors have argued that situational tests, such as a webcam test, owe some of their criterion-related validity to their assessment of job knowledge (e.g., McDaniel & Nguyen, 2001). Therefore, in **chapter 3** it was studied whether a webcam test intended to measure effectiveness in advising job seekers was able to explain unique variance in job performance over and above job knowledge. Results demonstrated that the webcam test incrementally predicted job placement success over and above a job knowledge test, suggesting the webcam test measures more than just job knowledge. Regression analyses also demonstrated that the unemployment rate of the province in which the consultant worked was significantly related to job placement success. Controlling for this effect of unemployment rate, and for age, gender, job tenure and the job knowledge test, the webcam test still was able to explain additional variance in job placement success. This finding confirms that the webcam test is a relevant additional predictor of job performance.

Chapter 4 investigated whether ITPs as measured with a multimedia SJT for leadership were able to incrementally predict observed leadership behaviors over and above personality scale scores and leadership experience. According to ITP theory, personality and experience have a causal effect on ITPs (Motowidlo et al., 2006a). Both personality and leadership experience have been found to be positively related to leadership behavior (e.g., Judge, Bono et al., 2002; Thomas & Cheese, 2005). Therefore it was investigated whether the relationship between ITPs and observed leadership behavior could be solely attributed to the causal effects of personality and

leadership experience on ITPs, or whether ITPs could explain unique variance in observed leadership behavior beyond the variance explained by personality traits and leadership experience. Results demonstrated that ITPs for extraversion were able to explain unique variance in peer ratings of leadership behavior and supervisor ratings of leadership behavior over and above leadership experience and the personality scale score of extraversion. However, ITPs for agreeableness were not able to explain unique variance in leadership behavior over and above leadership experience and the personality scale score of agreeableness. The ambiguous relationship between agreeableness and leadership may explain why ITPs for agreeableness were unable to incrementally predict ratings of leadership behavior. On the one hand, agreeable persons are likely to be more altruistic, which is an important trait for leaders. On the other hand agreeable persons are also more modest and have more need for affiliation (Yukl, 1998). These latter facets of agreeableness are negatively related to leadership.

Construct validity

This type of validity refers to the extent to which a selection test relates to other measures consistent with theoretically derived hypotheses (Carmines & Zeller, 1979). Many researchers have called for a focus towards the processes and constructs underlying situational tests (e.g., Lievens et al., 2008). For this reason, Chapters 2 and 4 addressed the construct validity of a webcam test and a multimedia SJT respectively.

Chapter 2 examined the relationships between scores on a webcam test for interpersonally oriented leadership skills and personality, cognitive ability, and previous job experience. Previous studies had provided evidence that paper-and-pencil leadership SJTs are related to the personality traits of extraversion and conscientiousness and to cognitive ability (Bergman et al., 2006; Oswald et al., 2004). However, it was expected that a webcam test would be less strongly related to cognitive ability than a paper-and-pencil SJT with a multiple-choice format. The arguments for this expectation are that a webcam test has no reading component and that its constructed-response item format measures participants' actual interpersonally oriented skills in job-related situations (Motowidlo et al., 2008). Therefore, it was hypothesized that scores on a webcam test for interpersonally oriented leadership skills would be more strongly related to the personality traits of extraversion and conscientiousness than to cognitive ability. Results were in line with this expectation. Furthermore, results showed that webcam test scores were related to leadership experience. People with job relevant experiences are probably more likely to have encountered the types of job-related situations presented in the webcam test and therefore may have learned how to respond successfully to these types of situations.

According to the ITP theory of Motowidlo et al. (2006b), individual differences in personality traits affect judgments of behavioral expressions in an SJT. ITPs may therefore implicitly measure personality traits. In **chapter 4** it was examined whether a multimedia SJT for leadership skills was able to capture individual differences in

ITPs by examining the relationship between ITPs and the associated personality scale scores and leadership experience. Specifically, it was hypothesized that ITPs for extraversion and ITPs for agreeableness as measured with a multimedia SJT for leadership skills would be positively related to leadership experience and to the personality scale scores of extraversion and agreeableness respectively. Results confirmed that a multimedia SJT for leadership skills indeed could be used as a measure of ITPs for targeted personality traits, as ITPs for extraversion and agreeableness were positively related to the personality scale scores of extraversion and agreeableness respectively. Furthermore, it was found that employees who had more experience as a leader held stronger beliefs about the effectiveness of extraversion in the leadership behaviors that were demonstrated in the SJT. However, no relationship was found between leadership experience and ITPs for agreeableness. Again, the multifaceted relationship between agreeableness and leadership may explain why employees with more experience as a leader did not hold stronger beliefs about the effectiveness of agreeableness in leadership behaviors than employees with less experience as a leader.

Acceptability

Much of the research on applicant reactions to multimedia tests has been rather descriptive and comparative in nature, rather than explanatory (e.g., Kanning et al., 2006; Richman-Hirsch et al., 2000). Chapters 5 and 6 tried to fill this void by examining the nature of the most commonly studied applicant reactions.

In **chapter 5** we examined the relationship of a number of testing-related and general individual differences with the most frequently studied dimension of applicant reactions, that is perceived job relatedness (Chan & Schmitt, 2004). Perceived job relatedness consists of two related, but distinguishable, dimensions, namely face validity and perceived predictive validity. Previous studies had shown that test content and test characteristics affect the perceived job relatedness of selection instruments (e.g., Chan & Schmitt, 1997), but still substantial variance in these perceptions remained unexplained. Among 153 psychology students it was examined whether individual differences are able to explain some of this variance in the perceived job relatedness of two often used computerized selection instruments, namely a cognitive ability test and a multimedia SJT intended to measure managerial skills. Specifically, the relationship of job relatedness perceptions with anxiety (test anxiety and computer anxiety), self-evaluations (test-taking self-efficacy, core self-evaluations, and subjective well-being), and personality (agreeableness, emotional stability, and openness to experience) were examined. Results indicated that computer anxiety, core self-evaluations, subjective well-being, agreeableness, emotional stability, and openness to experience affected the perceived job relatedness of a cognitive ability test and a multimedia SJT, but not systematically. For example, the face validity of the cognitive ability test was related to agreeableness, emotional stability, and openness to experience, while the face validity of the multimedia SJT was related to computer anxiety, core self-evaluations, subjective well-being, and

openness to experience. Openness to experience was found to be the most consistent predictor of job relatedness perceptions, implying that individuals who are more amenable to new experiences seem to react more positively to computerized selection instruments than individuals who are resistant to new experiences. These findings revealed that stable individual differences may account for a portion of variance in job relatedness perceptions, suggesting there may be a stable component to applicant reactions in addition to test-related factors.

In **chapter 6** we presented a final study in which pretest and posttest face validity perceptions, predictive validity perceptions, and fairness perceptions regarding a paper-and-pencil version and a computerized version of an in-basket exercise were compared among 205 applicants. Results showed that posttest predictive validity perceptions differed between the two test versions of the in-basket exercise. Participants who completed the paper-and-pencil in-basket exercise perceived the test as more predictively valid than the participants that completed the computerized in-basket exercise. However, a difference in difficulty of the computerized in-basket exercise and the paper-and-pencil in-basket exercise may have affected our results. Participants who completed the computerized version scored significantly lower on the in-basket exercise, indicating that the computerized version was more difficult than the paper-and-pencil version. Thus, in this study the computerization of the in-basket exercise may have adversely affected test performance and subsequently may have affected the posttest predictive validity perceptions. Yet, a comparison of the other applicant reactions yielded no significant differences between the two versions of the in-basket exercise.

Furthermore, the nature of these reactions and their relationship with test performance were examined by drawing upon the model of Chan, Schmitt, Sacco et al. (1998) on applicant reactions. In the majority of research applicant reactions are measured on a single occasion, either before the test (e.g., Rynes & Connerley, 1993; Schmit & Ryan, 1997) or after taking the test (e.g., Lievens & Sackett, 2006; Richman-Hirsch et al., 2000; Salgado & Moscoso, 2003). Yet, according to Chan, Schmitt, Sacco et al. pretest reactions and posttest reactions are influenced by different external variables. Results showed that pretest reactions and posttest reactions indeed could not be considered as interchangeable, because pretest reactions were affected by applicants' general beliefs in tests and posttest reactions were affected by applicants' test performance via self-assessed test performance.

Strengths, Limitations, and Suggestions for Future Research

The studies presented in this dissertation contribute to the literature on multimedia testing in a number of ways. First, the present dissertation presented three field studies, one among consultants in a large job centre (chapter 3), and two among assessment candidates of a large HRD consultancy firm (chapters 4 and 6). Most studies regarding the validity and acceptability of multimedia tests have been conducted among student samples (e.g., Richman-Hirsch et al., 2000; Wiechmann & Ryan, 2003). As students clearly differ from typical applicants in terms of previous

experience with selection instruments and self-presentation motives, field studies such as the presented ones remain critical to test the ecological validity of experimental research findings (Greenberg, 1990).

Second, the present dissertation provided support for the importance of J. P. Campbell's (1990) strategy of conceptually aligning predictors and criteria. In chapter 2, the predictor and criterion domain were carefully specified, as it was examined whether a webcam test intended to measure interpersonally oriented leadership was able to predict students' leadership-related behaviors, such as demonstrating skills in a group, motivating others, and coordinating groups and tasks. The webcam test showed higher validity for these observed learning activities than for GPA. In chapter 4, the predictive validity of ITPs as measured with the leadership SJT was examined by using a criterion that measures participants' leadership behavior. ITPs for extraversion showed more validity for predicting leadership behaviors than for other non-leadership behaviors. Thus, if predictors and criterion measures are matched in terms of the construct measured, selection instruments show better convergent and divergent validity.

Third, the present dissertation aimed to shed light on the nature of applicant reactions. Thus far, much of the research on applicant reactions has focused on descriptive questions, such as the comparison of favorability reactions across procedures and instruments (e.g., Hausknecht et al., 2004; Kluger & Rothstein, 1993; Rynes & Connerley, 1993). The findings in chapter 5 revealed that stable individual differences may account for a portion of variance in applicant reactions, suggesting there may be a stable component to applicant reactions in addition to test-related factors. The findings in chapter 6 revealed that pretest reactions and posttest reactions are affected by different factors. Pretest reactions were affected by applicants' general beliefs in tests, whereas posttest reactions were affected by applicants' test performance via self-assessed test performance.

Some limitations of the present dissertation are worth mentioning. First, because of the homogenous samples regarding ethnicity, it was not possible to investigate the potential adverse impact of multimedia tests. Using multimedia tests instead of paper-and-pencil tests has been suggested as one of the strategies to reduce adverse impact, because the use of multimedia reduces the reading demands, and subsequently may reduce the cognitive load of the test (Ployhart & Holtz, 2008). Future studies should examine whether multimedia tests result in less adverse impact compared to other selection instruments. We were also unable to investigate ethnicity differences in applicant reactions. Previous studies have demonstrated that applicant reactions differ across ethnic groups (e.g., Chan, 1997; Chan & Schmitt, 1997; Viswesvaran & Ones, 2004). For example, Viswesvaran and Ones (2004) found ethnic differences in the importance that was placed on different aspects of selection system characteristics that relate to fairness perceptions. Future research should examine whether these ethnicity differences also apply to applicant reactions to multimedia tests.

Second, to determine the validity of the multimedia tests, concurrent designs typically have been used. It is possible that the results from such concurrent validation

studies might not be generalizable to applicant samples. Applicants complete selection instruments in high stakes situations, which is likely to affect their motivation. In chapter 6 it indeed was found that test-taking motivation affects performance on multimedia tests. Although research has shown that there is little evidence of differences between predictive and concurrent validation designs (Barrett, Phillips, & Alexander, 1981), we recommend future studies to examine the validity of multimedia tests in actual applicant samples.

A final potential limitation worth mentioning is that in chapters 5 and 6 of the present dissertation, applicant reactions were measured before the participants received feedback on their test scores. These applicant reactions may relate to behaviors demonstrated by applicants during later stages of the selection process prior to the organization's decision (e.g., intentions to accept the job). However, because test feedback can influence applicant reactions (Bauer et al., 1998), we recommend future studies to also measure applicant reactions to multimedia tests after participants receive feedback on their test scores, as these perceptions may be related to more long-term behaviors (Ryan & Ployhart, 2000).

In future studies, it would be interesting to compare the construct validity and criterion-related validity of a multimedia SJT with a multiple choice item format with a multimedia SJT with a constructed-response item format, measuring the same construct with the same situational stimuli. By holding the predictor construct constant, conclusions then can be drawn about the effects of the response format.

Recently, there have been attempts to use 3D computer animation for the presentation of situational judgment scenarios (e.g., Hall, Fetzer, Tuzinski, & Freeman, 2010). It has been suggested that 3D animated SJTs may be even more realistic and therefore more valid than multimedia SJTs. A major advantage of using 3D animation instead of video clips is the possibility to make small alterations in the SJT scenarios without having to hire an entire film crew. Therefore, it appears to be worth examining the validity and acceptability of 3D animated SJTs.

Practical Implications

The present dissertation has demonstrated that multimedia tests can be of great value for personnel selection practices. First, it was demonstrated in chapters 2, 3 and 4 that multimedia tests are predictive of both job and academic performance. Moreover, it was demonstrated that multimedia tests are able to explain additional variance in performance over and above traditional instruments, such as personality questionnaires, cognitive ability tests, and job knowledge tests. Although, many organizations have already incorporated multimedia SJTs and webcam tests into their selection procedures, research regarding this type of instrument was running behind.

Second, in chapter 5 it was demonstrated that applicants react more positively to multimedia tests than to more traditional tests, such as cognitive ability tests. Organizations can benefit from selection instruments that generate positive applicant reactions, as previous studies have demonstrated that applicant reactions are related to intentions to accept the job, intentions to recommend the organization to others,

the likelihood of litigation against the outcome of the selection procedure, and perceived organizational attractiveness (Anderson et al., 2004; Chan & Schmitt, 2005; Gilliland, 1993; Ryan & Ployhart, 2000). However, the present dissertation has shown that applicant reactions are not only influenced by the selection instrument or medium itself, but also by factors outside the organization's control, such as applicants' computer anxiety, subjective well-being, or openness to experience (chapter 5). Thus, certain individuals may be more predisposed to react positively to selection instruments. Chapter 6 demonstrated that applicants' general beliefs in selection tests and their (perceived) test performance also affect their reactions. The nature of the applicant pool should therefore be carefully considered when designing interventions to improve applicant reactions. If negative applicant reactions are due to individual differences or general test beliefs instead of test content, modifying the test content or test administration medium will have little effect (Schmitt & Chan, 1999).

Third, the results of chapter 4 demonstrated that multimedia SJTs can be used as an implicit measure of personality traits. This finding may have important implications, as organizations have sought personality measures that are less affected by social desirability, faking, and self-presentation biases than explicit personality questionnaires (Fazio & Olson, 2003; Vaillant, 1998; Frost et al., 2007).

Fourth, the finding in chapter 2 that webcam test scores are not related to cognitive ability may also have important practical implications. Many organizations strive to create a diverse workforce. For this reason, organizations have been searching for selection instruments that are valid but at the same time minimize subgroup differences. Selection instruments with smaller cognitive loading produce smaller subgroup differences (Ployhart & Holtz, 2008).Therefore the use of a multimedia test as predictor measurement method can perhaps be an effective strategy to create a diverse workforce, as the present dissertation showed that a multimedia test is a valid selection instrument with a small cognitive loading.

From an applied point of view, multimedia tests could also have two limitations. First, some authors have suggested that multimedia tests could not be used to assess inexperienced workers, because some previous knowledge of the job is needed to address the situations adequately (Salgado & Lado, 2000). Yet, chapter 3 demonstrated that job tenure was not significantly related to scores on the webcam test.

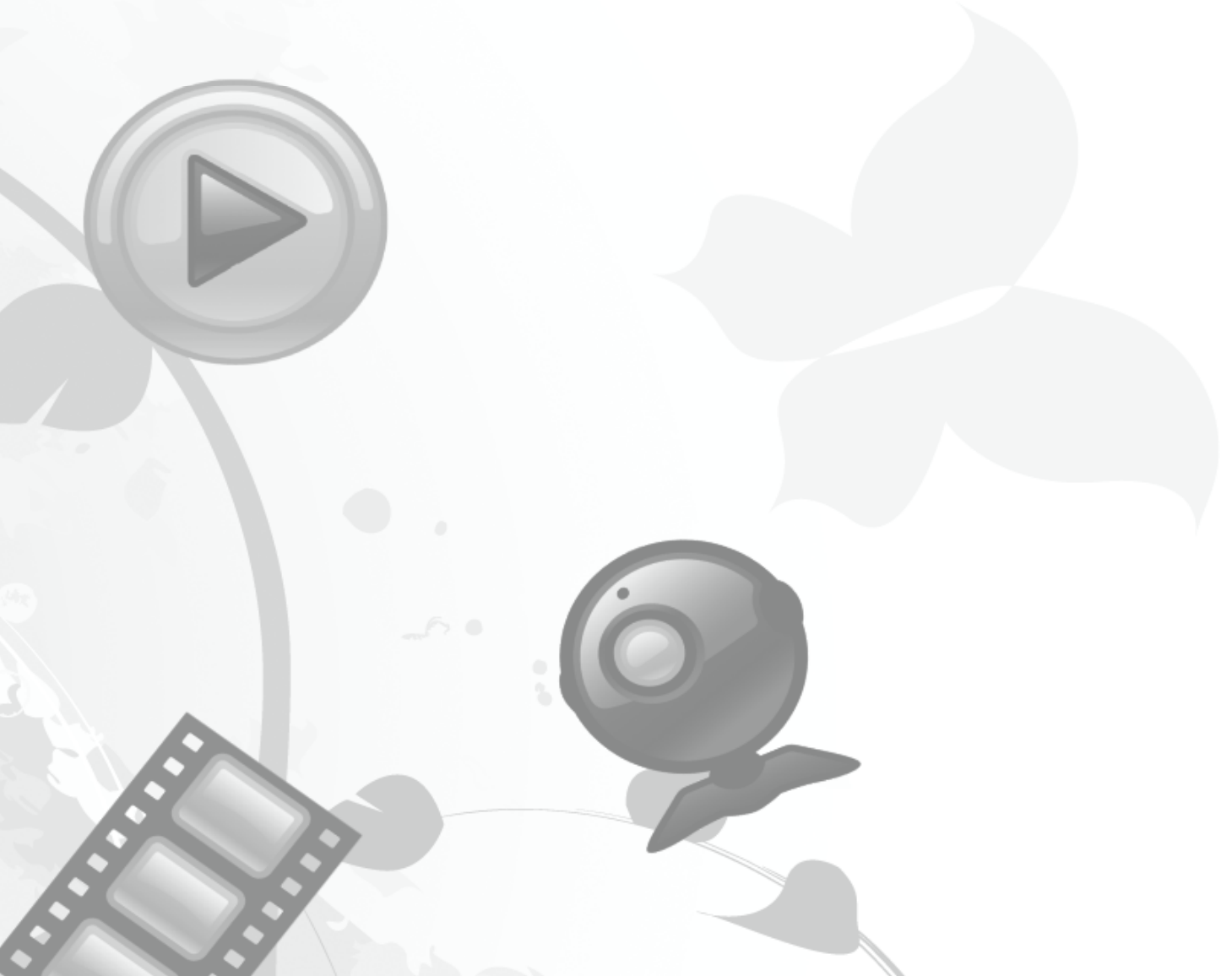
The second potential limitation of multimedia tests involves the costs of development, which is high compared to other selection instruments. Cost estimates per minute of filming vary from $2000 to $3000 (Dalessio, 1994). However, the cost effectiveness of any selection test is not only determined by the development costs, but also by its criterion-related validity (Cronbach & Gleser, 1965). As demonstrated in chapters 2, 3 and 4 multimedia tests show good criterion-related validities in the prediction of job and academic performance, implying that multimedia tests are worth their investment.

Conclusion

Although organizations have rushed to incorporate multimedia tests into their selection systems, research regarding these types of tests still was scarce. This dissertation aimed to address this shortcoming by presenting five empirical studies on the validity and acceptability of multimedia tests. To summarize, the present dissertation has demonstrated that multimedia tests can be useful and valuable predictors of academic and job performance beyond traditional measures as cognitive ability tests, personality questionnaires, and job knowledge tests. Also, as implicit measures of personality traits multimedia tests seem a valuable instrument for personnel selection practices. However, construct-driven multimedia tests were only able to predict conceptually aligned criterion measures. Therefore, it is important to clearly specify the criterion domain when incorporating multimedia tests into selection systems. Furthermore, it was found that multimedia tests are related to Big Five personality dimensions and job experience, but not to cognitive ability. As selection instruments with smaller cognitive loadings produce smaller subgroup differences, using multimedia tests may be an effective strategy to reduce adverse impact. It is important to verify and extend these findings in applicant settings.

Furthermore, the present dissertation demonstrated that applicants react more positively to multimedia tests than to more traditional tests, such as cognitive ability tests. However, not only the type of selection instrument or medium itself was found to affect applicant reactions, also individual differences, such as openness to experience, general belief in tests, and (perceived) test performance were found to affect applicant reactions. Moreover, pretest reactions and posttest reactions were affected by different factors. Pretest reactions were affected by applicants' general beliefs in tests, whereas posttest reactions were affected by applicants' test performance via self-assessed test performance. The nature of the applicant pool and the time of measurement of applicant reactions therefore should be carefully considered when designing interventions to improve applicant reactions. Further research on the effect of individual differences, beliefs in tests, and perceived test performance on applicant reactions is encouraged.

Samenvatting



Steeds meer organisaties maken gebruik van multimediatests bij de werving en selectie van hun personeel (Lievens et al., 2002). Met name het gebruik van multimedia situationele beoordelingstests (*situational judgment test* [SJT]) is de afgelopen decennia sterk toegenomen (McDaniel et al., 2007). Tijdens het maken van een multimedia SJT krijgen sollicitanten door middel van korte videofragmenten verschillende uitdagende situaties te zien die relevant zijn voor de functie die zij ambiëren. Op een belangrijk moment bevriezen de situaties en wordt er aan de sollicitanten gevraagd hoe zij zouden reageren in deze situaties. Dit doen zij door verschillende antwoordopties te evalueren (Weekley & Ployhart, 2006). Sinds kort wordt er in de selectiepraktijk ook een ander type multimediatest ingezet, namelijk een webcamtest. Een webcamtest kan gezien worden als een multimedia SJT met open vragen (Arthur & Villado, 2008). Tijdens het maken van een webcamtest krijgen sollicitanten net als in de multimedia SJT door middel van korte videofragmenten een aantal werkgerelateerde situaties te zien. Na het zien van de situaties wordt er echter aan de sollicitanten gevraagd om daadwerkelijk een reactie te geven, die gefilmd wordt met een webcam (Lievens et al., 2008).

Ondanks het feit dat organisaties in hun selectieprogramma's al veelvuldig gebruik maken van multimedia SJTs en webcamtests (Anderson, 2003), is er nog weinig wetenschappelijk onderzoek naar deze tests verricht. Dit proefschrift beschrijft vijf empirische studies met betrekking tot de validiteit en acceptatie van multimediatests en tracht daarmee dit hiaat op te vullen. In de volgende paragraaf zal een overzicht worden gegeven van de belangrijkste empirische bevindingen uit deze vijf studies.

Overzicht van Empirische Bevindingen

Alle selectie-instrumenten dienen beoordeeld te worden op basis van een aantal criteria, waaronder hun criteriumgerelateerde validiteit, hun incrementele validiteit, hun begripsvaliditeit en de acceptatie door sollicitanten (Cook, 2009). Met deze vier criteria als leidraad zal een overzicht worden gegeven van de empirische bevindingen uit de studies.

Criteriumgerelateerde validiteit

Deze vorm van validiteit heeft betrekking op de mate waarin een test gerelateerd is aan een bepaald extern criterium, zoals studiesucces of werksucces (Nunnally, 1978). In de hoofdstukken 2, 3 en 4 worden drie empirische studies beschreven waarin de criteriumgerelateerde validiteit van multimediatests is onderzocht. Over het algemeen ondersteunen de bevindingen de validiteit van multimediatests als voorspeller van zowel studiesucces als werksucces.

In **hoofdstuk 2** werd onderzocht of een webcamtest voor interpersoonlijk leiderschapsgedrag studiesucces kan voorspellen. Studiesucces bestond uit het gemiddelde studiecijfer en de geobserveerde leeractiviteiten van studenten, zoals hoe goed zij hun rol vervullen als voorzitter van werkgroepen. Aangezien eerdere onderzoeken reeds hadden aangetoond dat SJTs betere voorspellingen geven van interpersoonlijke criteria dan van cognitieve criteria (bijv. Lievens & Sackett, 2006; Oswald et al., 2004),

werd verwacht dat scores op de webcamtest voor interpersoonlijk leiderschapsgedrag een hogere validiteit zouden laten zien voor de geobserveerde leeractiviteiten dan voor het gemiddelde studiecijfer van de studenten. De data werden verzameld onder 153 psychologiestudenten. De resultaten ondersteunden de validiteit van de webcamtest als voorspeller van studiesucces. Scores op de webcamtest voorspelden de inzet van studenten tijdens werkgroepen, de kwaliteit van de voorbereiding voor deze werkgroepen, hoe goed studenten hun rol vervulden als voorzitter van werkgroepen, en hun leeractiviteiten in het algemeen. Zoals verwacht, lieten de scores op de webcamtest een hogere validiteit zien voor de geobserveerde leeractiviteiten dan voor het gemiddelde studiecijfer.

In **hoofdstuk 3** werd de criteriumgerelateerde validiteit onderzocht van een webcamtest die was ontwikkeld om de effectiviteit in de belangrijkste taak van werkadviseurs te voorspellen, namelijk het adviseren van werkzoekenden bij het vinden van een baan. Dit eerste veldonderzoek naar een webcamtest werd uitgevoerd bij een uitzendbureau. De steekproef bestond uit 188 werkadviseurs die deelnamen aan een certificeringsproces. Dit certificeringsproces bestond uit een assessment door middel van een webcamtest, een kennistest, een objectieve maat voor werksucces (het percentage cliënten van de adviseur dat een baan had gevonden), en een managersoordeel over hun werksucces (onder andere over de bijdrage die de werkadviseur had geleverd aan de opbrengsten van de afdeling). Er werd verwacht dat de scores op de webcamtest positief gerelateerd zouden zijn aan zowel de objectieve maat voor werksucces als het managersoordeel van werksucces. De resultaten boden deels steun voor de validiteit van de webcamtest. De scores op de webcamtest voorspelden de objectieve maat voor werksucces, maar niet het managersoordeel. Deze laatste bevinding zou verklaard kunnen worden door het feit dat de managers zich ervan bewust waren dat hun oordeel deel uitmaakte van het certificeringproces. Dit zou kunnen hebben geleid tot een bepaalde mildheid in hun oordelen, die de criteriumgerelateerde validiteit beïnvloed kan hebben.

In **hoofdstuk 4** werd onderzocht of zogenaamde *implicit trait policies* (ITPs; Motowidlo et al., 2006a), zoals gemeten met een multimedia SJT voor leiderschap, geobserveerd werkgedrag konden voorspellen. ITPs zijn impliciete overtuigingen met betrekking tot de effectiviteit van bepaalde persoonlijkheidstrekken (Motowidlo et al., 2006b). Een persoon kan bijvoorbeeld de overtuiging hebben dat de persoonlijkheidstrek consciëntieusheid over het algemeen erg effectief is. SJTs kunnen ITPs weergeven door de mate waarin bepaalde persoonlijkheidstrekken tot uiting komen in de antwoordopties te correleren met de waarderingen die de kandidaat heeft gegeven aan deze antwoordopties. Als een kandidaat bijvoorbeeld systematisch een positieve waardering geeft aan antwoordopties waarin een hoge mate van consciëntieusheid tot uiting komt, dan zal de kandidaat hoog scoren op ITPs voor consciëntieusheid. Motowidlo en Beier (2010) toonden aan dat ITPs, zoals gemeten met een SJT, een samengestelde maat van werksucces kunnen voorspellen. Net als in het onderzoek van Motowidlo en Beier, werd in hoofdstuk 4 getracht om de criteriumgerelateerde validiteit van ITPs in kaart te brengen. Echter, in tegenstelling tot het

onderzoek van Motowidlo en Beier en andere onderzoeken naar de criteriumgerelateerde validiteit van ITPs, werd gebruik gemaakt van een multimedia SJT die één bepaald begrip beoogt te meten (een zogenaamde constructgedreven SJT). Een constructgedreven SJT kent enkele voordelen ten opzichte van andere SJTs, namelijk dat de validiteit van de SJT te generaliseren is naar verschillende functies en dat de mogelijkheid wordt geboden om de predictor en het criterium conceptueel op elkaar af te stemmen. De multimedia SJT was ontwikkeld om leiderschapsvaardigheden te voorspellen. De antwoordopties varieerden in de mate waarin de persoonlijkheidstrekken extraversie en sociabiliteit tot uitdrukking werden gebracht. Verwacht werd dat ITPs voor extraversie en ITPs voor sociabiliteit sterker gerelateerd zouden zijn aan leiderschapsgedrag, een criterium dat conceptueel gezien was afgestemd op de SJT, dan aan andere typen werkgedragingen. De data werden verzameld onder 180 sollicitanten. De resultaten toonden aan dat ITPs voor extraversie leiderschapsgedrag voorspelden, zoals dat was geobserveerd door collega's en leidinggevenden. Zoals verwacht, waren de ITPs voor extraversie sterker gerelateerd aan leiderschapsgedrag dan aan andere typen werkgedragingen. Er werd echter geen relatie gevonden tussen ITPs voor sociabiliteit en leiderschapsgedrag. De complexe relatie tussen sociabiliteit en leiderschap zou dit resultaat kunnen verklaren.

Incrementele validiteit

Deze vorm van criteriumgerelateerde validiteit geeft aan in hoeverre een selectie-instrument ten opzichte van andere selectie-instrumenten iets toevoegt aan de voorspelling van een bepaald criterium (Hunsley & Meyer, 2003). In de hoofdstukken 2, 3 en 4 van dit proefschrift worden drie empirische studies beschreven waarin de incrementele validiteit van multimediatests is onderzocht. Over het algemeen ondersteunen de bevindingen de incrementele validiteit van multimediatests in de voorspelling van zowel studiesucces als werksucces.

In **hoofdstuk 2** werd onderzocht of een webcamtest voor interpersoonlijk leiderschapsgedrag incrementele validiteit heeft ten opzichte van een cognitieve capaciteitentest en een persoonlijkheidsvragenlijst in de voorspelling van geobserveerde leeractiviteiten van studenten. Eerder onderzoek had reeds aangetoond dat cognitieve capaciteiten en persoonlijkheid belangrijke voorspellers zijn van studiesucces (bijv. Lounsbury et al., 2003; Poropat, 2009). De waarde van de webcamtest neemt dan ook toe als deze incrementele validiteit zou hebben ten opzichte van deze traditionele voorspellers van studiesucces. De resultaten lieten zien dat de webcamtest in staat was om een uniek gedeelte van de variantie in studiesucces te verklaren. Hiermee is aangetoond dat in selectiecontexten een webcamtest een nuttige en valide aanvulling kan zijn ten opzichte van traditionele voorspellers.

Verschillende onderzoekers zijn van mening dat situationele tests, zoals webcamtests, hun criteriumgerelateerde validiteit te danken hebben aan het feit dat ze kennis meten (bijv. McDaniel & Nguyen, 2001). In **hoofdstuk 3** werd daarom onderzocht of een webcamtest in de voorspelling van werksucces incrementele validiteit heeft ten opzichte van een kennistest. Zoals hiervoor al is aangegeven, was deze webcamtest

ontwikkeld om de effectiviteit in de belangrijkste taak van werkadviseurs te voorspellen, namelijk het adviseren van werkzoekenden bij het vinden van een baan. De resultaten toonden aan dat de webcamtest incrementele validiteit had ten opzichte van de kennistest in de voorspelling van het percentage cliënten van de adviseurs dat een baan had gevonden. Dit resultaat suggereert dat de webcamtest meer meet dan alleen kennis. Regressieanalyses lieten ook zien dat de werkloosheidscijfers van de provincies waarin de adviseurs werkten gerelateerd was aan het percentage cliënten van de adviseurs dat een baan had gevonden. Hoe hoger de werkloosheidscijfers van de provincie waarin de adviseur werkte, hoe lager het percentage cliënten was waarvoor de adviseur een baan had gevonden. Gecontroleerd voor het effect van deze werkloosheidscijfers, maar ook voor leeftijd, geslacht, het aantal dienstjaren en scores op de kennistest, bleek de webcamtest nog steeds in staat om unieke variantie te verklaren in het criterium. Deze bevinding bevestigt dat de webcamtest een relevante aanvullende voorspeller van werksucces is.

In **hoofdstuk 4** werd onderzocht of ITPs, zoals gemeten met een multimedia SJT voor leiderschap, incrementele validiteit hebben ten opzichte van scores op persoonlijkheidsschalen en leiderschapservaring in de voorspelling van leiderschapgedrag. Volgens de ITP-theorie worden ITPs beïnvloed door persoonlijkheid en ervaring (Motowidlo et al., 2006a). Zowel voor persoonlijkheidstrekken als voor leiderschapservaring is reeds door anderen aangetoond dat deze gerelateerd zijn aan leiderschapsgedrag (bijv. Judge, Bono et al., 2002; Thomas & Cheese, 2005). Daarom werd onderzocht of de relatie tussen ITPs en geobserveerd leiderschapsgedrag toegeschreven kan worden aan de effecten van persoonlijkheidstrekken en leiderschapservaring op ITPs, of dat ITPs unieke variantie in leiderschapsgedrag kunnen verklaren naast de variantie die al verklaard wordt door persoonlijkheidstrekken en leiderschapservaring. De resultaten toonden aan dat ITPs voor extraversie unieke variantie konden verklaren in leiderschapsgedrag, zoals dit was geobserveerd door collega's en leidinggevenden. Deze variantie werd verklaard naast de variantie die door de persoonlijkheidstrek extraversie en door leiderschapservaring werd verklaard. Echter, ITPs voor sociabiliteit konden geen unieke variantie verklaren in leiderschapgedrag naast de variantie die al verklaard werd door de persoonlijkheidstrek sociabiliteit en leiderschapservaring.

Begripsvaliditeit

Deze vorm van validiteit geeft aan in hoeverre scores op een selectie-instrument gerelateerd zijn aan scores op andere instrumenten op basis van theoretisch opgestelde hypothesen (Carmines & Zeller, 1979). Het belang om de onderliggende processen en begrippen van situationele tests in kaart te brengen is door vele onderzoekers benadrukt (bijv. Lievens et al., 2008). In de hoofdstukken 2 en 4 worden daarom twee onderzoeken gepresenteerd naar de begripsvaliditeit van respectievelijk een webcamtest en een multimedia SJT.

In **hoofdstuk 2** werd de relatie onderzocht tussen aan de ene kant scores op een webcamtest voor interpersoonlijk leiderschapsgedrag en aan de andere kant de

persoonlijkheidstrekken extraversie en consciëntieusheid, cognitieve capaciteiten en werkervaring. Eerdere onderzoeken hadden reeds aangetoond dat scores op papieren versies van SJTs voor leiderschap gerelateerd zijn aan de persoonlijkheidstrekken extraversie en consciëntieusheid en aan cognitieve capaciteiten (Bergman et al., 2006; Oswald et al., 2004). Er werd echter verwacht dat scores op een webcamtest in mindere mate gerelateerd zouden zijn aan cognitieve capaciteiten dan scores op een papieren versie van een SJT. Deze verwachting was gebaseerd op het feit dat er in de webcamtest geen beroep gedaan wordt op de leesvaardigheden van de kandidaat. Bovendien worden de daadwerkelijke vaardigheden van kandidaten in werkgerelateerde situaties in kaart gebracht, doordat kandidaten in een webcamtest werkelijk een reactie dienen te geven (Motowidlo et al., 2008). Daarom werd verwacht dat scores op de webcamtest voor interpersoonlijk leiderschapsgedrag sterker gerelateerd zouden zijn aan de persoonlijkheidstrekken extraversie en consciëntieusheid dan aan cognitieve capaciteiten. De resultaten kwamen overeen met deze verwachting. De resultaten toonden daarnaast ook aan dat scores op de webcamtest gerelateerd zijn aan leiderschapservaring. Personen met relevante ervaring hebben een grotere kans om de situaties die gepresenteerd zijn in de webcamtest al een keer meegemaakt te hebben. Deze personen kunnen dus geleerd hebben hoe zij succesvol moeten handelen in dit soort situaties.

Volgens de ITP-theorie van Motowidlo en collega's (2006b), beïnvloedt de persoonlijkheid van kandidaten de waarderingen die zij geven aan de antwoordopties in een SJT. Om deze reden zouden ITPs op impliciete wijze de persoonlijkheid van kandidaten kunnen meten. In **hoofdstuk 4** werd onderzocht of een multimedia SJT voor leiderschap individuele verschillen in ITPs kunnen weergeven. Er werd specifiek verwacht dat ITPs voor extraversie en ITPs voor sociabiliteit, zoals gemeten met een multimedia SJT voor leiderschap, positief gerelateerd zouden zijn aan leiderschapservaring en respectievelijk de persoonlijkheidstrekken extraversie en sociabiliteit. De resultaten bevestigden deze verwachting. Een multimedia SJT voor leiderschap kan inderdaad gebruikt worden om ITPs weer te geven. ITPs waren positief gerelateerd aan de bijbehorende persoonlijkheidstrekken. Daarnaast toonden de resultaten aan dat werknemers met meer leiderschapservaring sterkere ITPs voor extraversie hadden. Leiderschapservaring was echter niet gerelateerd aan ITPs voor sociabiliteit. Ook hier zou de complexe relatie tussen sociabiliteit en leiderschap kunnen verklaren waarom werknemers met meer leiderschapservaring geen sterkere ITPs voor sociabiliteit hadden dan werknemers met minder leiderschapservaring.

Acceptatie

In de meeste onderzoeken naar reacties van kandidaten op multimediatests is getracht deze reacties te beschrijven of te vergelijken met reacties op andere instrumenten (bijv. Kanning et al., 2006; Richman-Hirsch et al., 2000). Er is echter nog weinig onderzoek verricht waarin verklaringen worden gezocht voor de manier waarop kandidaten reageren op multimediatests. In de hoofdstukken 5 en 6 worden twee studies gepresenteerd die dit hiaat trachtten op te vullen.

In **hoofdstuk 5** werd de relatie onderzocht tussen aan de ene kant een aantal testgerelateerde en algemene individuele verschillen en aan de andere kant de meest onderzochte reactie van kandidaten, namelijk de gepercipieerde relevantie van een test voor de toekomstige baan (Chan & Schmitt, 2004). Bij de gepercipieerde relevantie van een test zijn twee aspecten van die test van belang: enerzijds de *face validity* van de test, anderzijds de gepercipieerde voorspellende waarde van de test. Eerder onderzoek had reeds aangetoond dat de inhoud en de kenmerken van een test (bijv. het medium) invloed hebben op de gepercipieerde relevantie van de test (bijv. Chan & Schmitt, 1997). Toch bleef een groot deel van de variantie in deze meest onderzochte reactie van kandidaten nog onverklaard. Daarom werd onder 153 psychologiestudenten onderzocht of individuele verschillen een gedeelte van de variantie in de gepercipieerde relevantie van tests kunnen verklaren. De gepercipieerde relevantie werd gemeten ten aanzien van twee veelgebruikte gecomputeriseerde selectie-instrumenten, namelijk een cognitieve capaciteitentest en een multimedia SJT. Er werd specifiek gekeken naar de relatie tussen de gepercipieerde relevantie van de tests en angst (testangst en computerangst), zelfevaluaties (geloof in eigen kunnen [*self-efficacy*], *core self-evaluations* en subjectief welzijn) en drie persoonlijkheidstrekken (sociabiliteit, emotionele stabiliteit en openheid voor nieuwe ervaringen). De resultaten toonden aan dat computerangst, *core self-evaluations*, subjectief welzijn, sociabiliteit, emotionele stabiliteit en openheid voor nieuwe ervaringen de gepercipieerde relevantie van de cognitieve capaciteitentest en de multimedia SJT beïnvloedden, maar niet stelselmatig. Ter illustratie, de *face validity* van de cognitieve capaciteitentest was gerelateerd aan sociabiliteit, emotionele stabiliteit en openheid voor nieuwe ervaringen, terwijl de *face validity* van de multimedia SJT gerelateerd was aan computerangst, *core self-evaluations*, subjectief welzijn en openheid voor nieuwe ervaringen. Openheid voor nieuwe ervaringen was de meeste consistente voorspeller van de gepercipieerde relevantie van de tests voor de toekomstige baan. Dit impliceert dat mensen die meer open staan voor nieuwe ervaringen positiever reageren op gecomputeriseerde selectie-instrumenten dan mensen die nieuwe ervaringen schuwen.

In **hoofdstuk 6** werden de *face validity*, de gepercipieerde voorspellende waarde en de gepercipieerde rechtvaardigheid van een papieren versie en een gecomputeriseerde versie van een postbakoefening met elkaar vergeleken. Bij 205 sollicitanten werden deze reacties zowel voor de afname van de postbakoefening gemeten als na de afname van de postbakoefening. De resultaten toonden aan dat na de testafname de voorspellende waarde van de papieren postbakoefening hoger werd ingeschat dan de voorspellende waarde van de gecomputeriseerde postbakoefening. Deze resultaten kunnen echter verklaard worden door een verschil in de moeilijkheidsgraad van de papieren versie en de gecomputeriseerde versie van de postbakoefening. Sollicitanten die de gecomputeriseerde versie maakten scoorden namelijk significant lager op de postbakoefening. In dit onderzoek had de computerisering van de postbakoefening dus een negatief effect op testprestaties en daardoor waarschijnlijk een negatief effect op de gepercipieerde voorspellende waarde van de test. Ondanks de verschillen

in moeilijkheidsgraad werden er geen verschillen gevonden tussen de twee versies van de postbakoefening bij de andere reacties, te weten de *face validity* en de gepercipieerde rechtvaardigheid.

In hoofdstuk 6 werden ook enkele determinanten van reacties van kandidaten en hun relatie met testprestaties onderzocht door voort te borduren op het model van Chan, Schmitt, Sacco en collega's (1998). In de meeste onderzoeken zijn reacties van kandidaten slechts één keer gemeten, namelijk voordat de test werd afgenomen (bijv. Rynes, Connerly, 1993; Schmit & Ryan, 1997) of nadat de test werd afgenomen (bijv. Lievens & Sackett, 2006; Richman-Hirsch et al., 2000; Salgado & Moscoso, 2003). Echter, volgens Chan, Schmitt, Sacco en collega's worden vooraf gemeten reacties beïnvloed door andere factoren dan achteraf gemeten reacties. De resultaten toonden dit ook aan. De vooraf gemeten reacties werden vooral beïnvloed door het vertrouwen van sollicitanten in selectie-instrumenten in het algemeen. De achteraf gemeten reacties werden vooral beïnvloed door de testprestaties van de kandidaten.

Conclusies

Ondanks het feit dat organisaties in hun selectieprogramma's al veelvuldig gebruikmaken van multimediatests, was er nog weinig wetenschappelijk onderzoek naar deze tests verricht. De vijf empirische studies met betrekking tot de validiteit en acceptatie van multimedia tests die in dit proefschrift zijn beschreven, trachtten hier wat aan te doen. De studies in dit proefschrift toonden aan dat multimediatests zowel studiesucces als werksucces kunnen voorspellen. Ook als impliciete maat voor persoonlijkheid kunnen multimediatests gezien worden als een waardevol instrument voor de selectiepraktijk. Wanneer organisaties multimediatests willen implementeren in hun selectieprogramma's, is het echter belangrijk om het criteriumdomein nauwkeurig te specificeren De multimediatests bleken namelijk vooral conceptueel gerelateerde criteria te voorspellen. Daarnaast toonden de studies in dit proefschrift aan dat scores op multimediatests gerelateerd zijn aan persoonlijkheidstrekken en aan werkervaring, maar niet aan cognitieve capaciteiten.

De studies in dit proefschrift lieten tevens zien dat sollicitanten positiever reageren op multimediatests dan op meer traditionele tests, zoals cognitieve capaciteitentests. Echter, niet alleen het type selectie-instrument en het medium, maar ook factoren zoals openheid voor nieuwe ervaringen, het vertrouwen van sollicitanten in selectie-instrumenten in het algemeen en de testprestaties, bleken van invloed te zijn op de reacties van sollicitanten. Daarbij werden reacties die voor de testafname werden gemeten door andere variabelen beïnvloed dan reacties die na de testafname werden gemeten. Reacties voorafgaand aan de testafname werden beïnvloed door vertrouwen in selectie-instrumenten in het algemeen, terwijl reacties na de testafname werden beïnvloed door de testprestaties van de sollicitanten. Bij het ontwerpen van interventies die bedoeld zijn om de acceptatie van selectie-instrumenten te vergroten, moet daarom rekening worden gehouden met stabiele kenmerken van sollicitanten, zoals hun persoonlijkheid, en met het tijdstip waarop de reacties worden gemeten. Verder onderzoek is nodig om meer kennis te vergaren over de effecten van

individuele verschillen, het vertrouwen in selectie-instrumenten in het algemeen en de testprestaties op de reacties van sollicitanten ten aanzien van multimediatests.

References

Aguinas, H., Henle, C. A., & Beaty Jr, J. C. (2001). Virtual reality technology: A new tool for personnel selection. *International Journal of Selection and Assessment, 9*, 70-83.

Anderson, N. (2003). Applicant and recruiter reactions to new technology in selection: A critical review and agenda for future research. *International Journal of Selection and Assessment, 11*, 121-136.

Anderson, N., Lievens, F., Van Dam, K., & Ryan, A. M. (2004). Future perspectives on employee selection: Key directions for future research and practice. *Applied Psychology: An International Review, 53*, 487-501.

Arbuckle, J. L. (2007). Amos 16.0. Chicago, IL: SPSS.

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

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

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman and Company.

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

Bartram, D. (1994). Computer-based assessment. In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 9, pp. 31-69). New York, NY: John Wiley & Sons.

Bartram, D. (2005a). Computer-based testing and the internet. In A. Evers, N. Anderson & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 399-418). Malden, MA: Blackwell Pub.

Bartram, D. (2005b). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185-1203.

Bartram, D. (2006). Testing on the internet: Issues, challenges and opportunities in the field of occupational assessment. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the internet: Issues and advances* (pp. 13-37). New York, NY: John Wiley & Sons.

Bartram, D., & Bayliss, R. (1984). Automated testing: Past, present and future. *Journal of Occupational Psychology, 57*, 221-237.

Bauer, T. N., Maertz, C. P. J., Dolen, M. R., & Campion, M. A. (1998). Longitudinal assessment of application reactions to employment testing and outcome feedback. *Journal of Applied Psychology, 83*, 892-903.

Bauer, T. N., & Truxillo, D. M. (2006). Applicant reactions to situational judgment tests: Research and related practical issues. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 233-249). Mahwah, NJ: Erlbaum.

Bauer, T. N., Truxillo, D. M., Paronto, M. E., Weekley, J. A., & Campion, M. A. (2004). Applicant reactions to different selection technology: Face-to-face, interactive voice response, and computer-assisted telephone screening interviews. *International Journal of Selection and Assessment, 12*, 135-148.

Beckers, J. J., & Schmidt, H. G. (2003). Computer experience and computer anxiety. *Computers in Human Behavior, 19*, 785-797.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14*, 223-235.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*, 205-212.

Bernerth, J. B., Feild, H. S., Giles, W. F., & Cole, M. S. (2006). Perceived fairness in employee selection: The role of applicant personality. *Journal of Business and Psychology, 20*, 545-563.

Bloom, A. J., & Hautaluoma, J. E. (1990). Anxiety management training as a strategy for enhancing computer user performance *Computers in Human Behavior, 6*, 337-349.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology, 76*, 863-872.

Bornstein, R. F. (2002). A process dissociation approach to objective-projective test score interrelationships. *Journal of Personality Assessment, 78*, 47-68.

Brown, J. M., & Weiss, D. J. (1977). An adaptive testing strategy for achievement in test batteries (Research Report 77-6). Psychometrics program, Department of Psychology, University of Minnesota.

Brutus, S. (1995). The perception of selection tests: An expanded model of perceived job-relatedness. Unpublished doctoral dissertation. Bowling Green State University, Bowling Green, OH.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 687-732). Palo Alto, CA: Consulting Psychologists Press.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park, CA: Sage

Cascio, W. F. (1987). *Applied psychology in personnel management* (3th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*, 270-295.

Chan, D. (1997). Racial subgroup difference in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology, 82*, 311-320.

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.

Chan, D., & Schmitt, N. (2005). Situational judgment tests. In A. Evers, O. Smit-Voskuil & N. Anderson (Eds.), *Handbook of personnel selection* (pp. 219-246). Oxford, UK: Blackwell Publishers, Inc.

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

Chan, D., Schmitt, N., Jennings, D., Clause, C. S., & Delbridge, K. (1998). Applicant perceptions of test fairness: Integrating justice and self-serving bias perspectives. *International Journal of Selection and Assessment, 6*, 232-239.

Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology, 83*, 471-485.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410-417.

Cook, M. (2009). *Personnel selection: Adding value through people* (5th ed.). New York, NY: John Wiley & Sons.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana, IL: University of Illinois Press.

Dalessio, A. T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology, 9*, 23-32.

De Houwer, J. (2006). What are implicit measures and why are we using them? In R. W. H. J. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11-28). Thousand Oaks, CA: Sage.

De Meijer, L. A. L., Born, M. P., Van Zielst, J., & Van der Molen, H. T. (in press). The construct-driven development of a video-based situational judgment test for integrity: A study in a multi-ethnic police setting. *European Psychologist*.

Depreeuw, E. A. M. (1984). A profile of the test-anxious student. *International Review of Applied Psychology, 33*, 221-232.

Diener, E. (1994). Assessing subjective well-being: Progress and opportunities. *Social Indicators Research, 31*, 103-157.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71-75.

Drasgow, F., & Mattern, K. (2006). New tests and new items: Opportunities and issues. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the internet: Issues and advances* (pp. 59-75). New York, NY: John Wiley & Sons.

Drenth, P. J. D. (1965). *Test voor niet-verbale abstractie*. Amsterdam: Swets & Zeitlinger.

Elkins, T. J., & Phillips, J. S. (2000). Job context, selection decision outcome, and the perceived fairness of selection tests: Biodata as an illustrative case. *Journal of Applied Psychology, 85*, 479-484.

Epitropaki, O., & Martin, R. (2004). Implicit leadership theories in applied settings: Factor structure, generalizability, and stability over time. *Journal of Applied Psychology, 89*, 293-310.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297-327.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition*. New York, NY: McGraw-Hill.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York, NY: John Wiley & Sons Inc.

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.

Gilliland, S. W. (1992). *The perceived fairness of selection systems: An organizational justice perspective*: Unpublished doctoral dissertation, Michigan State University, East Lansing.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *The Academy of Management Review, 18*, 694-734.

Gilliland, S. W. (1994). Effects of procedural and distributive justice on reactions to a selection system. *Journal of Applied Psychology, 79*, 691-701.

GITP. (2010). *Manual G5R personality questionnaire*. Unpublished manuscript.

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., Braverman, E. P., & Chung, B. (1992). *Method versus content: The effects of different testing methodologies on subgroup differences*. Paper presented at the 7th Annual Conference of the Society for Industrial and Organizational Psychology, Montreal, Quebec, Canada.

Graziano, W. G., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. In R. Hogan, J. A. Johnson & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 795-825). New York: Academic Press.

Greenberg, J. (1990). Organizational justice: Yesterday, today, and tomorrow. *Journal of Management, 16*, 399-432.

Hall, S., Fetzer, M. S., Tuzinski, K., & Freeman, M. (2010). *3D computer animation: I-O finally catches up with IT*. Paper presented at the the 25th Annual Conference of the Society for Industrial and Organizational Psychology (SIOP), Atlanta, GA.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Application reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639-683.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Heinssen, R. K., Glass, C. R., & Knight, L. A. (1987). Assessing computer anxiety: Development and validation of the Computer Anxiety Rating Scale. *Computers in Human Behavior, 3*, 49-59.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test-anxiety. *Review of Educational Research, 58*, 47-77.

Highhouse, S., Stanton, J. M., & Reeve, C. L. (2004). Examining reactions to employer information using a simulated web-based job fair. *Journal of Career Assessment, 12*, 85-96.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*, 100-112.

Hogan, R., Curphy, G. J., & Hogan, J. (1994). What we know about leadership: Effectiveness and personality. *American Psychologist, 49*, 493-504.

Horvath, M., Ryan, A. M., & Stierwalt, S. L. (2000). The influence of explanations for selection test use, outcome favorability, and self-efficacy on test-taker perceptions. *Organizational Behavior and Human Decision Processes, 83*, 310-330.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling. Concepts, issues, and applications* (pp. 76-99). London: Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446-455.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Ironson, H. G., Guion, R. M., & Ostrander, M. (1982). Adverse impact from a psychometric perspective. *Journal of Applied Psychology, 67*, 419-432.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software International.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*, 765-778.

Judge, T. A., Erez, A., Bono, J. E., & Thoreson, C. J. (2003). The core self-evaluations scale: Development of a measure. *Personnel Psychology, 56*, 303-331.

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87*, 530-541.

Judge, T. A., Locke, E. A., & Durham, C. C. (1997). The dispositional causes of job satisfaction: A core evaluations approach. *Research in Organizational Behavior, 19*, 151-188.

Judge, T. A., Locke, E. A., Durham, C. C., & Kluger, A. N. (1998). Dispositional effects on job and life satisfaction: The role of core evaluations. *Journal of Applied Psychology, 83*, 17-34.

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.

Keller, T. (1999). Images of the familiar: Individual differences and implicit leadership theories. *The Leadership Quarterly, 10*, 589-607.

Kelloway, E. K. (1996). Common practices in structural equation modeling. In C. L. Copper & I. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 141-180). Chichester, UK: John Wiley.

Kluger, A. N., & Rothstein, H. R. (1993). The influence of selection test type on applicant reactions to employment testing. *Journal of Business and Psychology, 8*, 3-25.

Koch, B. P. N. (1998). *Handleiding G5*. Unpublished manuscript.

Koestner, R., Zuckerman, M., & Koestner, J. (1987). Praise, involvement, and intrinsic motivation. *Journal of Personality and Social Psychology, 53*, 383-390.

Kreitzberg, C. B., Stocking, M. L., & Swanson, L. (1978). Computerized adaptive testing: Principles and directions. *Computers & Education, 2*, 319-329.

Lambert, A. J., & Wedell, D. H. (1991). The self and social judgment: Effects of affective reaction and "own position" on judgments of unambiguous and ambiguous information about others. *Journal of Personality and Social Psychology, 61*, 884-897.

Lievens, F. (2006). International situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 279-300). Mahwah, NJ: Lawrence Erlbaum.

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

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment, 10*, 245-257.

Lievens, F., De Corte, W., & Brysse, K. (2003). Applicant perceptions of selection procedures: The role of selection information, belief in tests, and comparative anxiety. *International Journal of Selection and Assessment, 11*, 67-77.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37*, 426-441.

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., & Thornton, G. C., III. (2005). Assessment centers: Recent developments in practice and research. In A. Evers, O. Smit-Voskuijl & N. Anderson (Eds.), *Handbook of selection* (pp. 243-264). London: Blackwell.

Lievens, F., Van Dam, K., & Anderson, N. (2002). Recent trends and challenges in personnel selection. *Personnel Review, 31*, 580-601.

Lord, R. G., De Vader, C. L., & Alliger, G. M. (1986). A meta-analysis of the relation between personality and leadership perceptions: An application of validity generalization procedures. *Journal of Applied Psychology, 71*, 402-410.

Lord, R. G., Foti, R. J., & De Vader, C. L. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior and Human Performance, 34*, 343-378.

Lounsbury, J. W., Sundstrom, E., Loveland, J. M., & Gibson, L. W. (2003). Intelligence, "Big Five" personality traits, and work drive as predictors of course grades. *Personality and Individual Differences, 35*, 1231-1239.

Loyens, S. M. M., Rikers, R. M. J. P., & Schmidt, H. G. (2007). The impact of students' conceptions of constructivist assumptions on academic achievement and drop-out. *Studies in Higher Education, 32*, 581-602.

Lyubomirsky, S., King, L., & Diener, E. (2005). The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin, 131*, 803-855.

Macan, T. H., Avedon, M. J., Pease, M., & Smith, D. E. (1994). The effects of applicants' reaction to cognitive ability tests and an assessment center. *Personnel Psychology, 47*, 715-738.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structural modeling. *Psychological methods, 1*, 130-149.

Marcati, A., Gianluigi, G., & Peluso, A. M. (2008). The role of SME entrepreneurs' innovativeness and personality in the adoption of innovations. *Research Policy, 37*, 1579-1590.

Markus, H., Smith, J., & Moreland, R. L. (1985). Role of the self-concept in the perception of others. *Journal of Personality and Social Psychology, 49*, 1494-1512.

Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320-341.

Maurer, T. J. (2001). Career-relevant learning and development, worker age, and beliefs about self-efficacy for development. *Journal of Management, 27*, 123-140.

McBride, J. R. (1997). The marine corps exploratory development project: 1977-1982. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 59-67). Washington, DC: American Psychological Association.

McCrae, R. R., & Costa, P. T. (1989). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of Personality and Social Psychology, 56*, 586-595.

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

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730-740.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*, 103-113.

McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33*, 515-525.

McDaniel, M. A., Whetzel, D. L., Hartman, N. S., Nguyen, N. T., & Grubb, W. L. (2006). Situational judgment tests: Validity and an integrative model. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 183-203). Mahwah, NJ: Lawrence Erlbaum.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599-616.

McHenry, J. J., & Schmitt, N. (1994). Multimedia testing. In M. G. Rumsey & C. B. Walker (Eds.), *Personnel selection and classification* (pp. 193-232). Hillsdale, NJ: Lawrence Erlbaum.

Mead, A. D. (2001). *How well does web-based testing work? Results of a survey of users of NetAssess*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology (SIOP), San Diego, CA.

Meltzer, P. H. (1995). Videotest voor communicatieve vaardigheden. In F. J. R. C. Dochy & T. R. de Rijke (Eds.), *Assessment centers: Nieuwe toepassingen in opleiding, onderwijs en HRM* (pp. 109-122). Utrecht: Lemma.

Miller, D., Smith-Jentsch, K. A., & Afek, A. (2008). *Situational judgment tests as measures of implicit trait policies.* Paper presented at the 24th Annual Conference of the Society for Industrial and Organizational Pyschology (SIOP), New Orleans, LA.

Miller, D. T. (1978). What constitutes a self-serving attributional bias? A reply to Bradley. *Journal of Personality and Social Psychology, 36*, 1221-1223.

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., Brownlee, A. L., & Schmit, M. J. (2008). Effects of personality characteristics on kwowledge, skill, and performance in servicing retail customers. *International Journal of Selection and Assessment, 16*, 272-281.

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

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). A theoretical basis for situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 57-81). Mahwah, NJ: Lawrence Erlbaum.

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

Murphy, K. R., Thornton, G. C., & Prue, K. (1991). Influence of job characteristics on the acceptability of employee drug testing. *Journal of Applied Psychology, 76*, 447-453.

Nguyen, N. T., McDaniel, M. A., & Whetzel, D. L. (2005). *Subgroup differences in situational judgment test performance: A meta-analysis*. Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw Hill.

Olson-Buchanan, J. B., & Drasgow, F. (1999). Beyond bells and whistles: An introduction to computerized assessment. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 1-5). Mahwah, NJ: Lawrence Erlbaum.

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.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills *Personnel Psychology, 51*, 1-24.

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

Organ, D. W. (1994). Personality and organizational citizenship behavior. *Journal of Management, 20*, 465-478.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89*, 187-207.

Peterson, N. G., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. Linn (Ed.), *Educational Measurement* (Vol. 3, pp. 221-262). New York: MacMillian.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*, 33-40.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11*, 1-16.

Ployhart, R. E., & Harold, C. M. (2004). The applicant attribution-reaction theory (AART): An integrative theory of applicant attributional processing. *International Journal of Selection and Assessment, 12*, 84-98.

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.

Ployhart, R. E., & Ryan, A. M. (1997). Toward an explanation of applicant reactions: An examination of organizational justice and attribution frameworks. *Organizational Behavior and Human Decision Processes, 72*, 308-335.

Ployhart, R. E., & Ryan, A. M. (1998). The relative importance of procedural and distributive justice in determining applicants' reactions. *Journal of Applied Psychology, 83*, 3-16.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*, 322-338.

Potosky, D., & Bobko, P. (2004). Selection testing via the internet: Practical considerations and exploratory empirical findings. *Personnel Psychology, 57*, 1003-1034.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18*, 103-134.

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.

Prentice, D. A. (1990). Familiarity and differences in self- and other-representations. *Journal of Personality and Social Psychology, 59*, 369-383.

Reynolds, D. H., Sinar, E. F., & McClough, A. C. (2000). *Evaluation of an Internet-based selection procedure.* Paper presented at the 15th Annual Conference of the Society of Industrial and Organizational Psychology (SIOP), New Orleans, LA.

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.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and applicant reactions. *Journal of Occupational Psychology, 55*, 171-183.

Ryan, A. M., Greguras, G. J., & Ployhart, R. E. (1996). Perceived job relatedness of physical ability testing for firefighters: exploring variations in reactions. *Human Performance, 9*, 219-240.

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

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management, 26*, 565-606.

Rynes, S. L., & Connerley, M. L. (1993). Applicant reactions to alternative selection procedures. *Journal of Business and Psychology, 7*, 261-277.

Salgado, J. F., & Lado, M. (2000). *Validity generalization of video tests for predicting job performance ratings*. Paper presented at the 15th Annual Conference of the Society of Industrial and Oganizational Psychology.

Salgado, J. F., & Moscoso, S. (2003). Internet-based personality testing: Equivalence of measures and assesses' perceptions and reactions. *International Journal of Selection and Assessment, 11*, 194-205.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology* (pp. 165-199). London: Sage.

Sanchez, R. J., Truxillo, D. M., & Bauer, T. N. (2000). Development and examination of an expectancy-based measure of test-taking motivation. *Journal of Applied Psychology, 85*, 739-750.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Schmidt, F. L. (1994). The future of personnel selection in the U.S. Army. In M. G. Rumsey, C. B. Walker & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 333-350). Hillsdale, NJ: Erlbaum.

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., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of Job Experience and Ability on Job Knowledge, Work Sample Performance, and Supervisory Ratings of Job Performance. *Journal of Applied Psychology, 71*, 432-439.

Schmidt, H. G., & Moust, J. H. C. (2000). Factors affecting small-group tutorial learning: A review of research. In D. H. Evensen & C. E. Hmelo (Eds.), *Problem-based learning: A research perspective on learning interactions* (pp. 19-52). Mahwah, NJ: Lawrence Erlbaum.

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

Schmitt, N., & Chan, D. (1999). The status of research on applicant reactions to selection tests and its implications for managers. *International Journal of Management Reviews, 1*, 45-62.

Schmitt, N., & Chan, D. (2006). *Situational judgment tests: Method or construct?* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956-970.

Shotland, A., Alliger, G. M., & Sales, T. (1998). Face validity in the context of personnel selection: A multimedia approach. *International Journal of Selection and Assessment, 6*, 124-130.

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

Skarlicki, D. P., Folger, R., & Tesluk, P. (1999). Personality as a moderator in the relationship between fairness and retaliation. *Academy of Management Journal, 42*.

Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46*, 49-76.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

Stemler, S. E., & Sternberg, R. J. (2006). Using situational judgment tests to measure practical intelligence. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 107-131). Mahwah, NJ: Lawrence Erlbaum.

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use* (2nd ed.). Oxford, England: Oxford University Press.

Stricker, L. J. (1982). Interpersonal competence instrument: Development and preliminary findings. *Applied Psychological Measurement, 6*, 69-81.

Thibodeaux, H. F., & Kudisch, J. D. (2003). The relationship between applicant reactions, the likelihood of complaints, and organization attractiveness. *Journal of Business and Psychology, 18*, 247-257.

Thomas, R. J., & Cheese, P. (2005). Leadership: Experience is the best teacher. *Strategy & Leadership, 33*, 24-29.

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology, 87*, 1020-1031.

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2006). A field study of the role of big five personality in applicant perceptions of selection fairness, self, and the hiring organization. *International Journal of Selection and Assessment, 14*, 269-277.

Truxillo, D. M., Seitz, R., & Bauer, T. N. (2008). The role of cognitive ability in self-efficacy and self-assessed test performance. *Journal of Applied Social Psychology, 38*, 903-918.

Van Leeuwen, R. G. J. (2004). Handleiding Captain. *Unpublished manuscript*.

Van Vianen, A. E. M., Taris, R., Scholten, E., & Schinkel, S. (2004). Perceived fairness in personnel selection: Determinants and outcomes in different stages of the assessment procedure. *International Journal of Selection and Assessment, 12*, 149-159.

Viswesvaran, C., & Ones, D. S. (2004). Importance of perceived personnel selection system fairness determinants: Relations with demographic, personality, and job characteristics. *International Journal of Selection and Assessment, 12*, 172-186.

Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology, 52*, 1236-1247.

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49*, 436-458.

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

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*, 679-700.

Weekley, J. A., & Ployhart, R. E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance, 18*, 81-104.

Weekley, J. A., & Ployhart, R. E. (2006). *Situational judgment tests: Theory, measurement, and application*. Mahwah, NJ: Lawrence Erlbaum.

Weiss, D. J. (1973). The stratified adaptive computerized ability test (Research Report 73-3). Psychometrics program, Department of Psychology, University of Minnesota.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372-376.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*, 188-202.

Wiechmann, D., & Ryan, A. M. (2003). Reactions to computerized testing in selection contexts. *International Journal of Selection and Assessment, 11*, 215-229.

Wiggins, J. S. (1995). *Interpersonal adjective scales: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Yeo, G., & Neal, A. (2008). Subjective cognitive effort: A model of states, traits, and time. *Journal of Applied Psychology, 93*, 617-631.

Yukl, G. (1998). *Leadership in Organizations*. Upper Saddle River, NJ: Prentice Hall.

Dankwoord

Ik ben ontzettend dankbaar voor alle hulp die ik van een groot aantal personen heb gekregen tijdens het maken van dit proefschrift. Zonder hun steun, advies, aanmoedigingen, vertrouwen, tijd en geduld was het nooit gelukt om dit proefschrift in drie jaar af te ronden. Allereerst wil ik natuurlijk mijn promotoren bedanken, Marise Born en Henk van der Molen. Ondanks jullie overvolle agenda's, hadden jullie altijd tijd om met mij te overleggen en om feedback te geven op mijn werk. Marise, bedankt voor de fijne begeleiding en voor alle kansen die je me hebt geboden de afgelopen jaren. Door jouw steun en vertrouwen, heb ik de kans gekregen om aan dit AiO-avontuur te beginnen. Je hebt me de afgelopen jaren enorm gemotiveerd en gestimuleerd om kritisch na te denken over mijn werk. Tegelijkertijd gaf je me altijd alle ruimte om mijn eigen ideeën uit te werken. Ik had geen betere promotor kunnen wensen! Henk, ik heb onze halfjaarlijkse overleggen zeer op prijs gesteld. Bedankt voor je goede adviezen en je motiverende woorden.

Ook wil ik mijn copromotor, Alec Serlie, bedanken voor de mogelijkheid om dit project bij GITP uit te voeren. Alec, ik vond mijn tijd bij GITP heel waardevol, niet alleen vanwege alle onderzoeksmogelijkheden, maar ook vanwege alle personen die ik heb leren kennen. Zoals ik al vaker heb gezegd, ik had het niet willen missen!

Ik zou ook graag alle commissieleden willen bedanken: Arnold Bakker, Filip Lievens, Henk van der Flier, Hans Hoekstra, Eva Derous en Edwin van Hooft. Bedankt dat jullie deel wilden uitmaken van mijn promotiecommissie en ondanks jullie drukke agenda's de tijd hebben kunnen vinden om mijn proefschrift te lezen en te beoordelen.

De afgelopen jaren heb ik veel steun gehad van mijn kamergenootjes: Kiki, Suzanne en Lisette. Bedankt voor jullie luisterend oor en voor alle gezelligheid! Bedankt dat jullie altijd voor me klaar stonden en dat we onze ervaringen en frustraties hebben kunnen delen. Ik zal jullie erg missen als kamergenootjes... Kiki en Suzanne, ik ben erg blij dat jullie mijn paranimfen wilden zijn!

Ook wil ik graag alle (oud) collega's bij het instituut voor psychologie aan de Erasmus Universiteit bedanken: Gera, Benjamin, Maria, Arnold, Heleen, Eva, Edwin, Lonneke, Daantje, Despoina, Marjan, Lieke, Kimberley, Jeroen, Nevra, René, Dimitri, Matthijs, Mirella, Jennifer en Evie. Bedankt voor alle hulp bij mijn project, voor de inspirerende praatjes en discussies, en natuurlijk ook voor de prettige werksfeer! Gera, Benjamin en Maria, jullie wil ik in het bijzonder bedanken. Ik heb veel aan onze meetings gehad! Ik vond het fijn dat ik altijd bij jullie binnen kon lopen om ideeën te bespreken of gewoon voor een gezellig praatje.

Hierbij wil ook graag alle (oud) collega's bij GITP bedanken: Annemarie, Joost, Anneke, Wai-Wah, Rob, Paul, Miriam, Emiel, Hans, Nicole, Joris, Chris en alle stagiairs. Ik heb met veel plezier de afgelopen jaren met jullie samengewerkt. Bedankt voor de fijne werksfeer, jullie frisse ideeën en jullie hulp bij onder meer het ontwikkelen van instrumenten, het genereren van rapporten en het samenstellen van databestanden. Joost, bedankt voor de fijne samenwerking de afgelopen jaren! We hebben van veel (stage)projecten een succes kunnen maken! Ook wil ik je hartelijk bedanken voor de tijd en energie die je hebt gestoken in het ontwerpen van de omslag van dit proef-

schrift! Anneke, bedankt voor je hulp bij alle praktische problemen die ik tegenkwam tijdens mijn onderzoek. Maar ik wil je ook vooral bedanken voor de gezelligheid die je altijd bracht op de afdeling!

Hierbij wil ik ook GITP bedanken voor de financiële steun en alle adviseurs, testzaalcoördinatoren en medewerkers van het secretariaat voor hun hulp bij de dataverzameling.

Bij de dataverzameling heb ik ook hulp gekregen van enkele studenten, die ik heb mogen begeleiden bij hun mastertheses: Marit, Charlotte, Lineke, Nicole en Roxanna. Ik wil jullie bedanken voor al jullie harde werk! Ik vond het een eer om jullie te mogen begeleiden en wil jullie heel veel succes wensen in jullie verdere carrière.

Ook wil ik graag de student-assistenten bedanken, die mij de afgelopen jaren hebben geholpen. Nienke, bedankt voor al je hulp bij het zoeken naar kandidaten die wilden meewerken aan ons onderzoek. Dit bleek geen gemakkelijke klus te zijn… Maar door jouw harde werk en de vele, vele mailtjes die je hebt gestuurd en beantwoord is het toch gelukt! Ook wil ik graag Marit, Arn, Sonja, Michiel en Jorrit bedanken voor het nakijken van de opnames uit de webcamtest.

Paul van der Maesen en Barend Koch, bedankt voor jullie steun en advies de afgelopen jaren! Jullie hebben mijn enthousiasme voor multimediatests gewekt! Erg fijn dat ik altijd gebruik kon maken van de SQ-tests en webcamtests. Ik hoop in de toekomst nog veel met jullie te mogen samenwerken.

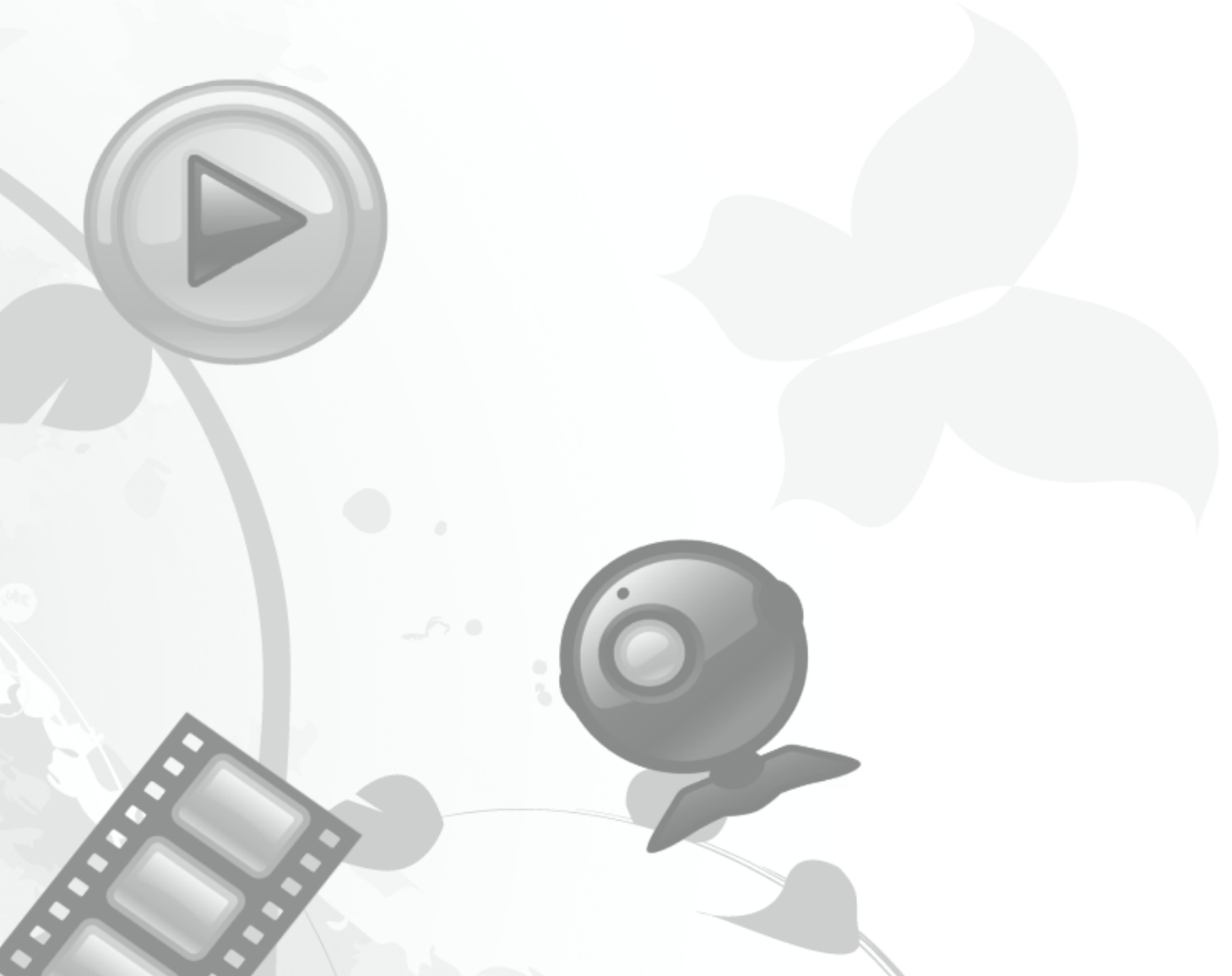
Tot slot, wil ik mijn familie en vrienden bedanken voor hun liefde, steun en interesse voor mijn (soms onbegrijpelijke) werk. Bij jullie vond ik de afgelopen jaren de ontspanning en afleiding die ik nodig had. Pa en ma, bedankt voor alles wat jullie me hebben gegeven! Lieve Barry, jij bent de belangrijkste persoon voor mij. Zonder jou was het nooit gelukt! Bedankt voor alle mooie momenten die we samen hebben meegemaakt en nog zullen meemaken.

Iedereen ontzettend bedankt!

Curriculum Vitae

Janneke Oostrom was born on March 10th, 1984, in Krimpen aan den IJssel, The Netherlands. She completed her secondary education in 2001 at the Comenius College in Capelle aan den IJssel. Hereafter, she started studying Psychology at the Erasmus University Rotterdam. She received her Master's degree in Industrial and Organizational Psychology in 2005 (cum laude). Her published master thesis concerned an evaluation of an aggression management training program to cope with workplace violence in the healthcare sector. This master thesis was awarded with the Unilever Research Prize. From September 2005 till August 2007, she worked as a tutor in Industrial and Organizational Psychology at the Erasmus University Rotterdam and as a consultant for Van der Maesen & Koch HRM-advies. In September 2007, Janneke started as a PhD student at the Institute of Psychology at the Erasmus University Rotterdam, studying the topic of new technology in personnel selection. The results of the PhD project, which was co-financed by GITP International BV, are reported in the present dissertation. As a PhD student she taught a number of psychology and practical courses and supervised several bachelor and master theses. At the time of this writing Janneke is employed as an assistant professor at the Institute of Psychology at the Erasmus University Rotterdam.

Kurt Lewin Institute dissertation series



The "Kurt Lewin Institute Dissertation Series" started in 1997. Since 2008 the following dissertations have been published:

2008-1: Marijke van Putten: *Dealing with missed opportunities. The causes and boundary conditions of inaction inertia*
2008-2: Marjolein Maas: *Experiential Social Justice Judgment Processes*
2008-3: Lonneke de Meijer: *Ethnicity effects in police officer selection: Applicant, assessor, and selection-method factors*
2008-4: Frederike Zwenk: *Voice by Representation*
2008-5: Margreet Reitsma: *The Impact of Linguistically Biased Messages on Involved Receivers*
2008-6: Marcus Maringer: *Feeling one thing, seeing another: Emotion comparison effects in person judgments*
2008-7: Hanneke Heinsman: *The competency concept revealed: Its nature, relevance, and practice*
2008-8: Joris Lammers: *Toward a more social social psychology of power*
2008-9: Daniël Fockenberg: *Between Good and Evil: Affective Priming in Dynamic Context*
2008-10: Arne van den Bos: *Why we stereotype influences how we stereotype: self-enhancement and comprehension effects on social perception*
2008-11: Lidewij Niezink: *Considering Others in Need: On Altruism, Empathy and Perspective Taking*
2008-12: Aad Oosterhof: *Better together: Antecedents and consequences of perceived expertise dissimilarity and perceived expertise complementarity in teams*
2008-13: Femke ten Velden: *Negotiation in dyads and groups: The effects of social and epistemic motives*
2008-14: Maike Wehrens: *How did YOU do? Social comparison in secondary education*
2008-15: Kyra Luijters: *Making Diversity Bloom: Coping Effectively with Cultural Differences at Work*
2008-16: Ilona de Hooge: *Moral emotions in decision making: Towards a better understanding of shame and guilt*
2008-17: Lindred L. Greer: *Team Composition and Conflict: The Role of Individual Differences*
2008-18: Sezgin Cihangir: *The Dark Side of Subtle Discrimination: How targets respond to different forms of discrimination*
2008-19: Giel Dik: *On the contagiousness of others' goals: The role of perceiving effort*
2008-20: Lotte van Dillen: *Dealing with negative feelings: The role of working memory in emotion regulation*
2008-21: Marijn Poortvliet: *Information exchange examined: An interpersonal account of achievement goals*
2008-22: Sjoerd Pennekamp: *Dynamics of disadvantage: Uncovering the role of group-based anger*

2008-23: Chris Reinders Folmer: *Cooperation and communication: Plastic goals and social roles*
2009-1: Marijke Leliveld: *Ethics in Economic Decision-Making*
2009-2: Monique Pollmann: *Accuracy and Bias in Person Perception*
2009-3: Krispijn Faddegon: *Regulatory Focus in Group Contexts*
2009-4: Lieven Brebels: *Mirror, mirror on the wall… Procedural fairness as an evaluative and regulatory looking-glass self*
2009-5: Daphne Wiersema: *Taking it personally: Self-esteem and the protection of self-related attitudes*
2009-6: Judith D.M. Grob: *Dial E for Emotion: Context and Consequences of Emotion Regulation*
2009-7: Katherine Stroebe: *Is this about me? Responding to subtle discrimination - beyond an individual versus group perspective*
2009-8: Menno Vos: *Identity patterns in diverse work groups: Improving social integration outcomes through relational identities*
2009-9: Lennart Renkema: *Facing Death Together: Understanding The Consequences of Mortality Threats*
2009-10: Michael Vliek: *Group-based social comparison processes: An intragroup level of analysis*
2009-11: Karlijn Massar: *Unconscious rivals: The automatic evaluation of rivals in jealousy-evoking situations*
2009-12: Bart Terwel: *Origins and consequences of public trust: Towards an understanding of public acceptance of carbon dioxide capture and storage*
2009-13: Emma ter Mors: *Dealing with information about complex issues: The role of source perceptions*
2009-14: Martijn Veltkamp: *On the Instigation of Implicit Motivation: How Deprivation and Positive Affect Cause Motivated Behavior*
2009-15: Marret K. Noordewier: *Consistency and the unexpected*
2009-16: Sytske van der Velde: *Imitation of Emotion: How meaning affects the link between imitation and liking*
2009-17: Jacomijn Hofstra: *Attaching Cultures: The role of attachment styles in explaining majority members' acculturation attitudes*
2009-18: Jacqueline Tanghe: *Affect in Groups: Convergence, Conditions and Consequences*
2009-19: Anne Marike Lokhorst: *Using Commitment to Improve Environmental Quality*
2009-20: Jonathan van 't Riet: *Framing Health Communication Messages*
2009-21: Suzanne Pietersma: *Persuasive Health Communication: A Self-Perspective*
2009-22: Remco Wijn: *A functional perspective on the justice judgment process and its consequences*
2009-23: Niels van de Ven: *The bright side of a deadly sin: The psychology of envy*
2009-24: Anthon Klapwijk: *The Power of Interpersonal Generosity*
2010-1: Maarten Wubben: *Social Functions of Emotions in Social Dilemmas*

2010-2: Joyce Rupert: *Diversity faultlines and team learning*
2010-3: Daniel Lakens: *Abstract Concepts in Grounded Cognition*
2010-4: Luuk Albers: *Double You? Function and Form of Implicit and Explicit Self-Esteem*
2010-5: Matthijs Baas: *The Psychology of Creativity: Moods, Minds, and Motives*
2010-6: Elanor Kamans: *When the Weak Hit back: Studies on the Role of Power in Intergroup Conflict*
2010-7: Skyler Hawk: *Changing Channels: Flexibility in Empathic Emotion Processes*
2010-8: Nailah Ayub: *National Diversity and Conflict: The Role of Social Attitudes and Beliefs*
2010-9: Job van der Schalk: *Echoing Emotions: Reactions to Emotional Displays in Intergroup Context*
2010-10: Nevra Cem: *Organizational citizenship behavior and counterproductive work behavior: Cross-cultural comparisons between Turkey and the Netherlands*
2010-11: Ron Broeders: *On Situated and Embodied Knowledge Regarding Moral Issues*
2010-12: Margriet Braun: *Dealing with a deviant group member*
2010-13: Dennis Bleeker: *Representing or defecting? The pursuit of individual upward mobility in low status groups*
2010-14: Petra Hopman: *Group Members Reflecting on Intergroup Relations*
2010-15: Janneke Oostrom: *New Technology in Personnel Selection: The Validity and Acceptability of Multimedia Tests*