RUNNING HEAD: WEBCAM TESTING

Webcam Testing: Validation of an Innovative Open-Ended Multimedia Test

Janneke K. Oostrom*, Marise Ph. Born*, Alec W. Serlie**, and Henk T. van der Molen*

* Erasmus University Rotterdam

** Erasmus University Rotterdam / GITP

Contact address:

Janneke K. Oostrom

Erasmus University Rotterdam

Faculty of Social Sciences, T13-08

PO-Box 1738

3000 DR Rotterdam, The Netherlands

oostrom@fsw.eur.nl

+ 31 10 4082933 (phone)

+ 31 10 4089009 (fax)

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 up and above job knowledge in predicting job placement success. The webcam test, therefore, seems a promising type of instrument for personnel selection.

KEYWORDS: WEBCAM TESTING, MULTIMEDIA TESTING, VIDEO TESTING, PERSONNEL SELECTION

Acknowledgements

We wish to thank Paul E.A.M. van der Maesen de Sombreff en Barend P.N. Koch for providing the webcam test materials and for their valuable comments on an earlier version of this article.

Webcam Testing: Validation of an Innovative Open-Ended Multimedia Test

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, van Dam, & Anderson, 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 an open response 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, Dunnette, & Carter, 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, Drasgow, Donovan, Henning, & Juraska, 2006; Lievens & Sackett, 2006; McDaniel, Hartman, Whetzel, & Grubb, 2007). McDaniel et al. 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, Pereira, Wiechmann, Schmitt, & Harvey, 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 up and above traditional predictors (Chan & Schmitt, 2005), suggesting SJTs capture a unique part of job performance. For example, Clevenger et al. 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, 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 (Schmidt, Hunter, & Outerbridge, 1986).

Multimedia Testing

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 (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, Olson-Buchanan, and Drasgow (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 (1997) 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 up 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 video test with orally-given free responses was significantly higher than the criterion-related validities of the multimedia test 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, 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). Interrater 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. 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 up 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 up 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 = .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 = .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 = .94) did not differ

significantly from the other consultants (M = 4.7, SD = .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 HRM 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. Cronbach’s 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". Cronbach's 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 ABOUT HERE ---

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 Cronbach's 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 ($z = -.57$, *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 ABOUT HERE ---

We tested our third hypothesis, which stated that the webcam test would incrementally predict job performance up 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 ($\beta = .17$, $F = 3.82$, $p < .05$). When the webcam test was added in the next step, it explained an additional 4 % of the variance in job placement success ($\beta = .20$, $F = 3.68$, $p < .05$). We also conducted a

hierarchical regression analysis to examine whether the job knowledge test had incremental validity 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 ($\beta = .20$, $F = 4.85$, $p < .05$). 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 ($\beta = .17$, $F = 2.67$, *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 the 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 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; Stricker), 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., 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 up 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 .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, Hooper, and Jackson (2006) 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 up 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 (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 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 up 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.

References

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*(2-3), 121-136.

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*(3), 223-235.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*(2), 205-212.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of Supervisory Job Performance Ratings. *Journal of Applied Psychology, 76*(6), 863-872.

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*(1), 143-159.

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.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*(3), 410-417.

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

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*(2), 115-123.

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.

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*(4), 245-257.

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*(5), 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*. London: Blackwell.

Lievens, F., van Dam, K., & Anderson, N. (2002). Recent trends and challenges in personnel selection. *Personnel Review, 31*(5), 580-601.

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

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*(1), 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*(4), 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*(1&2), 103-113.

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: Erlbaum.

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.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*(6), 640-647.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91*(4), 749-761.

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*(4), 823-854.

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.

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*. 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), 1-24.

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*(2), 187-207.

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), 1-16.

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*(6), 880-887.

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.

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*(2), 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*(3), 432-439.

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*(1), 45-62.

Schmitt, N., & Ostroff, C. (1986). Operationalizing the "behavioral consistency" approach: Selection test development based on a content-oriented strategy. *Personnel Psychology, 39*(1), 91-108.

Stricker, L. J. (1982). Interpersonal competence instrument: Development and preliminary findings. *Applied Psychological Measurement, 6*(1), 69-81.

Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology, 52*(6), 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*(2), 436-458.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*(1), 25-49.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*(3), 679-700.

Weekley, J. A., & Ployhart, R. E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance, 18*(1), 81-104.

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

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	.62	-.24*	(-)							
3. Gender	1.57	.50	-.36**	.17*	(-)						
4. Unemployment rate	5.60	0.78	-.09	-.05	-.05	(-)					
5. Job tenure	4.65	.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	.51	-.08	.10	.09	-.12	-.09	.13	.13	.25*	(.82)

Note. Cronbach's alpha of the scales 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. $N = 188$

* $p < .05$

** $p < .01$

Table 2.

Hierarchical Regression Analysis.

	Job placement success ($N = 90$)				Manager's appraisal ($N = 188$)			
Predictors	β	R^2	ΔR^2	F	β	R^2	ΔR^2	F
Step 1		.12	.12	2.73*		.03	.03	1.20
Age	.24				-.01			
Gender	.22				.06			
Job tenure	-.10				-.06			
Unemployment rate	-.10				-.11			
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.

* $p < .05$

** $p < .01$