Situational Tests in Student Selection: An Examination of Predictive Validity, Adverse Impact, and Construct Validity

Filip Lievens and Pol Coetsier*

Ghent University

The Flemish Admission Exam 'Medical and Dental Studies' is comprised of four cognitive ability tests and four situational tests, namely two work samples (i.e., a lecture and a medical text) and two video-based situational judgement tests (i.e., a physician–patient interaction and a medical expert discussion). On the basis of the Admission Exam scores of 941 candidates (359 men, 582 women) this study shows that situational tests significantly can predict better than cognitive ability tests, with lecture and text emerging as significant predictors. When situational tests are combined with cognitive ability tests, there are no mean gender differences. Situational tests also enable us to measure a broader range of constructs. For example, in this study, the personality factor Openness is related to better situational test performance. Overall, this study demonstrates that situational tests may be a useful complement to traditional student selection procedures.

Introduction

When schools, institutions and universities face a large number of applicants for places available, it is understandable that they adopt some kind of selection procedure. Traditionally, the selection procedure for admission to medical and dental studies was based on prior academic achievement (e.g., Green, Peters, and Webster 1993; McManus 1982; Montague and Odds 1990), knowledge of science-related subjects (e.g., Montague and Odds 1990; Tomlinson, Clack, Pettingale, Anderson, and Ryan 1977), and cognitive abilities (e.g., Roessler, Lester, Butler, Rankin, and Collins 1978; Vu, Dawson-Saunders, and Barrows 1987). In general, these cognitively oriented variables turned out to be good predictors of the academic performance of medical students, especially in the so-called pre-clinical years (e.g., Green, Peters, and Webster 1991; Minnaert 1996; Mitchell, Haynes, and Koenig 1994; Powis 1994). These results are not unique to medical studies as it is well documented that cognitive ability also plays a dominant role in academic achievement in general (Neisser *et al.* 1996; Sternberg and Kaufman 1998). Besides cognitively oriented variables, personality factors have also been used to predict medical student performance (e.g., Aldrich 1987; Ferguson, Sanders, O'Hehir and James, 2000; Gough and Hall 1982; Hobfoll, Anson and Antonovsky 1982; Hojat *et al.* 1993). Although the evidence as to which personality factors are important for success in medical student performance is mixed, inclusion of personality factors has been found to significantly add to the prediction of medical academic success (Powis 1994; Shen and Comrey 1997).

What many of the aforementioned predictors and predictor instruments have in common that they are 'sign-based'. This means that they are in the first place geared at measuring some dispositions or constructs (e.g., verbal intelligence, persistence, etc.) (Wernimont and Campbell 1968). Given this long-standing tradition of sign-based predictors in medical student selection, it was not surprising that measures of cognitive ability and knowledge of the sciences were also included in the Flemish Admission Exam 'Medical and Dental Studies', which was first set up and organized in 1997 (Ministerie van de Vlaamse Gemeenschap 1996). More specifically, one part of the Admission Exam was designed to evaluate applicants' mastery of four science-related

* Address for correspondence: Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium. E-mail: filip.lievens@rug.ac.be

© Blackwell Publishers Ltd 2002, 108 Cowley Road, Oxford OX4 1JF, UK and
350 Main Street, Malden, MA 02148, USA.

subjects: chemistry, biology, physics, and mathematics.[1] Another part was comprised of four cognitive ability tests: reasoning, memory association, visual information processing, and pattern recognition. A final part of the Flemish Admission Exam 'Medical and Dental Studies' consisted of situational tests.

Situational tests are based on the sample approach to personnel selection (Wernimont and Campbell 1968). For instance, in the Admission Exam, medical student candidates were presented with examples of situations, which they were likely to encounter in the future as students *and* physicians. To this end, two miniaturized work samples (i.e., a videotaped lecture of a professor and a silent reading protocol with a medical subject matter) and two video-based situational judgement tests (i.e., physician–patient interaction and medical expert discussion) were developed. These situational tests did not aim to measure specific constructs (see also Motowidlo *et al.* 1990). Due to practical considerations including the standardized administration of tests to a large group of applicants and the need for a fast scoring system, responses to the situational tests were captured via multiple-choice questions.

To date, the use of such situational tests in medical student selection has remained unexplored (Powis 1994; Roberts and Porter 1990). Therefore, this study will examine the effectiveness of situational tests (i.e., miniaturized work samples and video-based situational judgement tests) in a student selection context in terms of three perspectives: predictive validity, adverse impact, and construct validity.

Situational Tests: Short Overview and Research Needed

In the personnel selection literature situational tests have emerged as an important and useful complement to the more traditional sign-based predictor instruments. Anastasi and Urbina define a situational test as 'one that places the test taker in a situation closely resembling or simulating a "real-life" criterion situation' (1997, p. 450). Given this definition, examples of situational tests include accomplishment records (e.g., Hough 1984), situational judgement tests (e.g., Motowidlo, Dunnette and Carter 1990), situational interviews (e.g., Latham, Saari, Pursell and Champion 1980), video-based situational judgement tests (e.g., Weekley and Jones 1997), work samples (e.g., Robertson and Kandola 1982), and assessment centre exercises such as role-plays (e.g., Thornton 1992). Situational tests have also become known in the educational literature under the aliases of performance assessment, alternative assessment or authentic assessment (Baker, O'Neil and Linn 1993; Linn, Baker and Dunbar 1991; Messick 1994; Sackett 1998; Wiggins 1989).

In general, the personnel selection literature has shown that situational tests have good predictive validities. This is shown by looking at meta-analyses of situational judgement tests (McDaniel, Morgeson, Finnegan, Campion and Braverman, in press), situational interviews (e.g., McDaniel, Whetzel, Schmidt and Maurer 1994), video-based situational judgement tests (Salgado and Lado, 2000), work samples (Schmidt and Hunter 1998), and assessment centres (Gaugler, Rosenthal, Thornton and Bentson 1987). The notion of behavioural consistency (Schmitt and Ostroff 1986), which posits that the behaviour of candidates in situations similar to those encountered on the job will provide good predictions of actual job behaviour, has been suggested as the most straightforward explanation for the positive predictive validity results of situational tests. Although the predictive validity of situational tests seems to be well established in personnel selection, it is less known whether the effectiveness of situational tests can also be extended to student selection in general and to medical student selection in particular.

A second research issue regarding situational tests pertains to adverse impact in terms of gender. For instance, prior research on situational judgement tests found that women typically scored higher than men did. This was evidenced by the different samples in Motowidlo *et al.* (1990) and Motowidlo and Tippins (1993) showing that women outperformed men, with effect sizes varying from .11 to .32. In a similar vein, Weekley and Jones (1997; 1999) found that women scored higher than men did by .31 and .19 standard deviations. Some assessment centre studies (Neubauer 1990; Schmitt 1993; Shore 1992) also reported a subtle gender bias favouring female candidates. An unresolved question is whether these subgroup differences on situational tests in favour of women average out subgroup differences favouring men on measures of cognitive ability (e.g., spatial orientation and visualization, Jensen 1998) (see Pulakos and Schmitt 1996, for a similar argument for reducing adverse impact in terms of race).

Finally, it is still unclear which constructs are associated with performance on video-based situational judgement tests. Because these tests evaluate a variety of knowledge, skills, and abilities relevant to the target job, they are typically multidimensional in nature (Chan and Schmitt 1997). Prior studies mainly focused on cognitive-based correlates of situational judgement test performance (e.g., cognitive ability measures, GPA, etc.). In the recent meta-analysis of McDaniel *et al.* (in press) the correlation between written situational judgement tests and cognitive-based correlates was estimated to be .53, with lower correlations usually associated with video-based versions (see also Weekley and Jones 1997). Another possibility, which has remained virtually unexplored, is that personality factors are also related to performance in situational judgement tests because many of these situations are interpersonally oriented. Conforming to these assumptions, Nguyen and McDaniel (2001) reported moderate correlations between situational judgement test performance and three factors of the Five-Factor Model of personality, namely agreeableness, conscientiousness, and emotional stability.

Present Study: Aims

The overall aim of this study is to examine the effectiveness of situational tests in student selection. This general objective can be broken down in three specific objectives. First, we examine whether the two miniaturized work samples and the two video-based situational judgement tests included in the Flemish Admission Exam 'Medical and Dental Studies' provide valid predictions of students' medical school performance in the first year. On the basis of the notion of behaviour consistency, we especially expect that two situational tests, namely the lecture and the medical text, will emerge as significant predictors of medical students' first year grades (our criterion measure, see below). We also expect that the situational tests will show incremental validity over the cognitive ability measures.

Second, we examine adverse impact via inspection of mean score differences between men and women at the test level as well as at the level of the composite Admission Exam Score (i.e., a combination of the scores obtained on both cognitive ability measures and situational tests). As suggested above, we expect that, although adverse impact in terms of gender will exist at the test level, this will not be the case for the composite score. Besides looking at mean differences between men and women, we also examine whether there exists evidence of differential prediction (Bartlett, Bobko, Mosier and Hannon 1978; Cleary 1968; Dunbar and Novick 1988).

Third, we investigate which constructs are associated with performance on the situational tests included in the admission exam. To this end, we place the situational tests in a nomological net with cognitive ability measures. We also place the situational tests in a nomological net with the Five-Factor Model of personality. On the basis of prior research (McDaniel *et al.* in press) we expect the situational tests to have significant and substantial correlations with the cognitive ability measures. However, because two situational tests, namely, the physician–patient interaction and medical expert discussion, aim to capture candidates' reactions to interpersonal situations, we also expect that personality factors will play a role in the performance on these two tests in particular.

Method

Sample

The total sample consisted of 941 candidates (359 men and 582 women), who attended the Admission Exam 'Medical and Dental Studies' in Flanders. The average age of the candidates was eighteen years and three months.

In the predictive validity study only participants who had passed the Admission Exam, had entered the first year of medical and dental studies in one of the five Flemish universities, and had obtained a final score at the end of the first year, were included. In total, we were able to obtain the first year scores of 610 students (227 men and 383 women; mean age = eighteen years and two months).

In the construct validity study only participants who had passed the Admission Exam, had entered the first year of medical and dental studies, and had attended the course in which the authorized Flemish translation (Hoekstra, Ormel and De Fruyt 1996) of the NEO-PI-R (Costa and McCrae 1992) was administered, were included. Specifically, 529 students (185 men and 344 women; mean age = eighteen years and two months) completely filled in the NEO-PI-R. There were no significant differences between this group and the group entering medical and dental studies on the study variables.

Predictors: The Admission Exam Tests and Scores

Cognitive ability tests. These measures were not specifically developed for the Admission Exam. Instead, four existing cognitive ability measures were chosen. For test security reasons we cannot mention the source of these cognitive ability tests. For the same reason we cannot present sample items. Interested researchers may contact the authors to obtain more information.

The first cognitive ability measure was a 'reasoning' test, which consisted of 54 questions with five response alternatives. The problems in this test were formulated in either verbal, numeric, or figure terms. Prior research demonstrated the good reliability and predictive validity of this reasoning test for medical students (Minnaert 1996). In particular, Minnaert (1996) reported an internal consistency of .84 and a validity coefficient of .36 for predicting the final scores obtained in the first year of medical and dental studies. Hence, the Admission Exam commission decided to weigh this test more in the total Admission Exam score (see Table 1 for specific weights).

The second test 'visual information processing' (32 items) measured the ability to quickly scan and interpret complex figures. According to prior research provided in the test booklet the internal consistency of this test was .77.

Table 1: Means, standard deviations, and intercorrelations among study variables

	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Gender	–	–	–																
Cognitive ability tests																			
2. Reasoning (54[a], .50[b])	26.27	5.94	−.05	–[c]															
3. Visual information (32[a], .20[b])	11.76	5.22	−.13**	.28**	.83														
4. Memory association (20[a], .10[b])	9.10	3.86	.14**	.24**	.03	.70													
5. Pattern recognition (52[a], .20[b])	23.86	10.80	−.05	.31**	.29**	.09**	.92												
Situational tests																			
6. Lecture (40[a], .33[b])	30.26	4.31	−.01	.31**	.14**	.15**	.17**	.55											
7. Medical text (20[a], .17[b])	15.18	2.37	.05	.33**	.10**	.16**	.13**	.37**	.56										
8. Physician-patient interaction (30[a], .25[b])	23.45	3.01	.16**	.25**	.02	.19**	.07*	.29**	.31**	.41									
9. Medical expert discussion (30[a], .25[b])	24.17	2.92	.08*	.22**	.06	.21**	.05	.29**	.22**	.32**	.48								
10. Cognitive ability test score (10)[d]	4.54	1.00	−.07*	.81**	.61**	.37**	.69**	.31**	.30**	.21**	.20**	–							
11. Situational test score (10)[d]	7.76	.73	.09**	.40**	.12**	.25**	.16**	.79**	.63**	.67**	.64**	.37**	–						
12. Admission exam score (20)	12.30	1.45	.00	.77**	.48**	.39**	.56**	.62**	.53**	.49**	.46**	.88**	.76**	–					
NEO-PI-R																			
13. Extraversion	166.21	19.82	.09*	−.09*	−.01	.00	.03	−.10*	−.05	−.03	.02	−.04	−.07	−.07	.89				
14. Agreeableness	170.24	19.50	.22**	−.11**	−.06	−.01	−.09*	−.07	−.08	−.02	−.05	−.13**	−.08	−.14**	.14**	.90			
15. Conscientiousness	166.26	20.15	.03	.00		−.01	−.05	−.05	−.03	−.06	−.02	−.09*	−.04	−.08	−.07	.07	.24**	.91	
16. Emotional stability	138.16	22.10	.16**	−.05	−.14**	.03	.01	−.04	−.02	.08	.03	−.07	.01	−.05	−.24**	−.03	−.31**	.91	
17. Openness	169.35	18.66	.22**	.07	−.02	.08	.05	.02	.13**	.16**	.11*	.07	.14**	.13**	.37**	.25**	.03	.00	.87
Criterion																			
18. Final first year score (20)	11.14	3.48	.02	.30**	−.03	.13**	.11**	.19**	.19**	.10**	.02	.24**	.20**	.29**	−.10*	.01	.20*	.02	.04
Final first year score, corrected for direct restriction of range				.33	−.03	.12	.12	.20	.21	.12	.02	.27	.23	.35					

Note: *Sample sizes varied between 529 and 941. Internal consistency coefficients (alphas) are on the diagonal. Statistical significance was determined prior to correcting correlations for restriction of range.*

[a] *This value refers to the total number of items (or the maximal score) of each test.*

[b] *This value indicates the weight of each test.*

[c] *Because of the stringent test time limit, there were very few students, who completed all of the items. Therefore, alpha could not be computed. Students received a score on the Reasoning test by summing correct answers (applying the penalty for guessing) This score was used in the subsequent analyses.*

[d] *These scores were obtained by summing the test score, which was first multiplied by its weight.*

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

In the third test 'memory association', 15 names of patients (in five groups of three) had to be memorized. Besides the patient names, their age, their job title, their personal characteristic, and their diagnosis were also included. The reproduction phase contained 20 questions dealing with these patient descriptions. According to prior studies with a similar memory association test, which were provided in the test booklet, the internal consistency of this test equalled .70.

The fourth test 'pattern recognition' measured the cognitive ability to determine which simple figure was part of a complex figure. Fifty complex figures were included and per complex figure five possible simple figures were presented. According to prior research provided in the test booklet, the internal consistency of this test was .80.

All four cognitive ability measures were multiple-choice tests with five response alternatives. For each test specific time limits were set.

Situational tests. As already noted, the situational tests were specifically developed for the Admission Exam. The first two tests, namely the videotaped lecture and the written text with a medical subject matter ('medical text'), were miniaturized samples of important student tasks. For reasons of realism, we decided to use a real lesson and a real course text as stimulus materials. To this end, a professor delivering a lecture (lasting about 30 minutes) was filmed. In reality the professor used to give this lecture in the second year of medical studies. In a similar vein, the seven-page text was extracted from a larger course syllabus. Two professors in medicine assisted us in developing a list of relevant questions and response options. The questions covered only the lecture (text) content and were evenly distributed over the whole lecture (text). Correct answers were also determined by scrutiny of the lecture (text) content. Pilot testing of these questions was not possible because of test security reasons. The Admission Exam Commission also did not allow us to improve the psychometric properties of the tests on the basis of received applicant data (e.g., discard 'bad' questions).

The other situational tests (i.e., 'physician–patient interaction' and 'medical expert discussion') were video-based situational judgement tests. The design of these tests was similar to procedures used in previous studies (Lievens, 2000; Motowidlo *et al.* 1990; Weekley and Jones 1997). In a first step a representative group of critical incidents were gathered for these two situations. To this end, we inspected the relevant literature (e.g., Tate 1994) and asked five experienced physicians (mean age = 41 years; mean working experience = 15.2 years) and five professors in general medicine (mean age = 38 years; mean working experience = 9.8 years) to provide examples indicative of effective and ineffective job behaviour in the respective situations. The literature review and the interviews with these subject matter experts yielded a list of 376 examples of behaviour (after eliminating redundancies).

Second, scripts were written. Care was taken to preserve realism and smoothness by nesting critical behaviours among innocuous material. The scripts depicted the word-for-word dialogue between the parties involved. Two professors teaching physicians' consulting practices tested the scripts for realism. On the basis of their suggestions one script had to be rewritten.

Next, semi-professional actors were selected to play the various roles. These actors were videotaped delivering their scripted performances. An experienced physician attended the set to guarantee realism. The actors were filmed in a recording studio equipped with props to simulate the two situations. The videotape was filmed using a two-camera shot. After professional editing, the physician–patient interaction ran for about six minutes and the medical expert discussion eleven minutes.

Fourth, questions (i.e., situational items) and responses (i.e., item options) were derived from the videotaped performances and critical behaviours. For each videotaped performance, 30 multiple-choice questions were formulated. Again, pilot testing and calibration of these questions were not possible. A sample question of the physician–patient interaction was the following:

'If you were the physician on the videotape, which of the following would be a better sentence to open the conversation with the patient?'

1. What symptoms do you have?
2. Tell me why you came in.
3. Are you here again, what's wrong?
4. Do the problems of last time still bother you?

In the last step expert judgements were used to develop scoring rules. In particular, eleven experienced physicians (mean age = 43 years; mean working experience = 16 years) received the written scripts, the videotaped performances, the questions, and the response alternatives. Their task was to read the scripts, observe the videotaped performances, and independently indicate the best response to each question. The experts observed the videotapes under optimal conditions. This meant, for instance, that they could view the videotaped performances repeatedly and rewind them. We analysed the expert answers to see if they agreed on the best response option. The results were satisfactory. Cohen's (1960) kappa, which is a coefficient of chance-corrected inter-rater agreement for nominal scales, equalled .75 for the physician–patient interaction, and .73 for the medical expert discussion. In a subsequent meeting between the experts the discrepancies were discussed and resolved. To this end, about 20 per cent of the original questions and response alternatives had to be changed. Lievens and

Coetsier (1998) present a more thorough description of the development of the situational tests.

All questions of the situational tests were of the multiple-choice type, with four response alternatives. Again, specific time limits were set for each test.

Admission Exam scores. For each of the eight tests of the Admission Exam a final score was computed by summing the number of correct answers. There was a small penalty for guessing, namely each incorrect answer received a penalty of 0.1 point. Next, a weighted sum of the four cognitive ability measures and a weighted sum of the four situational test scores were computed. These weights, which are presented in Table 1, were determined by the Admission Exam commission. The maximum score on each of these two weighted sum scores (i.e., Cognitive Ability Test Score and Situational Test Score) was 10. The Admission Exam commission decided that candidates had to obtain at least 6 on 10 on each weighted sum score to pass the Admission Exam. Finally, the Admission Exam Score was obtained by summing these two weighted sum scores.

Candidates who passed the exam received a certificate. This certificate guaranteed entry to the university in which they wanted to start their medical studies. Hence, there was no further selection on the part of the universities.

Personality Inventory

We used an authorized Flemish translation (Hoekstra, Ormel and De Fruyt 1996) of the long version of the NEO-PI-R (Costa and McCrae 1992). The NEO-PI-R is a measure of the Five Factor Model of personality. Each of the Five Factors is further divided into six facets, each measured by 8 items, resulting in 240 items. The response scale ranged from 1 (strongly agree) to 5 (strongly disagree). A factor analysis (principal axes with varimax-rotation) performed on our data resulted in five factors (eigenvalues from 2.20 to 3.6), which explained 53 per cent of the variance. Twenty-eight of the 30 facets had a significant loading of .40 or higher on the factor, which they purported to measure. In addition, all scales were found to be internally consistent, with Cronbach's alpha varying from .87 (Openness) to .91 (Emotional Stability and Conscientiousness). These results are in line with prior large-scale Dutch studies (De Fruyt 1996; Hoekstra *et al.* 1996), supporting the underlying structure of the NEO-PI-R in terms of the Five Factor Model. This enabled us to compute a score per subject on each of the five factors and the 30 facets. This score was the mean self-rating on the scales, which belonged to a factor (facet).

Criterion Measure

In the predictive validity study the final scores of the students at the end of the first year medical and dental studies served as the criterion. The final score of a student at the end of the first year was the average of the scores obtained by a student on the various courses. Students could obtain a maximum score of 20.

To obtain information on the reliability of this criterion measure, we computed the internal consistency of the final score with the scores on the courses as items. Across the various universities Cronbach's alpha varied from .87 to .91. In terms of construct validity our criterion measure is probably heavily influenced by general mental ability because science courses are primarily taught in the first year. In addition, we should note that the actual content of this first year differed across universities (with respect to the courses and professors). However, closer inspection of the courses taught across the universities showed only slight variations.

Procedure of the Admission Exam

The predictors were gathered during the Flemish Admission Exam 'Medical and Dental Studies' (1997). On the first day of this admission exam, candidates completed the four cognitive ability tests: reasoning, visual information processing, memory association, and pattern recognition (in this order). On the second day, candidates completed the four situational tests. First, the videotaped lecture was shown. Candidates were expected to take notes on copies of the transparencies used in the lecture. They could use their notes to answer the follow-up questions. Next, they received the medical text and completed the questions dealing with this text. Finally, candidates completed the two video-based situational judgement tests, namely 'physician–patient interaction' and 'medical expert discussion'. Prior to each of these tests, they received background information (e.g., patient's medical history and actual problems).

To obtain the personality inventory data, the NEO PI-R was administered to the students during classes in each of the five Flemish universities. This administration took place in the first year of Medical and Dental studies. Students were informed about the purpose of the study and it was announced that they would receive individual feedback, which was available through their student number. We also emphasized that study participation was voluntary and that students could end their participation at any time. They were assured that the results only served research purposes and would not influence their exam results. The administration of the personality inventory lasted between 30 and 50 minutes.

The criterion measures (i.e., first year final scores) were retrieved from archival records of the five Flemish universities.

Results

Descriptive Statistics

Table 1 presents the means and standard deviations of this study's variables, together with their intercorrelations. The internal consistencies of the various tests are also displayed. The mean score on the Admission Exam was 12.30 (SD = 1.45). Some 683 of the 941 candidates successfully passed the Admission Exam (selection rate of 72.58 per cent). The Situational Test Score (M = 7.76) was significantly higher than the Cognitive Ability Test Score (M = 4.54). The standard deviation of the Situational Test Score was also smaller (SD = .73 vs. SD = 1.00 for the Cognitive Ability Test Score).

Internal consistency coefficients were acceptable (between .70 and .92) for the measures of cognitive ability. This was not the case for the situational tests (between .41 and .56). Note, however, that because situational tests typically measure heterogeneous content, internal consistency is not an appropriate reliability coefficient (as opposed to test–retest reliability) (Chan and Schmitt 1997; Clause, Mullins, Nee, Pulakos and Schmitt 1998; Motowidlo and Tippins 1993).

Predictive Validity

As a first issue in terms of predictive validity we examined which of the situational tests emerged as predictors of students' first year scores. Inspection of Table 1 (last row) shows that the correlation between the Admission Exam Score and the final first year score was .35 ($p < .01$) (corrected for direct restriction of range, Thorndike's 1949, case 2). To give these correlations some practical value, we counted how many of the participants, who passed the Admission Exam, successfully completed the first year of medical and dental studies. After the first exam period 56.76 per cent of the students successfully completed their first year exams. After the second exam period, which included only students failing in the first exam period, this percentage increased to 72.95 per cent.

Both the weighted Cognitive Ability Test Score ($r = .27$, $p < .01$) and the Situational Test Score ($r = .23$, $p < .01$) showed significant (corrected) correlations with the final first year score. Among the specific tests, the reasoning test yielded the largest correlation ($r = .33$, $p < .01$). For the situational tests the lecture ($r = .20$, $p < .01$) and the medical course text ($r = .21$, $p < .01$) showed the largest correlations. Note also that the correlation between Conscientiousness and first year grades in medical and dental studies was .20 ($p < .01$). Extraversion correlated −.10 ($p < .05$) with first year scores. As mentioned above, the NEO-PI-R was not part of the Admission Exam (we administered this inventory during the first year).

Next, we investigated whether the situational tests showed incremental validity over the cognitive ability measures. In a hierarchical regression analysis (see Table 2) the four cognitive ability measures were entered as a first block in the regression equation and explained 10.4 per cent of the criterion variance, F (4, 605) = 17.62, $p < .001$. The reasoning test ($\beta = .26$, $p < .001$) emerged as the most important predictor. The regression weight of the visual information processing test was also significant but in the opposite direction ($\beta = -.09$, $p < .05$). Consistent with our expectations, the four situational tests, which were entered as a second block into the regression equation, explained an additional significant portion of the variance, namely 3.1 per cent, F (8, 601) = 11.74, $p < .001$. Both the videotaped lecture ($\beta = .11$, $p < .01$) and the medical text ($\beta = .11$, $p < .01$) emerged as significant predictors among the situational tests. This is also in line with our expectations.

Because range restriction may also affect the regression results, we applied the Lawley (1943, cited in Bobko 1995 and Ree, Carretta, Earles and Albert 1994) multivariate range restriction correction to the entire matrix of correlations (see Table 1) and used this corrected matrix as input for the hierarchical regression analysis. The range corrected Rs were slightly higher than in the aforementioned analysis (R = .136 after first step and R = .169 after second step). However, the regression weights were similar to the ones reported above.

Finally, we used the formula put forth by Cattin (1980) and estimated cross-validity. This value equalled .34 ($R^2 = .11$) illustrating that the validity estimates of the Admission Exam might also be meaningful in other and different samples.

Adverse Impact (Gender)

Table 3 presents the descriptive statistics on the various test and composite scores, broken down by candidate gender. Effect sizes are also displayed. Regarding the cognitive ability tests, in line with our expectations, men scored better than women on the visual information processing test by .25 standard deviation and women obtained significantly higher scores on the weighted situational test score ($d = .18$). Further, there were significant mean differences favouring women over men for the physician–patient interaction ($d = .30$) and the medical expert discussion ($d = .16$). As expected, however, these mean gender differences at the test level averaged out in the composite Admission Exam Score, which showed no significant gender difference. This was also evidenced by comparing pass rates of men and women (71.59 per cent vs. 72.85 per cent).

Besides looking at mean subgroup differences, adverse impact was also examined via the regression model of test bias (American Educational Research Association, American Psychological Association and National Council on Measurement in Education 1998; Society for Industrial and Organizational Psychology 1987). This Cleary (1968)

Table 2: Summary of hierarchical regression analysis of tests of admission exam on final score in first year medical and dental studies (N = 609) with and without correction for multivariate range restriction

	No correction for range restriction			Correction for range restriction		
Test	*b*	*SE b*	−*β*	*b*	*SE b*	−*β*
Step 1						
Reasoning	.18	.02	.29**	.20	.02	.34**
Visual Information	−.06	.03	−.09*	−.06	.03	−.09*
Memory Association	.05	.03	.06	.06	.03	.07
Pattern Recognition	.02	.01	.06	.02	.01	.07
Step 2						
Reasoning	.16	.03	.26**	.16	.03	.27**
Visual Information	−.06	.03	−.09*	−.06	.03	−.09*
Memory Association	.04	.03	.05	.05	.04	.05
Pattern Recognition	.02	.01	.06	.02	.01	.06
Lecture	.09	.03	.11**	.09	.03	.11**
Medical text	.18	.06	.11**	.18	.06	.12**
Physician–patient interaction	.03	.05	.03	.03	.05	.03
Medical expert discussion	−.07	.05	−.06	−.07	.05	−.06

Note: When no correction for multivariate range restriction was applied, $R^2 = .104$ for Step 1; $\Delta R^2 = .031$ for Step 2 ($p < .01$). When correction for multivariate range restriction was applied, $R^2 = .136$ for Step 1; $\Delta R^2 = .033$ for Step 2 ($p < .01$).
*$p < .05$; **$p < .01$.

Table 3: Means and standard deviations of admission exam tests and scores for men and women

Test	Men (N = 359)		Women (N = 582)		*t*-value	Effect size
	M	SD	M	SD		
Reasoning	26.62	6.34	26.06	5.68	1.36	−0.09
Visual information	12.56	5.33	11.27	5.10	3.71**	−0.25
Memory association	8.36	3.90	9.55	3.77	−4.64**	0.31
Pattern recognition	24.40	10.76	23.53	10.82	1.20	−0.08
Lecture	30.27	4.46	30.25	4.22	.07	0.00
Medical text	15.04	2.39	15.27	2.35	−1.44	0.10
Physician–patient interaction	22.88	3.09	23.80	2.90	−4.53**	0.30
Medical expert discussion	23.89	2.97	24.34	2.88	−2.33*	0.16
Cognitive ability test score	4.61	1.06	4.50	0.97	1.58	−0.11
Situational test score	7.67	0.77	7.81	0.71	−2.69**	0.18
Admission exam score	12.28	1.56	12.31	1.37	−.25	0.02

Notes: *$p < .05$; **$p < .01$. Positive effect sizes reflect differences that favour women, whereas negative effect sizes reflect differences that favour men.

model of test bias investigates whether mean subgroup differences in test scores are related to mean subgroup differences in the criterion. In line with this model we examined evidence for differential prediction through a hierarchical regression analysis of the criterion (i.e., final first year scores) on a predictor (i.e., a specific test of the Admission Exam), then gender, and then the product of the predictor and gender (Bartlett *et al.* 1978). We carried out such a hierarchical regression analysis for each of the eight tests of the Admission Exam. In none of these hierarchical regression analyses did gender or the product of gender and a specific test add significantly to the variance explained. These results suggest that there was no evidence of differential validity or over-/under-prediction as a function of candidate gender.

Construct Validity

As noted above, the construct validity of the situational tests was examined by placing them in a nomological net with measures of cognitive ability and personality. This was done through canonical correlation analysis, which investigates the degree of relationship between two sets of variables. For each set of variables a linear composite of the variables is determined so that these two linear composites maximally correlate. The linear composite of a set of variables is also known as the canonical variable. The correlation of the canonical variable of one set of variables with the canonical variable of the other set of variables is the canonical correlation coefficient. Note that the following always presents the solution with the highest canonical coefficient. The main advantage of canonical correlation analysis over bivariate correlations (see Table 1) is that the multivariate treatment of the data takes the interrelationships among the variables into account. More information on canonical correlation analysis is given by Tabachnick and Fidell (1996).

Two canonical correlation analyses were performed. In the first canonical correlation analysis we examined the relationship between the set of situational tests and the set of cognitive ability measures. We expected that the lecture and the medical text would be related to measures of cognitive ability. The results are displayed in Table 4. The canonical correlation coefficient equalled .44 ($p < .001$). This means that the measures of cognitive ability and the situational tests share 19.5 per cent common variance. Inspection of the standardized canonical variate coefficients reveals that with respect to the set of cognitive ability measures this substantial canonical correlation was mainly determined by the reasoning test (.93). In line with our predictions the lecture (.74) and the medical text (.77) mainly determined this canonical correlation on the situational test side. Note that we use here a cut-off of > .70 (i.e., more than 50 per cent shared variance) for interpreting the standardized canonical correlations.

In the other canonical correlation analysis the situational tests were related to the Big Five scores on the NEO-PI-R. The canonical correlation coefficient between the linear personality composite and the linear situational test composite equalled .26 ($p < .001$). In other words, only 7 per cent of the variance of the situational tests was explained by the personality factors. Therefore, the results presented in Table 5 should be interpreted with caution (Tabachnick and Fidell 1996). The standardized canonical variate coefficients show that the variance in the canonical variable X (personality factors) was primarily determined by the factor Openness (.74). This means that the Openness scale was important to determine the maximal correlation between the two canonical variables. The physician–patient interaction and also the medical text determined the canonical variable Y (situational tests). In ancillary analyses (available from the authors) we looked at which Openness facets correlated with the situational tests. Virtually all Openness facets had significant correlations (ranging from $r = .10$ to .15) with the videotaped physician–patient interaction and the weighted situational test score.

Table 4: Results of canonical correlation analysis between cognitive ability measures (canonical variable X) and situational tests (canonical variable Y)

	Standardized Canonical Variate Coefficient
Cognitive ability set (Canonical Variable X)	
Reasoning	.93
Visual information	.28
Memory association	.56
Pattern recognition	.37
Situational test set (Canonical Variable Y)	
Lecture	.74
Medical text	.77
Physician-patient interaction	.64
Medical expert discussion	.58
Canonical correlation between X and Y	.44**

Notes: N = 941. ***p* < .01.

Table 5: Results of canonical correlation analysis between NEO-PI-R (canonical variable X) and situational tests (canonical variable Y)

	Standardized Canonical Variate Coefficient
Personality set (Canonical Variable X)	
Extraversion	−.15
Agreeableness	−.29
Conscientiousness	−.30
Emotional stability	.13
Openness	.74
Situational test set (Canonical Variable Y)	
Lecture	.39
Medical text	.77
Physician-patient interaction	.70
Medical expert discussion	.55
Canonical correlation between X and Y	.26**

Notes: N = 529. ***p* < .01.

Discussion

This study focused on the development and validation of the situational tests included in the Flemish Admission Exam 'Medical and Dental Studies'. First of all, our results show that, in the specific context of student selection, situational tests yield reasonable validity estimates and significantly predict over traditional cognitive ability measures. Among the situational tests the videotaped lecture and the medical text emerge as significant predictors. The physician–patient interaction and the medical expert discussion do not emerge as significant predictors.

Although the incremental explained variance of 4 per cent may seem small, given the high costs associated with situational test development, we believe that this result is more encouraging than disappointing. This is because the criterion measure used (final first year score) was mainly comprised of scores on science-related subjects. This heavily cognitive-based criterion also provides a good explanation for the lower predictive validity of the physician–patient interaction and medical expert discussion. When the final score in the clinical years serves as the criterion measure, higher predictive validities for the physician–patient interaction and the medical expert discussion can be expected (Glaser, Hojat, Veloski, Blacklow and Goepp 1992). This exemplifies the need for other predictive validity studies of situational tests in different samples (i.e., other populations) with less *g*-loaded criteria.

Besides the nature of our criterion measure there are at least two other explanations for our predictive validity results. In particular, previous studies have revealed that experience is related to situational test performance (Smith and McDaniel 1998; Weekley and Jones 1997; 1999). Therefore, the fact that participants, who have never experienced situations such as an interaction with a patient or a medical expert discussion, and are asked to indicate how they would handle these situations, may also explain why these tests do not emerge as significant predictors. Consistent with this explanation, it is striking that only scores on tests consisting of situations, which candidate medical students have already experienced (i.e., following a lecture and studying course material), are predictive.

Another explanation is related to the differences in response fidelity of the situational tests included. Along these lines, Funke and Schuler (1998) demonstrated that response fidelity instead of stimulus fidelity moderated the criterion-related validity of situational tests. In our study it is noteworthy that the video-based situational judgement tests, which typically have lower response fidelity (i.e., physician–patient interaction and medical expert discussion), show also lower predictive validity. In these video-based situational judgement tests students have to pick the correct response alternative instead of constructing the answer or acting it out. Alternatively, the miniaturized work samples, which have higher response fidelity (i.e., lecture and course text), have also higher predictive validity. Here participants respond to the stimulus in a more realistic way as they take notes of the lecture and probably make a summary of the course text. However, because the situational tests differ both in terms of content and response fidelity (Chan and Schmitt 1997), definite conclusions about the viability of this last explanation are not possible.

With respect to predictive validity, it is also striking that the personality factor Conscientiousness does nearly as well as the situational tests in predicting first year scores. The correlation between Conscientiousness and first year grades in medical and dental studies is .20. In addition, Extraversion correlated −.10 with first year scores. The significant correlation between Conscientiousness and first year grades corroborates previous results in the specific field of academic medicine (Ferguson, Sanders, O'Hehir and James, 2000) and in the general field of academic achievement (Blickle 1996; Busato, Prins, Elshout and Hamaker, 2000; De Fruyt and Mervielde 1996; De Raad 1996; De Raad and Schouwenberg 1996; Geisler-Brenstein, Schmeck and Hetherington 1996; Goff and Ackerman 1992; Rothstein, Paunonen, Rush and King 1994; Wolfe and Johnson 1995). These results do not confirm the decision of the Flemish Admission Exam Commission to leave personality questionnaires out of the admission exam.

Second, this study shows that there are no mean gender differences when situational tests are combined with cognitive ability measures. In other words, the mean gender differences at the test level are balanced out if *both* situational tests and cognitive ability measures are used. Recently, Pulakos and Schmitt (1996) reached similar conclusions in terms of diminishing adverse impact in terms of race. They found that subgroup differences between Blacks and Whites could be considerably reduced by assessing a broad array of both cognitive and noncognitive abilities (e.g., interpersonal skills) (see also Schmitt, Rogers, Chan, Sheppard and Jennings 1997).

The gender differences found at the test level are consistent with prior research. For instance, parallel with research on mental ability testing (Jensen 1998), men score higher on visualization and pattern recognition and women score higher on memory tasks. In addition, women score higher on all four situational tests (see also Motowidlo *et al.* 1990; Motowidlo and Tippins 1993; Weekley and Jones 1997; 1999). An important remaining question is why situational tests seem to give a slight edge to women. According to Weekley and Jones (1999), the interpersonal nature of many problems in situational tests tends to favour women. Future research is needed to confirm this explanation.

A third conclusion is that the use of situational tests in student selection enables us to measure a broader range of skills and abilities. Although the videotaped lecture and the medical text are still related to cognitive ability (i.e., reasoning test), our canonical correlation analyses also show that personality factors and particularly Openness contribute to higher scores on situational tests. This was especially true for the video-based situational judgement test, which required candidate medical students to react to situations of an interaction between a physician and a patient. According to Hoekstra *et al.* (1996), people high on Openness are more imaginative, intellectually curious, and independent. They typically like variation, have a broader range of interests, and are more willing to learn new things. Such personality characteristics have been found to lead to better performance in the so-called pre-clinical years (Peng, Khaw and Edariah 1995). Therefore, it is good news that they are related to higher performance on one of the tests used to select medical students.

In sum, because traditional sign-based selection procedures used in admission exams are often criticized for their narrow coverage, alternatives have been sought. In this study we examined one of the possible alternatives, namely, the use of situational tests. This study shows that situational tests are a useful complement to the traditional student selection procedures in terms of enhancing predictive validity, reducing adverse impact (regarding gender), and broadening the constructs measured.

Acknowledgements

This research was supported by a grant as Postdoctoral Fellow of the Fund for Scientific Research-Flanders (Belgium) (F.W.O.). We would like to thank Michael McDaniel for his helpful comments on a previous version of this manuscript. We are also indebted to Piet Janssen for his valuable help throughout the project.

Parts of this study were presented at the European Congress of Work and Organizational Psychology, Espoo-Helsinki, Finland (1999, May).

Note

[1] Eventually, this part was suspended by a decision of a special court in Belgium ('Arbitragehof').

References

Aldrich, C.K. (1987) Psychiatric interviews and psychological tests as predictors of medical students' success. *Journal of Medical Education*, **62**, 658–664.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, and NCME) (1998) *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Anastasi, A. and Urbina, S. (1997) *Psychological Testing*. 7th edn. Upper Saddle River, NJ: Prentice-Hall.

Baker, E.L. O'Neil, H.F. and Linn, R.L. (1993) Policy and validity prospects for performance based assessment. *American Psychologist*, **48**, 1210–1218.

Bartlett, C.J., Bobko, P., Mosier, S.B. and Hannon, R. (1978) Testing for fairness with a moderated multiple regression strategy: An alternative for differential analysis. *Personnel Psychology*, **31**, 233–241.

Blickle, G. (1996) Personality traits, learning strategies, and performance. *European Journal of Personality*, **10**, 337–352.

Bobko, P. (1995) *Correlation and Regression: Principles and Applications for Industrial/Organizational Psychology*. New York: McGraw-Hill.

Busato, V.V., Prins, F.J., Elshout, J.J. and Hamaker, C. (2000) Intellectual ability, learning style, personality, achievement and academic success of psychology students in higher education. *Personality and Individual Differences*, **29**, 1057–1168.

Cattin, P. (1980) Estimation of a regression model. *Journal of Applied Psychology*, **65**, 407–414.

Chan, D. and Schmitt, N. (1997) Video-based versus paper-and-pencil method of assessment in situational judgement tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, **82**, 143–159.

Clause, C.C., Mullins, M.E., Nee, M.T., Pulakos, E.D. and Schmitt, N. (1998) Parallel test form development: A procedure for alternative predictors and an example. *Personnel Psychology*, **51**, 193–208.

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

Cohen, J.A. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **109**, 37–46.

Costa, P.T. and McCrae, R.R. (1992) *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Odessa: PAR.

De Fruyt, F. (1996) Personality and vocational interests. Relationship between the Five-Factor Model of Personality and Holland's RIASEC typology. Unpublished doctoral dissertation, University of Ghent, Belgium.

De Fruyt, F. and Mervielde, I. (1996) Personality and interests as predictors of educational streaming and achievement. *European Journal of Personality*, **10**, 405–425.

De Raad, B. (1996) Personality traits and education. *European Journal of Personality*, **10**, 185–200.

De Raad, B. and Schouwenberg, H.C. (1996) Personality in learning and education: A review. *European Journal of Personality*, **10**, 303–336.

Dunbar, S.B. and Novick, M.R. (1988) On predicting success in training for men and women: Examples from Marine Corps clerical specialties. *Journal of Applied Psychology*, **73**, 545–550.

Ferguson, E., Sanders, A., O'Hehir, F. and James, D. (2000) Predictive validity of personal statements and the role of the five-factor model of personality in relation to medical training. *Journal of Occupational and Organizational Psychology*, **73**, 321–344.

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

Gaugler, B.B., Rosenthal, D.B., Thornton, G.C. and Bentson, C. (1987) Meta-analysis of assessment center validity. *Journal of Applied Psychology*, **72**, 493–511.

Geisler-Brenstein, E. Schmeck, R.R. and Hetherington, J. (1996) An individual difference perspective on students' diversity.

Higher Education, **31**, 73–96.

Glaser, K., Hojat, M., Veloski, J.J., Blacklow, R.S. and Goepp, E.C. (1992) Science, verbal or quantitative skills: Which is the most important predictor of physician competence? *Educational and Psychological Measurement*, **52**, 395–406.

Goff, M. and Ackerman, P.L. (1992) Personality-intelligence relations: Assessment of typical intellectual engagement. *Journal of Educational Psychology*, **84**, 537–552.

Gough, H.G. and Hall, W.B. (1983) Prediction of performance in medical school from the California Psychological Inventory. *Journal of Applied Psychology*, **48**, 218–226.

Green, A., Peters, T.J. and Webster, J.T. (1991) An assessment of academic performance and personality. *Medical Education*, **25**, 343–348.

Green, A., Peters, T.J. and Webster, D.J.T. (1993) Preclinical progress in relation to personality and academic profiles. *Medical Education*, **27**, 137–142.

Hobfoll, S.E., Anson, O. and Antonovsky, A. (1982) Personality factors as predictors of medical student performance. *Medical Education*, **16**, 251–258.

Hoekstra, H.A., Ormel, J. and De Fruyt, F. (1996) *NEO persoonlijkheidsvragenlijsten NEO-PI-R en NEO-FFI*. Handleiding. Lisse: Swets and Zeitlinger.

Hojat, M., Robeson, M., Damjanov, I., Veloski, J.J., Glaser, K. and Gonnella, J.S. (1993) Students' psychosocial characteristics as predictors of academic performance in medical school. *Academic Medicine*, **68**, 635–637.

Hough, L.M. (1984) Development and evaluation of the 'accomplishment record' method of selecting and promoting professionals. *Journal of Applied Psychology*, **69**, 135–146.

Jensen, A.R. (1998) *The g Factor: The Science of Mental Ability*. Westport, CT: Praeger.

Latham, G.P., Saari, L.M., Pursell, E.D. and Campion, M.A. (1980) The situational interview. *Journal of Applied Psychology*, **65**, 422–427.

Lievens, F. (2000) Development of an empirical scoring scheme for situational inventories. *European Review of Applied Psychology*, **50**, 117–124.

Lievens, F. and Coetsier, P. (1998) A different look at selection of candidate medical students: Development of video-based simulations [In Dutch]. *Tijdschrift voor Hoger Onderwijs*, **16**, 117–131.

Linn, R.L., Baker, E.L. and Dunbar, S.B. (1991) Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, **20**, 15–21.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A. and Braverman, E.P. (in press) Predicting job performance using situational judgement tests: A clarification of the literature. *Journal of Applied Psychology*.

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

McManus, I.C. (1982) A-level grades and medical school admission. *British Medical Journal*, **284**, 1654.

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, **23**, 13–23.

Ministerie van de Vlaamse Gemeenschap (1996) Decreet houdende wijziging van het decreet van 12 juni 1991 betreffende de universiteiten in de Vlaamse Gemeenschap. *Belgisch Staatsblad 19.09.96*.

Minnaert, A. (1996) Academic performance, cognition, metacognition and motivation. Assessing freshmen characteristics on task: A validation and replication study in higher education. Unpublished doctoral dissertation, University of Louvain, Belgium.

Mitchell, K., Haynes R. and Koenig J. (1994) Assessing the validity of the updated Medical College Admission Test. *Academic Medicine*, **69**, 394–401.

Montague, W. and Odds, F.C (1990) Academic selection criteria and subsequent performance. *Medical Education*, **24**, 44–47.

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

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

Neisser, U., Boodoo, G., Bouchard, Jr, T.J., Wade Boykin, A., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J. and Urbina, S. (1996) Intelligence: Knowns and unknowns. *American Psychologist*, **51**, 77–101.

Neubauer, R. (1990) Women in the career assessment center – a victory?' [In German]. *Zeitschrift für Arbeits und Organisationspsychologie*, **34**, 29–36.

Nguyen N.T. and McDaniel, M.A. (2001) Constructs assessed in situational judgement tests: A meta-analysis. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA, April.

Peng, R., Khaw, H.H. and Edariah, A.B. (1995) Personality and performance of preclinical medical students. *Medical Education*, **29**, 283–288.

Powis, D.A. (1994) Selecting medical students. *Medical Education*, **28**, 443–469.

Pulakos, E.D. and Schmitt, N. (1996) An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance*, **9**, 241–258.

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

Roberts, G.D. and Porter, A.M.W. (1989) Medical student selection in time for change: Discussion paper. *Journal of the Royal Society of Medicine*, **82**, 288–298.

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

Roessler, R., Lester, J.W., Butler, W.T., Rankin, B. and Collins, F. (1978) Cognitive and non-cognitive variables in the prediction of preclinical performance. *Journal of Medical Education*, **53**, 678–681.

Rothstein, M.G., Paunonen, S.V., Rush, J.C. and King, A. (1994) Personality and cognitive ability predictors of performance in graduate Business School. *Journal of Educational Psychology*, **86**, 516–530.

Sackett, P.R. (1998) Performance assessment in education and professional certification: Lessons for personnel selection? In M.D. Hakel (ed.), *Beyond Multiple Choice Tests* (pp. 113–129) Mahwah, NJ: Lawrence Erlbaum.

Salgado, J.F. and Lado, M. (2000) Validity generalization of video tests for predicting job performance ratings. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Schmidt, F.L. and 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.

Schmitt, N. (1993) Group composition, gender, and race effects on assessment center ratings. In H. Schuler, J.L. Farr and M. Smith (eds), *Personnel Selection and Assessment: Individual and Organizational Perspectives* (pp. 315–332), Hillsdale, NJ: Lawrence Erlbaum.

Schmitt, N. and Ostroff, C. (1986) Operationalizing the 'behavioural consistency' approach: Selection test development based on a content-oriented approach. *Personnel Psychology*, **39**, 91–108.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L. and Jennings, D. (1997) Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology*, **82**, 719–730.

Shen, H. and Comrey, A.L. (1997) Predicting medical students' academic performances by their cognitive abilities and personality characteristics. *Academic Medicine*, **72**, 781–786.

Shore, T.H. (1992) Subtle gender bias in the assessment of managerial potential. *Sex Roles*, **27**, 499–515.

Smith, K.C. and McDaniel, M.A. (1998) Criterion and construct validity evidence for a situational judgement measure. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Society for Industrial and Organizational Psychology (1987) *Principles for the Validation and Use of Personnel Selection Procedures*. College Park, MD: Author.

Sternberg, R.J. and Kaufman, J.C. (1998) Human abilities. *Annual Review of Psychology*, **49**, 479–502.

Tabachnick, B.G. and Fidell, L.S. (1996) *Using Multivariate Statistics*. New York: HarperCollins College Publishers.

Tate, P. (1994) *The Doctor's Communication Handbook*. Oxford and New York: Radcliffe Medical Press.

Thorndike, R.L. (1949) *Personnel Selection: Test and Measurement Techniques*. New York: Wiley.

Thornton, G.C. III (1992) *Assessment Centers in Human Resource Management*. Reading, MA: Addison-Wesley.

Tomlinson, R.W.S., Clack G.B., Pettingale, K.W., Anderson, J. and Ryan, K.C. (1977) The relative role of 'A' level chemistry, physics and biology in the medical course. *Medical Education*, **11**, 103–108.

Vu, N.V., Dawson-Saunders, B. and Barrows, H.S. (1987) Use of Medical Reasoning Aptitude Test to help predict performance in medical school. *Journal of Medical Education*, **62**, 325–335.

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

Weekley, J.A. and Jones, C. (1999) Further studies of situational tests. *Personnel Psychology*, **52**, 679–699.

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

Wiggins, G. (1989) A true test: Toward more authentic and equitable assessment. *Phi Delta Kappa*, **70**, 703–713.

Wolfe, R.N. and Johnson, S.D. (1995) Personality as a predictor of college performance. *Educational and Psychological Measurement*, **55**, 177–185.