The use of Situational Judgment Tests in Admission to Higher Education: Validity and Coaching Effects

Tine Buyse

Promotor: Prof. Dr. F. Lievens

Proefschrift ingediend tot het behalen van de academische graad van Doctor in de Psychologie

2011

DANKWOORD

Een doctoraat maak je zeker niet alleen. Ik ben dan ook heel veel mensen dank verschuldigd.

Mijn eerste woord van dank gaat naar Filip, mijn promotor. Meer dan 10 jaar samenwerken is niet iets wat ik zomaar in enkele zinnen kan neerschrijven. We hebben deze weg samen afgelegd. Bedankt dat ik altijd bij je mocht komen uitrazen, voor onze leuke uitstapjes naar Brussel en elders. Bedankt voor je geloof in mij, je enthousiasme en vooral je steun bij al de keuzes die ik doorheen de jaren maakte.

Dank aan de meer dan 30.000 kandidaten die sinds 1997 deelnamen aan het toelatingsexamen en die ik niet persoonlijk ken. Zij hebben onwetend de belangrijkste bijdragen aan dit werk geleverd.

Graag bedank ik ook alle ex-collega's van de vakgroep PP09 voor hun belangstelling, het kritisch meedenken, het luisteren en de gezelligheid. Het waren 10 mooie jaren.

Papa en mama, bedankt voor alles!!

Pa en ma, familie en vrienden, bedankt voor jullie interesse, steun en voor alle leuke dingen die we samen doen.

Een heel bijzonder woord van dank gaat naar Aude en Jade. Voor het verdragen van alle woensdagen, weekends en vakantiedagen dat ik niet helemaal jullie mama kon zijn, voor jullie liefde en knuffels. Omdat jullie mijn kleine kippetjes zijn.

En uiteindelijk mijn lieve Jan, bedankt voor je steun, begrip en liefde en voor alle avonden samen in de zetel met de laptop op schoot. Bedankt omdat je altijd achter me staat en omdat je van mij de gelukkigste vrouw ter wereld maakt.

CONTENTS

CHAPTER 1

GENERAL INTRODUCTION

In the first chapter of this dissertation, we provide a general introduction to the subject by giving an overview of the literature on situational judgment tests (SJTs). First, we define an SJT in general terms. Second, we discuss features and psychometric properties of SJTs. We discuss and review the large body of literature on the use of SJTs in employment settings and the rare studies in a high-stakes selection context. Next, we describe the setting of this dissertation: the admission to medical and dental studies in Flanders. An overview of the origin, the development, and the procedure of the Admission Exam is discussed, and is compared to admission systems in other countries. All of this exemplifies the common thread running through this dissertation: the use of SJTs in high-stakes selection settings.

INTRODUCTION

"*I held his admission interview in the medical school cafeteria. I sensed his passion to become a physician. He communicated easily. He described the strong sense of connection he had felt with the patients at the free clinic at which he had volunteered. While I wasn't yet sure what a great physician was, I had an intuitive sense he would become one. Yet the decision was "His science grades aren't strong enough. Reject." I felt personally bruised. But then, I was only the student [chosen as a full member on the school's admissions committee]—what did I know?" (Barr, 2010a, p. 678)*

This anecdote is only one of many examples which indicate what society expects of a 'good' doctor. Both technical knowledge and interpersonal skills are important. Powis (2010) states that any competency list for a generic medical practitioner should comprise excellent academic ability and good cognitive skills but practitioners should also have well developed decision making skills, professional integrity, and excellent interpersonal skills, in addition to being accomplished and confident communicators who can empathize with patients. Makoul & Curry (2007) recommend that, in order to improve quality of care, initiatives could include more systematically assessing interpersonal skills during the admissions process and ensuring that clinical skills assessments include a communications component. However, today still many medical selection systems rely only on tests that measure cognitive knowledge and ability. In other countries, interviews and personality tests are widely used to measure interpersonal characteristics. In recent years, SJTs have drawn the attention of many researchers. There is recent evidence that SJTs might be valuable supplements to extant cognitive tests in admission contexts.

Therefore, the main objective of this doctoral dissertation is to examine the potential use of SJTs in medical admission contexts. This first chapter provides an introduction to SJTs

and presents an overview of relevant previous research. The admission context in which our SJT was used is described, and compared to the international context. Next, large-scale results of selection instruments in admission contexts are discussed. On the basis of this literature review, the research questions of the present dissertation are identified at the end of the chapter.

SITUATIONAL JUDGMENT TESTS

Definition

SJTs present applicants with different work-related situations. Applicants have to indicate the appropriate response alternative from a list of different response options (Motowidlo, Dunnette, & Carter, 1990; Motowidlo, Hanson, & Crafts, 1997; Motowidlo & Tippins, 1993; Weekley & Jones, 1999). The answers to SJT questions typically require common sense, experience, and common knowledge, more than logic reasoning abilities or high intelligence. Therefore, SJTs are categorized as non-cognitive tests. The first prototype of an SJT dates back to 1926, namely the 'Judgment in Social Situations' which was a subtest of the George Washington Social Intelligence Test (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). This test is probably the first widespread and largely evaluated SJT. In World War II, psychologists tried to measure the insight and judgment of soldiers and in the 1960s, tests were developed to measure the leadership potential of applicants (McDaniel et al., 2001). Examples are the 'Practical Judgment Test (Cardall, 1942), the 'How Supervise?' (File, 1945; File & Remmers, 1948), and the 'Supervisory Practices Test' (Greenberg, 1963). However, the widespread use of SJTs was practically nonexistent until the modern version of the SJT was "reinvented" by Motowidlo et al. (1990). Thanks to these researchers, a new interest in SJTs emerged and still exists today.

Features

The last two decades, these modern SJTs are used in research settings and in applicant selection situations. While modern SJTs vary on many characteristics, they have a few features in common. First, SJTs are based on the assumption that behavior is consistent. According to this "behavioral consistency principle" (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968), the best predictor of applicants' future behavior is past behavior. More specifically, the performance on a realistic selection test (closely corresponding to the future job) will be consistent and therefore predictive of later job performance. Second, SJTs present applicants with realistic situations. They give applicants a realistic job preview. However, the specific way that the situation is presented to the applicant can vary. The realistic situations can be shown on video or computer. SJTs can also be presented as paper-and-pencil or written tests. Third, SJTs mostly use the multiple-choice answering format. Again, the different options can be presented on paper, or digitally. Applicants are not asked to act out their chosen response. In this respect, SJTs differ from assessment centers, where the candidate is asked to act out his/her response. SJTs are highly standardized and can be administrated to large groups (unlike assessment centers). Thus, in SJTS, there are many ways to present situations and alternative answers to the applicants. SJTs can differ a great deal on these features. Moreover, there are two ways to present response instructions: knowledge based instructions and behavioral tendency instructions (McDaniel & Nguyen, 2001; Ployhart & Ehrhart, 2003). Knowledge based instructions ask candidates to identify the right answer ("What should you do?"). On the other hand, behavioral tendency instructions ask the candidate how he or she would react in a particular situation ("What would you do?"). Prior research has provided much insight in the differences between these two formats. In general, higher mean scores are found on SJTs with knowledge based instructions (McDaniel,

et al., 2001). The meta-analysis of McDaniel et al. also found that knowledge instructions and behavioral tendency instructions have equal criterion-related validity.

Development and scoring

Motowidlo et al. (1990, 1997) describe the three typical stages necessary in developing an SJT. First, through a thorough job analysis, critical incidents that are encountered on the job are collected from subject matter experts (SMEs). SMEs are people who know the job very well (supervisors, customers, experienced workers) (Flanagan, 1954). Critical incidents emphasize very good or very bad behaviors in work situations. The test developer groups these incidents into similar content areas, selects representative scenarios from each content area (Motowidlo et al., 1997), and constructs item stems of similar length and format. In the second phase, SMEs are asked to generate different responses to each work situation. They have to identify what they would most likely do or what they think is the best thing to do. SMEs should be able to identify the best response and other, less excellent possible reactions. After this, the test developer can sort all response alternatives on a range of effectiveness. In most cases, four response alternatives are constructed. Finally, the scoring key is developed. SJT scoring keys are often developed using another pool of SMEs or excellent employees. These experts judge the effectiveness of each response alternative, or they identify the best and the worst response. The development of the scoring key described above, is the expert-based scoring approach. Various other scoring methods for multiple-choice SJTs, such as the empirically-derived scoring key, are discussed in the SJT literature (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; Hogan, 1994; Weekley & Jones, 1997, 1999; Weekley, Ployhart, & Holtz, 2006).

As can be seen, SMEs are used in each phase of development. The realism of the stems and response options is high when experts are used. Moreover, a large group of experts generates a large pool of incidents and possible responses for each situation.

Research

Since the reinvention of the SJT by Motowidlo et al. (1990), many studies have examined the effectiveness of SJTs. Therefore, the strengths and weaknesses of SJT use are rather easy to describe. The efficacy and efficiency of SJTs is described below in an evidence-based overview. Different psychometric criteria are discussed including reliability, criterion-related and incremental validity, adverse impact, and coaching and practice effects.

Reliability. This refers to the consistency of the test scores in different conditions (over time, concerning item content). As SJTs are designed to measure multiple constructs, internal consistency estimations are not appropriate indications of reliability. In most cases, SJTs are multidimensional at the item level (Clause, Mullins, Nee, Pulakos, & Schmitt, 1998). Many researchers report internal consistency coefficients of SJTs. The meta-analysis of McDaniel et al. (2001) presents coefficients varying from .43 to .94 (average .60). Chan and Schmitt (2002) report a value of .73 (40 item SJT). Test-retest reliabilities are more appropriate, but often not available. Ployhart, Porr, and Ryan (2004) report a test-retest reliability of .84.

Criterion-related validity. Many studies have investigated the criterion-related validity of SJTs. In their meta-analysis, McDaniel et al. (2001) analyzed the criterion-related validities of SJTs across 95 studies and concluded that SJTs are valid predictors of job performance (corrected *r* of .34). However, most studies in this meta-analysis were concurrent in design and did not involve the use of SJTs in operational settings. Moreover, there was a marked difference between the mean validity coefficient for predictive study designs (corrected *r* of .18) and that for concurrent study designs (corrected *r* of .35). A second meta-analysis by

McDaniel, Hartman, Whetzel, and Grubb (2007) reported a mean corrected validity of .26. Again, the number of concurrent study designs was large (114 out of 118). Third, Christian, Edwards, and Bradley (2010) showed that the validity of SJTs was higher for predicting conceptually-related performance dimensions, eventually underscoring the importance of matching predictor and criterion. Only 6 out of 84 studies included in this meta-analysis were predictive validity studies.

Incremental Validity. Research has indicated that SJTs significantly add to the prediction of job performance over cognitive ability, the Big Five, job knowledge, and job experience (Chan & Schmitt, 2002; McDaniel et al., 2001; Weekley & Jones, 1997, 1999; Weekley & Ployhart, 2006). In the meta-analysis of McDaniel et al. (2007), SJTs accounted for additional variance (varying from 1% to 2%) over both cognitive ability and personality. Chan and Schmitt (2002) found an incremental validity varying from 3% to 8% over and above cognitive ability, the Big Five, and job experience. These results were replicated in educational settings i.e. the prediction of performance in university (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Lievens, Buyse, & Sackett, 2005a). Hence, SJTs can be an important addition to the selection battery. Patterson, Baron, Carr, Plint, and Lane (2009) studied the use of an SJT for selection into postgraduate general practitioners training in the UK. This SJT focused on three non-clinical selection criteria: empathy, integrity, and coping with pressure. The SJT was the best single predictor of performance in a selection center that used work-relevant simulations to target both clinical and non-clinical domains. Furthermore, the SJT offered the most incremental validity over other methodologies. These findings have important implications for the development of selection methodologies in the assessment of non-clinical domains.

Construct-related Validity. It is commonly accepted that an SJT is a method to evaluate a variety of professional knowledge, capacities, and competencies. SJT items may refer to a

wide range of situations and answering an SJT involves using experience, personality, and common sense. Low internal consistency coefficients point in the direction that SJTs measure different constructs. To determine the construct validity of an SJT, their correlation with other selection instruments has been investigated. In the meta-analysis of McDaniel et al. (2001), it was found that SJTs show a significant, moderate correlation (r=.46) with cognitive ability, even though there was substantial variability around this estimate. The meta-analysis of McDaniel et al. (2007) revealed that the type of response instruction seems to be a key factor, as it was found to affect the cognitive loading of SJTs. That is, SJTs with knowledge instructions had a higher cognitive loading. Alternatively, SJTs with behavioral tendency instructions had a higher personality loading. Taken together, the extent to which SJTs measure a specific construct, varies greatly. Hence, an SJT is best viewed as a multidimensional measurement method with which one can assess a variety of work related knowledge, skills and abilities (KSAs), rather than as a method with which one can measure a particular individual differences construct.

Adverse impact. Do SJTs disadvantage certain groups (race or gender)? Differences in mean scores between racial subgroups are typically smaller than those reported for cognitive ability tests. Whetzel, McDaniel, and Nguyen (2008) conducted a meta-analysis to examine the value of SJTs in reducing subgroup differences. With respect to race, differences in mean SJT scores between subgroups were typically smaller than those reported for various ability tests, including cognitive ability. The difference between Whites and minority members was without exception in favor of White participants who scored .38, .24 and, .29 SD higher than Black, Hispanic, and Asian participants, respectively. Past research has shown that females score slightly better than males on SJTs (O'Connell, McDaniel, Grubb, Hartmann, & Lawrence, 2002; Weekley & Jones, 1997, 1999). Whetzel et al.'s meta-analysis (2008) confirmed that women in general outperform men on SJTs, although the female advantage in

SJT performance was rather limited (*d*=.11). One explanation states that women tend to score higher on agreeableness and conscientiousness (McCrae & Terracciano, 2005). These two personality traits are commonly measured by SJT items.

Face validity. A great advantage of SJTs is the fact that applicants react very positive and perceive these tests as job-related and relevant. This adds to the applicant's judgment of the procedural justice of the selection process. Kanning, Grewe, Hollenberg, and Hadouch (2006) examined the factors of SJT presentation on test-taker perceptions. They concluded that SJTs that are interactive and used a video-based modality for the presentation of stimuli as well as for the response options received the highest ratings as compared to other SJTs that varied in other ways on these factors. Positive applicant reactions are important because they play a crucial role in motivation and performance in selection (Chan & Schmitt, 1997; Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001). Moreover, positive reactions give the employer a good image. This image has an important influence on the attraction of the applicant to the organization (Lievens & Highhouse, 2003). Furthermore, positive applicant reactions increase the chances of hiring the best applicants, avoid the possibility of costly litigation and contribute to the organization's reputation (Gilliland & Steiner, 1999; Ryan & Ployhart, 2000).

Coaching and practice effects. With a test gaining as much attention as the SJT, chances of a coaching business arising are big. Various test coaching programs and the Internet provide candidates with strategies to improve their test scores and get selected. For the organization or selection committee, especially the teaching of tricks and gimmicks has negative consequences: The actual test score does no longer provide an accurate picture of the true ability of the applicant. In the past, the effects of coaching were primarily studied in relation to cognitively-oriented tests in educational settings. In this context, research found that coaching produced small but practically meaningful increases in performance (Bangert-

Drowns, Kulik, & Kulik, 1983; Becker, 1990). So far, little research on coaching effects has been conducted in relation to SJTs. The results of Cullen, Sackett, and Lievens (2006) indicate that some SJTs are susceptible to coaching (*d*=.24). This indicates that caution must be taken when using SJTs in selection. A question similar to the coaching problem, deals with practice effects. Can applicants reach higher scores when they retest on an SJT? The results of Lievens, Buyse, and Sackett (2005b) show that retest effects for SJTs are not higher than retest effects for cognitively-oriented tests.

Conclusion

The large literature on SJTs provides many insights into the strengths and weaknesses of this selection instrument. However, the key limitations that were mentioned by McDaniel et al. (2001), namely predominantly low-stakes settings and concurrent or experimental designs apply to the entire research on SJTs. In such settings, respondents are mostly incumbents or test subjects who are not extremely motivated to take the test. In most studies, the SJT is not used to make actual selection decisions. Therefore, studies in operational high-stakes selection settings are needed to draw firm conclusions on the use of SJTs as additional selection instruments, over and above cognitively-oriented tests.

THE HIGH-STAKES SETTING: ADMISSION TO MEDICAL AND DENTAL STUDIES IN FLANDERS

History

In the fall of 1995, the Belgian federal government was faced with an excess number of doctors and dentists. In order to deal with this, a law was voted in which the maximum amount of graduating doctors and dentists per year was determined. These federal intentions were communicated to the two communities in Belgium: the Dutch speaking region (Flanders)

and the French speaking region (Wallonia). This law influenced the policy of these two regions. In Flanders, the ministry of Education and Training and the ministry of Health and Wellbeing -who were both responsible for the execution of this federal law- decided to install a Technical Commission. This commission had to determine how Flanders would restrict the flow of students in medical and dental education. It was decided to do this via an Admission Exam. This Admission Exam had to 1) discriminate students on their chances of succeeding medical and dental education and 2) give serious indications of their later performance as medical doctor or dentist. The Technical Commission thoroughly discussed many options but high priority was given to the specific content of the selection test and to the method of scoring and evaluating it. The first three years were deemed experimental so that continuous evaluation and corrections were possible. The commission recommended a standardized selection procedure for all candidates, organized at the same time, at the same place. It was decided that each year, two sessions would be held and all systems for data processing should be computerized. Therefore, the answers to the questions in the selection test had to be determined a priori. No further selection on the part of universities was allowed. Up until this day, the Admission Exam is the only criterion that decides whether a student can start medical or dental education. For candidates, the Admission Exam is a huge obstacle in attaining their goal.

The Technical Commission reached a consensus concerning the content of the Admission Exam. The first part of the Admission Exam captured the knowledge and insight of candidates in Sciences. The second part of the Admission Exam involves information gathering and processing abilities. The first Admission Exam took place in 1997.

Content and requirements to succeed

Before reaching a consensus, the Technical Commission discussed many possible subtests. Three suggested alternatives were not withheld, namely an interview, a manual dexterity test, and a nursing internship. This was mostly due to the expected size of the applicant group. There was no consensus in the Technical Commission regarding a personality test. Therefore, no personality inventory was administered. After the Technical Commission, an Exam Commission was set up to oversee the actual development and organization of the Admission Exam. It contained many members of the Technical Commission but was expanded with clinical and scientific subject matter experts. Over the years, the number and content of subtests of the Admission Exam has changed a few times. On these grounds, different periods can be distinguished.

Period 1: 1997

When the ministry of Education and Training announced the Admission Exam, many students claimed that they did not know about the Admission Exam when they chose their main courses in high school two years before and went to court. Due to the decision of this higher court of Justice in Belgium (Arbitragehof cfr. Constitutional Court), the first part of the Admission Exam (knowledge in sciences, further called KIW) was not administered in 1997. Students only had to take the second part of the exam: information gathering and processing (further called IVV). This part consisted of two main subparts which each contained four tests. The first subpart of IVV comprised four cognitive ability tests: reasoning, memory association, visual information processing, and pattern recognition. The second subpart of IVV consisted of 4 situational tests: a lecture on a medical subject, a silent reading text, an interaction with a patient, and a discussion in a multidisciplinary team. The development and content of these eight IVV tests is described in the following paragraphs.

Cognitive ability tests. These measures were not specifically developed for the Admission Exam. Instead, four existing cognitive ability tests were chosen. For test security reasons, the source of these measures and example items cannot be presented. The first cognitive ability test was a reasoning test, which consisted of 54 questions with five response options. The problems in this test were formulated in verbal, numeric, or figural terms. Prior research demonstrated the good reliability and predictive validity of this test for medical students. In particular, Minnaert (1996) reported an internal consistency coefficient of .84 and a validity coefficient of .36 for predicting first-year GPA in medical studies. Because of these good psychometric properties, the Admission Exam Commission decided to weigh this test more in the total Admission Exam score (see table 1 for specific weights). The visual information processing test measured the ability to quickly scan and interpret complex figures. It consisted of 32 items with five response alternatives. In the third test, memory association, 15 names of patients had to be memorized. Besides the names, their age, job title, personal characteristics, and diagnosis were also included. The reproduction phase contained 20 questions dealing with these patient descriptions. The pattern recognition test measured the cognitive ability to determine which simple figure was part of a more complex figure. This test contained 50 items and per item five possible simple figures were provided in a test booklet. According to prior research provided in the test booklets, the internal consistency of these three tests was satisfactory. For each test, specific time limits were set.

Situational tests. These four tests were specifically developed for the Admission Exam. The first two tests i.e. the videotaped lecture and the written text with a medical subject matter, were miniaturized samples of important student tasks. Real lessons and course texts were used. To this end, a professor delivering a lecture (30 minutes) was filmed and a seven-page text was extracted from a course syllabus. A list of relevant questions and response options were developed. The other two situational tests (i.e., interaction with a patient and medical

team discussion) were video-based SJTs. An approach similar to other studies (see e.g., Weekley & Jones, 1997) was used for developing these SJTs. In a first step, a representative group of critical incidents were gathered for these two situations. To this end, the relevant literature was inspected and experienced physicians and professors in general medicine were questioned so they could provide examples indicative of effective and ineffective job behavior in the respective situations. This exercise yielded a list of 376 usable examples of behavior. Second, scripts were written. Two professors teaching physicians' consulting practices tested the scripts for realism. The scripts depicted the word-to-word dialogue between the parties involved. Using a similar approach, questions and response options were derived. Third, semi-professional actors were selected to play the various roles while being videotaped. An experienced physician attended the set to guarantee realism. For each videotaped test, 30 multiple-choice questions were formulated. In the last step, expert judgments were used to develop the scoring key. Cohen's (1960) kappa, which is an indication for inter-rater agreement, was satisfactory (always exceeded .70). Discrepancies were easily resolved through discussion. All questions of the situational tests were of the multiple-choice type with four response options. Due to test security reasons, pilot testing of these items was not possible, nor was it allowed to discard items or use different scoring rules. Again, specific time limits were set for each test.

Admission Exam scores. For each of the eight tests 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 tests were computed. These weights were determined by the Admission Exam Commission and are presented in table 1. The maximum score on each part was 10. Candidates had to obtain at least 6 out of 10 on each part to pass the Admission Exam. The final Admission Exam score was obtained by summing

both weighted sum scores. Candidates who passed received a certificate which guaranteed entry to medical or dental education.

Table 1. Admission Exam for Medical and Dental Studies in Flanders 1997-1999

Knowledge in Sciences (KIW)			Information gathering and processing (IVV)		
Test	Number of items	Weight	Test	Number of items	Weight
			Cognitive ability tests		
Biology	15	.50	Reasoning	54	.50
Physics	15	.50	Visual information processing	32	.20
Chemistry	15	.50	Memory association	20	.10
Mathematics	15	.50	Pattern recognition	52	.20
			Situational tests		
			Videotaped lecture	40	.33
			Written Text	20	.17
			Videotaped Interaction between doctor and patient	30	.25
			Videotaped Team discussion	30	.25

Note. In 1997, KIW was not administered due to a decision of a higher Court of Law

Period 2: 1998 and 1999

In 1998 and 1999, still experimental years, KIW and IVV were both administered. KIW tested students' knowledge of four science tests: biology, physics, chemistry, and mathematics, who each had 15 items with four possible answers. The difficulty of these subtests was adapted to the average level of difficulty in the last years of Flemish high schools. Each year, a professor who was a subject matter expert in that particular science subject (and member of the Admission Exam Commission), developed the items and possible answers (for both sessions). The Commission discussed the difficulty of the items.

Candidates had three hours to solve the items. IVV contained the eight subtests described above.

After these three experimental years, the Admission Exam Commission wanted to evaluate the Exam and propose improvements for the future. That is why they invited two experts on the admission subject to inspect the Flemish Admission Exam and give their expert opinion. These recommendations and the results of the commission's own research activities, gave rise to a few adjustments in the conception of the Admission Exam.

Period 3: 2000-2002

Table 2 shows that a few of the IVV subtests were removed from the Admission Exam after the three experimental years. From 2000 on, the reasoning test was the only cognitive ability test in the Admission Exam. IVV further contained the written text about a medical subject matter (no longer from a course syllabus but developed from scratch) and the videotaped interaction between a doctor and a patient. The conditions to pass the exam were made less stringent. Candidates had to obtain at least 5 out of 10 for KIW and at least 5 out of 10 for IVV. In total, however, they still had to obtain 12 out of 20.

Table 2. Admission Exam for Medical and Dental Studies in Flanders since 2000

Knowledge in Sciences (KIW)			Information gathering and processing (IVV)		
Test	Number of items	Weight	Test	Number of items	Weight
			Cognitive ability tests		
Biology	10	.50	Reasoning	50	.70
Physics	10	.50	Situational tests		
Chemistry	10	.50	Written Text	30	.70
Mathematics	10	.50	Videotaped Interaction between doctor and patient	30	.60

Period 4: 2003-2005

Two major changes were introduced in the Admission Exam of 2003. First, passing conditions were eased once more. Students still had to obtain at least 5 out of 10 for both KIW and IVV but their total score to pass the Admission Exam was lowered to 11 out of 20. Second, due to many technical problems and high production and administration costs, it was decided that the videotaped interaction between the doctor and patient would be transformed into a paper-and-pencil test. Therefore, all dialogues were written into full text and the SJT was fully administered on paper (i.e., both stimulus and responses). To increase realism, photographs were added to the test booklet.

Period 5: 2006-present

Due to issues in developing both situational tests, especially guaranteeing the same difficulty index in both sessions per year, it was decided to change their format. First, the silent reading text, which was initially one long text with 30 questions, was changed to seven short texts (one page) with each 5 or 6 questions (30 in total). Consequently, candidates can choose which text they read first and which text they possibly ignore. Hence, difficulty no longer depends on one single text and a greater variety in medical subjects is possible. Second, the interaction between a doctor and patient no longer consisted of one single interaction with 17 critical incidents. Since this subtest consists of 30 questions, candidates are now confronted with 30 different, and independent situations. Situations can deal with interactions between doctor and patient, but also with interactions between nurse and patient, doctor and nurse, doctor, child and parent, dentist and patient and so on. Hence, the context in which doctor and patient interact, is broadened and other significant care takers are introduced.

Conclusion

Throughout the 13 years of its existence, various changes to the Admission Exam have been made. At first sight, these changes may seem substantial. However, a closer look reveals that the basic design, namely using cognitive and non-cognitive measures was never abandoned. The early alterations (in 1999) mainly resulted from comments made by the two admission system experts. These changes made the exam more practicable and efficient by reducing the number of tests and items. Later adjustments (presentation format of SJT and silent reading text) were made by the test developers in order to safeguard the validity of these tests. Hence, in the studies described in this dissertation, Admission Exams of different years, and with different contents are used. Differences between different cohorts and differences in the difficulty of the tests were resolved by standardizing the Admission Exam scores per year.

COMPARISONS TO INTERNATIONAL MEDICAL AND DENTAL ADMISSION

In the previous part of this chapter, we described the context of medical and dental admission in Flanders. However, the entry criteria, structure, teaching methodology, and curriculum offered at medical and dental schools vary considerably around the world. One aspect that medical and dental schools around the world have in common, is that they are often highly competitive and most of them use a form of selection to decrease the inflow of students. In the following, we describe general differences between admission to medicine and dentistry in Flanders, and admission systems used around the world.

First, in European countries (like Belgium), the study of medicine is mostly completed as an undergraduate degree. However, in many other countries, medical education is moving closer to the US/Canadian model. In these countries, medical degrees require at least several years of previous study at university. Therefore, students who want to enter medical school often have already completed a bachelor with a curriculum with a heavy

emphasis on sciences. In these cases, the mean age of students entering medical school is higher than in Flanders where most of the candidates are approximately 18 years old.

Second, although medical schools around the world confer medical degrees, in many countries a medical doctor or dentist may not legally practice medicine until (s)he is licensed by the local government. This may require passing an extra test (licensing examination) or paying a fee. In Flanders, up until now, every student who graduates after seven (medicine) or after five (dentistry) years, is allowed to practice the profession.

Next, in Flanders, the Admission Exam is centrally organized and administrated. There is no further selection on the part of the universities. Every student who passes the Admission Exam, can enter his/her preferred university. However, in many countries, medical and dental schools construct and apply their own entrance examinations and they decide on an independent basis who gets accepted.

Fourth, in Flanders students who want to study medicine and dentistry take the same Admission Exam. Only after passing the exam, students have to indicate which study they aspire by enrolling in the medical or dental school of their choice. Consequently, the Admission Exam Commission never knows in advance how many medical or dental students will start the education. There is no “numerus fixus” as every student that succeeds for the Admission Exam, is allowed to start medical or dental education. In Flanders, this regulation has led to a major lack of dentists since every year up to 90% of students passing the Admission Exam, choose to study medicine.

Another major difference between Flanders and other countries is the use of the SJT. If other countries measure interpersonal and communication skills, in most cases they do so by using an interview or by including a personality test. As far as we know, Belgium is the only country that uses an SJT for actual college admission decisions.

Finally, the international comparison learns that in many countries a lot of weight is given to applicants' past academic records. Former grades (high school or undergraduate) determine students' chances of acceptance for medical (or dental) education. In many cases, most attention is given to grades attained in a final secondary school leaving exam (e.g., Germany and Ireland). Some countries apply secondary grades and add specific requirements (e.g., the UK). Past academic grades have proven their predictive value in higher education as recent reviews have shown that the undergraduate grade point average (GPA) is moderately related to subsequent academic performance (McGaghie, 2002; Salvatori, 2001). Similarly, the Medical College Admission Test (MCAT) has an acceptable predictive value for pre-clinical performance (Julian, 2005). Note that the relation with professional performance is not straightforward. A very early study of Price, Taylor, Richards and Jacobsen (1964) clearly demonstrates that performance in formal education, as measured by grade point averages, comes out as a factor almost completely independent of all the factors having to do with performance as a physician. However, in the case of Flanders, it was the concern of the Belgian government that requesting students' grades could give the impression that their grades influenced their chances of succeeding in the Admission Exam. Therefore, the use of secondary school grades was not allowed. In view of the scientific evaluation, the grades were requested after the Admission Exam and students could cooperate on a voluntary basis.

CURRENT DEBATE IN MEDICAL AND DENTAL ADMISSION RESEARCH

Selection in higher education typically serves two purposes: (1) to reduce the large number of otherwise qualified and capable applicants to match the number of places available, and (2) to enroll students thought most likely to succeed in what is an arduous program of study and to subsequently become effective members of the profession. However, selecting those students who will do well academically in the early part of medical school, or selecting

those students who will make the best physicians after medical school do not necessarily have the same outcome. Barr (2010b) claims that success in pre-medical sciences gives rise to success in pre-clinical sciences encountered early in medical school, but success in pre-medical sciences has little predictive value regarding eventual success as a clinician.

The literature on medical and dental school admissions consistently draws a distinction between cognitive and non-cognitive abilities (Benbassat & Baumal, 2007). The former refer to intellectual prowess, typically measured by GPA or performance on standardized tests of knowledge (e.g., MCAT in US or GAMSAT in UK). The latter quality, non-cognitive aptitude, is typically used to encapsulate all the other qualities that might be desired in an applicant (Eva et al., 2009). The list of these non-cognitive abilities is very large. Although there is some variation in opinion with respect to the relative weight assigned to cognitive and non-cognitive measures of potential at medical school admissions, there has always been widespread agreement that it is desirable to broaden the scope of assessment beyond academic achievement.

Traditionally, non-cognitive traits are inferred from applicants' performance in interviews and on specific assignments such as small group discussions of a problem (Collins, White, Petrie, & Willoughby, 1995), multiple mini interviews (MMI) (Eva, Rosenfeld, Reiter, & Norman, 2004), simulated tutorials (Kulatunga-Moruzi & Norman, 2002), as well as from applicants' scores on personality tests and letters of recommendation. Ferguson, James, and Madeley (2002) summarized the consistent findings regarding the use of personality measures. They found that within medicine extraversion predicted success in paediatric objective examinations and conscientiousness was a positive predictor of preclinical achievement, even when controling for previous academic performance (β=.58). Lievens, Ones, and Dilchert (2009) found that over time extraversion, openness, and conscientiousness factor and facet scale scores showed increases in operational validity for predicting grade

point averages in medical education. Openness and extraversion gained importance for later academic performance. Conscientiousness appeared to be an increasing asset for medical students. Manuel, Borges and Gerzina (2005) tested whether personality factors were associated with the clinical skills of second-year medical students, and were able to confirm such an association. Students' communication skills correlated positively with warmth, emotional stability, and perfectionism.

The interview is even more widely used for capturing non-cognitive skills (Albanese, Snow, Skochelak, Huggett, & Farrell, 2003; Kreiter, Yin, Solow, & Brennan, 2004). Half a century ago, Gee and Cowles (1957) already state that as long as there is no objective way to determine the criteria for good physicians of whatever variety, and as long as there are no objective ways to evaluate some of the traits that are allegedly prerequisites for becoming a good physician, the interview is the only tool for estimating traits. Interviews can assess various personal attributes considered appropriate to a career in medicine. These attributes include the ability to communicate, cooperativeness, evidence of active participation, open-mindedness, self-confidence. However, results of the validity of interviews are not consistently positive. In general, limited predictive validity is found (Streyffeler, Altmaier, Kuperman, & Patrick, 2005). The meta-analysis of Goho and Blackman (2006) concluded that selection interviews have a modest capacity in predicting clinical performance in healthcare disciplines and they probably have limited practical value. Other systematic reviews also indicate that interviews add little to the selection process (Ferguson et al., 2002). The results of Wilkinson et al. (2008) were consistent with this prior research, concluding that prior academic performance accounted for 23% of variance in undergraduate medical performance and that interviews had only limited value.

More generally, the study of Wilkinson et al. (2008) has instigated the old debate as to whether it makes sense to include measures of non-cognitive skills in admission. That is, the

need to incorporate non-cognitive measures in the selection of medical students is no longer agreed on by all researchers in the field. For instance, Benbassat and Baumal (2007) posited that the use of the vast majority of non-cognitive admission criteria is not evidence-based and that these criteria should not be a component of the selection process for medical schools. Others stated that selecting for interpersonal relationship skills is only to be recommended when selecting general practitioners and psychiatrists (Arnold, 2008). In addition, over the last years, the idea that only using prior academic achievement and measures of cognitive knowledge does not exclude students with interpersonal skills, has gained ground. For example, some authors argued that it is important to acknowledge that academic ability and other key (non-cognitive) attributes are not necessarily inversely correlated (Norman, 2004), or mutually exclusive. Indeed, there is evidence that the two are positively correlated (Eva & Reiter, 2004). Thus, selecting solely or predominantly on academic performance may in fact also lead to the admission of students with attractive non-cognitive attributes (Wilkinson et al., 2008). Clever people are not known to be systematically less humane than others (Brown, 2008).

Following the paper of Wilkinson et al. (2008) and the resulting debate, many researchers expressed their concern about this evolution in medical selection. Harding and Wilson (2008) argued that abandoning interview selection methodology represents a regressive step in medical student selection. They mentioned that some interviews (like MMI) have demonstrated promising reliability and validity (Reiter, Eva, Rosenfeld, & Norman, 2007; Roberts et al., 2008). Therefore, Powis (2008) repeated that medical schools have to select students on the basis of more than their academic achievements at school. Measuring cognitive ability is a step in the right direction, but it does not tackle the admission of people from lower socioeconomic groups or those whose education has been compromised by attending poorer schools. Bore, Munro, and Powis (2009) developed a comprehensive model

for selecting medical students. This model is grounded in the theoretical and empirical selection and assessment literature. Their goal was to develop a model that results in ethically defensible selection decisions. The model includes a method of using scores from cognitive and non-cognitive measures. The Admission Exam for medical and dental studies in Flanders is an example of such a model as it captures the traditional cognitive factors and a method (SJT) to measure the non-cognitive (interpersonal) skills. In light of the ongoing debate about the relevance of assessing non-cognitive factors, this makes it worthwhile to scrutinize the performance of this SJT in the Admission Exam in Flanders.

RESEARCH OBJECTIVES

This chapter emphasizes that in the field of medical selection (and even in the broader field of selection in higher education) many disagreements and discussions exist. The present dissertation tries to answer some of the questions in this ongoing debate. Previous research documents that cognitive predictors are the most important predictors in selection for higher education. Conversely, interviews as measures of non-cognitive capacities have not shown consistent results. So, there is a clear need for other measures that enable to assess non-cognitive factors. The Admission Exam in Flanders is the only one worldwide that uses an SJT to measure interpersonal skills of applicants to medical studies. That is also the reason why this dissertation focuses on the SJT. This is reflected in the following four research questions.

First, we want to investigate whether it is possible to use the SJT (and the other tests of the Admission Exam) for selecting students in both medical and dental education. As stated before, one of the big differences between the Admission Exam in Flanders and the ones in the rest of the world is the use of the same Admission Exam for two different majors (medical and dental studies). Hence, Chapter 2 addresses the first research question by

examining ***whether the same admission exam tests can be used for different academic majors.*** This question has major practical implications. At present, there already exists a shortage of dentists. If students who aspire a career in dentistry have lower scores and therefore fewer chances to pass the exam, the Flemish health care faces a major challenge in the future.

Second, in the past, most studies regarding the criterion-related validity of SJTs were concurrent in design and did not involve the use of SJTs in operational high-stakes settings (Christian et al, 2010; McDaniel et al, 2001; McDaniel et al, 2007). Therefore, the present dissertation presents two studies of SJTs used in high-stakes contexts. These studies are described in chapters 3 and 4 and use predictive validation designs of an SJT used in an operational high-stakes setting. In particular, chapter 3 addresses ***research question 2***: ***What is the predictive validity of an SJT measuring interpersonal skills in dental education?***

As opposed to dental education, the predictive validity of the Admission Exam for the first years of medical education has been examined in the past (Lievens, Buyse, & Sackett, 2005a). However, the ultimate goal of medical selection is to choose those applicants who will do well as professionals. In the later years of medical education and in the profession, the impact and importance of interpersonal skills increases as interactions with patients augment. Therefore, chapter 4 focuses on the following research question (RQ 3): ***What is the long-term predictive validity of an SJT measuring interpersonal skills in medical education?***

This long-term predictive validity of SJTs might be a major concern for anyone interested in using this method in selection. In fact, the lab study of Cullen et al. (2006) indicates that some SJTs are susceptible to coaching. Thus, caution must be exerted when using SJTs on a long-term basis. In high-stakes selection, coaching effects might jeopardize the ultimate goal of the selection procedure. If applicants learn to use tricks and gimmicks to give the right answer and receive a higher score, this does not mean that they possess the

necessary interpersonal skills. Furthermore, paid coaching programs are not accessible to every student. So, students who seek coaching might differ from students who don't seek coaching activities. Research question 4 deals with these pre-existing group differences and coaching effects on SJTs and cognitive tests. The study in chapter 5 examines ***whether SJTs are susceptible to coaching effects (RQ 4).*** To account for pre-existing group differences between coached and non-coached groups, a methodological innovation is that we use propensity scoring.

REFERENCES

Albanese, M., Snow, M., Skochelak, S. Huggett, K., & Farrell, P. (2003). Assessing personal qualities in medical school admissions. *Academic Medicine, 78,* 313-321.

Arnold, P.C. (2008). Letters to the editor. *Medical Journal of Australia, 189,* 236.

Bangert-Drowns, R.L., Kulik, J.A., & Kulik, C.C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research, 53,* 571-585.

Barr, D.A. (2010a). Science as superstition: selecting medical students. *The Lancet, 376,* 678-679.

Barr, D.A. (2010b). Questioning the premedical paradigm. Enhancing Diversity in the medical profession a century after the Flexner report. John Hopkins University Press, Baltimore.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60,*373-417.

Benbassat, J., & Baumal, R. (2007). Uncertainties in the selection of applicants for medical school. *Advances in Health Sciences Education, 12,* 509-521.

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

Brown, C.A. (2008). Selecting medical students. *British Medical Journal, 336,* 786.

Bore, M., Munro, D., & Powis, D.A. (2009). A comprehensive model for the selection of medical students. *Medical Teacher, 31*, 1066-1072.

Cardall, A.J. (1942). *Preliminary manual for the Test of Practical Judgment*. Chicago: Science Research.

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

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15*, 233-254.

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

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

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Collins, J.P., White, G.R., Petrie, K.J., & Willoughby, E.W. (1995). A structured panel interview and group exercise in the selection of medical students. *Medical Education, 29*, 332-336.

Cullen, M.J., Sackett, P.R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14*, 142-155.

Eva, K.W., & Reiter, H.I. (2004). Where judgement fails: pitfalls in the selection process for medical personnel. *Advances in Health Sciences Education, 9*, 161-174.

Eva, K.W., Reiter, H.I., Trinh, K., Wasi, P., Rosenfeld, J., & Norman, G.R. (2009). Predictive validity of the multiple mini-interview for selecting medical trainees. *Medical Education, 43,* 767-775.

Eva, K.W., Rosenfeld, J., Reiter, H.I., & Norman, G.R. (2004). An admissions OSCE: The multiple mini interview. *Medical Education, 38,* 314-326.

Ferguson, E., James, D., & Madeley, L. (2002). Factors associates with success in medical school: systematic review of the literature. *British Medical Journal, 324,* 952-957.

File, Q.W. (1945). The measurement of supervisory quality in industry. *Journal of Applied Psychology, 29,* 381-387.

File, Q.W., & Remmers, H.H. (1948). *How Supervise? Manual 1971 revision*. New York: Psychological Corporation.

Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-359.

Gee, H.H., & Cowles, J.T. (1957). *Appraisal of Applicants to Medical School: Report of the Fourth Teaching Institute of the Association of American Medical Colleges*. Evanston, IL: Association of American Medical Colleges), 60.

Gilliland, S.W., & Steiner, D.W. (1999). Applicant reactions. In R.W. Eder & M.M. Harris (Eds.), *The employment interview handbook* (pp. 69-82). Thousand Oaks, CA: Sage.

Goho, J., & Blackman, A. (2006). The effectiveness of academic admission interviews: an exploratory meta-analysis. *Medical Teacher, 28,* 335-340.

Greenberg, S.H. (1963). *Supervisory Judgment Test manual*. Washington, D.C.: U.S. Civil Service Commission.

Harding, D.W., & Wilson, I.G. (2008). Medical school selection criteria and the prediction of academic performance. *Medical Journal of Australia,189* ,234-235.

Hogan, J.B. (1994). Empirical keying of background data measures. In G.S. Stokes & M.D. Mumford (Eds.), *Biodata handbook: Theory, research, and use of biographical*

information in selection and performance prediction, 69-107. Palo Alto, CA: CPP Books.

Julian, E.R. (2005). Validity of the Medical College Admission Test for predicting medical school performance. *Academic Medicine, 80*, 910-917.

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

Kreiter, C.D., Yin, P., Solow, C., & Brennan, R.L. (2004). Investigating the reliability of the medical school admission interview. *Advances in Health Sciences Education, 9*, 147-159.

Kulatunga-Moruzi, C., & Norman, G.R. (2002). Validity of admissions measures in predicting performance outcomes: the contribution of cognitive and non-cognitive dimensions. *Teaching and Learning in Medicine, 14*, 34-42.

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

Lievens, F., Buyse, T., & Sackett, P.R. (2005b). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58*, 981-1007.

Lievens, F., & Highhouse, S. (2003). The relation of instrumental and symbolic attributes to a company's attractiveness as an employer. *Personnel Psychology, 56*, 75-102.

Lievens, F., Ones, D.S., & Dilchert, S. (2009). Personality scale validities increase throughout medical school. *Journal of Applied Psychology, 94*, 1514-1535.

Makoul, G., & Curry, R.H. (2007). The value of assessing and addressing communication skills. *The Journal of the American Medical Association, 298*, 1057-1059.

Manuel, R.S., Borges, N.J., & Gerzina, H.A. (2005). Personality and clinical skills: Any correlation? *Academic Medicine, 80,* S30-S33.

McCrae, R.R., Terraciano, A., & 79 members of the Personality Profiles of Cultures Project (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology, 88,* 547-561.

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

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

McDaniel, M.A., & Nguyen, N.T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9* (1-2), 103-113.

McGaghie, W.C. (2002). Student selection. In G.R. Norman, C.P.M. van der Vleuten & D.I. Newble (Eds.), *International Handbook of Research in Medical Education,* 303-335. Dordrecht, Boston, London: Kluwer Academic Publishers.

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.

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

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

Motowidlo, S.J., Hanson, M.A., & Crafts, J.L. (1997). Low-fidelity simulations. In D.L. Whetzel & G.R. Wheaton (Eds.), Applied measurement methods in industrial psychology (pp. 241-260). Palo Alto, CA: Davies-Black Publishing.

Norman, G. (2004). The morality of medical school admissions. *Advances in Health Sciences Education, 9,* 79-82.

O'Connell, M.S., McDaniel, M.A., Grubb, W.L.III., Hartman, N.S., & Lawrence, A. (2002, April). *Incremental validity of situational judgment tests for task and contextual performance.* Paper presented at the 17th annual conference of the Society of Industrial Organizational Psychology, Toronto, Canada.

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

Patterson, F., Baron, H., Carr, V., Plint, S., & Lane, P. (2009). Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. *Medical Education, 43,* 50-57.

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

Ployhart, R.E., Porr, W., & Ryan, A.M. (2004, April). *New developments in SJT's: Scoring, coaching and incremental validity.* Paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Powis, D. A. (2008). Selecting medical students [editorial]. *Medical Journal of Australia, 188,* 323-324.

Powis, D. (2010). Improving the selection of medical students. *British Medical Journal, 340,* 708.

Price, P.B., Taylor, C.W., Richards, J.M., & Jacobsen, T.L. (1964). Measurement of physician performance. *Journal of Medical Education, 39,* 203-211.

Reiter, H.I., Eva, K.W., Rosenfeld, J., Norman, G.R. (2007). Multiple mini-interviews predict clerkship and licensing examination performance. *Medical Education, 41,* 378-384.

Roberts, C., Walton, M., Rothnie, I., Crossley, J., Lyon, P., Kumar, K., & Tiller, D. (2008). Factors affecting the utility of the multiple mini-interview in selecting candidates for graduate-entry medical school. *Medical Education, 42,* 396-404.

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

Salvatori, P. (2001). Reliability and validity of admissions tools used to select students for the health professions. Advances in Health Sciences Education, 6, 159-175.

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

Streyffeler, L., Altmaier, E. Kuperman, S., & Patrick, L. (2005). Development of a medical school admissions interview phase 2: predictive validity of cognitive and non-cognitive attributes. *Medical Education Online, 10,* 14.

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

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

Weekley, J.A., & Ployhart, R.E. (2006). *Situational Judgment Tests: Theory, measurement and application*. San Francisco, Jossey Bass.

Weekley, J.A., Ployhart, R.E., & Holtz, B.C. (2006). *On the development of situational judgment tests: Issues in item development, scaling, and scoring.* Mahwah, NJ: Lawrence Erlbaum.

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

Wilkinson, D., Zhang, J., Byrne, G.J., Luke, H., Ozolins, I.Z., Parker, M.H., & Peterson, R.F. (2008). Medical school selection criteria and the prediction of academic performance. *Medical Journal of Australia, 188,* 349-354.

Whetzel, D.L., McDaniel, M.A., & Nguyen, N.T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance, 21*, 291-309.

CHAPTER 2

Admission Systems to Dental School in Europe: A closer look at Flanders[1]

Dental education in Europe faces enormous challenges. One deals with the admission to dental school. Although admission procedures vary considerably across Europe, a characteristic of some systems is that the same procedure is used across students who will ultimately pursue different majors (medical or dental). This is based on the assumptions that there is no significant difference in these students' scores and that the requirements for medicine and dentistry are equal. This study examines these assumptions in the admission exam "Medical and Dental Studies" in Flanders. Students who pass may choose whether they start medical or dental education. Over an 8-year period (2000-2007), admission exam scores of students starting medicine (n=4492) were compared to those of students starting dentistry (n=547). Second, the validity of this exam is examined for both medical and dental education. It was found that students starting dentistry had a significantly lower total score on the admission exam than students starting medicine. Differences were especially striking for the cognitive part of the admission exam. For both medical and dental students the admission exam score was a valid predictor of academic grades in the first three years, although correlations were lower for dental education. These results have implications for admission procedures in countries where the same system is used for both majors. The findings that students who have a lower score choose dental education and that the validity of the exam is slightly lower for dentistry, raise questions about using the same admission exam for two obviously different majors.

[1] Buyse, T., Lievens, F., & Martens, L. (2010). Admission Systems to Dental School in Europe: A closer look at Flanders. *European Journal of Dental Education, 14*, 215-220.

INTRODUCTION

In 1999, the Bologna Declaration aimed to make the EU higher education community more transparent to place the EU as a world leader in higher education and to compete with the global market for students (1,2). In the domain of dental education, the aim was to harmonise the activity of the dental schools in achieving the EU standard for a graduate to be registered within the European Union as a dentist. With dental education moving toward a more European and even global context, it is time to examine the challenges that will test undergraduate education for dentists of the future (3).

So far, the discussion on dental education in Europe has mainly focused on the objectives of dental education and on the ways information and new skills should be provided to students (4-6). A common vision is that those selected as the dentists of the future should be capable learners, fascinated by knowledge and research, open-minded, communicative and socially competent, and open to the promotion of health and to all preventive and curative aspects of their chosen profession (7). Clearly, such dental curriculum objectives provide a firm basis for designing dental education. Similarly, these objectives play a key role to conceptualise admission procedures that can reach these objectives because the initial quality of students who choose dental education also influences the results of the educational efforts undertaken.

Due to historic, economic and cultural reasons the requirements for admission to dental education and the specific admission procedures used vary widely between the countries of Europe (7,8). Some countries allow everyone to start in the first year (e.g., France). Selection into the second year of dental (and in the latter country medical) school is then made on the basis of the results of competitive end-of-year examinations. Most countries, however, operate a numerus clausus which is set by the national government. In one system, countries (e.g.,

Germany, Ireland, and Norway) determine specific minimum academic entrance requirements in terms of high school grades. In Ireland, for example, entry into university education (including dental school) is based solely on academic performance in the Leaving Certificate Examination at the end of formal school education (9). In Germany, main attention is being paid to the grade of the final school leaving exam (called Abitur). In Norway, the criteria for admission to the dental faculty are outstanding school records (especially on mathematics, physics, and chemistry) (10).

Another system (e.g., the UK, Sweden, and Portugal) combines high school grades with national/local tests to select dental students. Most of the UK universities base the selection of dental students on prior academic performance as well as on the performance on the UKCAT (UK Clinical Aptitude Test) or GAMSAT (Graduate Medical School Admissions Test), with some universities even using extra procedures such as a structured interview (11). In Sweden, the national admission centre uses secondary school matriculation scores or scores from a university standard aptitude test (12). Some dental schools use admission tests and interviews in combination with either grades or USAT (university standard aptitude test) and one dental school also relies on the assessment of manual dexterity (13). In those cases in Sweden where both test/interview and grades are used, the outline is different between admission to medicine and to dentistry. In Portugal, students have to obtain excellent scores in the entrance exam and brilliant secondary school course grades. In the Netherlands, grades in high school play a key role because popular subjects such as medicine or dental medicine have a numerus fixus. Medical schools select a proportion of entrants via interview and other methods, but the remaining candidates are identified through a lottery (weighted by academic attainment) among school leavers (14).

In yet another system, countries like Finland and Flanders (the Dutch speaking part of Belgium) pay little attention to high school grades but choose their university students on the basis of an entrance exam. For example, in Finland, despite a nationwide final exam in high school (matriculation examination), the majority of student selections for university is based on entrance exams. As every university has internal autonomy, the entrance procedures vary widely but nearly all universities use a quota. Contrary to this country, one common government-run admission exam is organised in Flanders for students who want to study medicine or dentistry. The cut-off for allowing students into both studies is also identical. There is no numerus clausus. Everyone who succeeds (i.e. reaches the cut-off score) can enrol in their university of choice and can choose whether to study either medicine or dentistry. There is no specific number of places in each school and students claim their choice for medicine or dentistry only after passing the exam. Ever since the admission exam was institutionalised, most passing students chose medicine (in some years up to 90%). Previous studies showed this Flemish admission exam to be valid for predicting future grades (15-18).

A characteristic of the Flemish admission exam is that the same admission exam procedure (e.g. same tests, same cut-off score) is used across students who will ultimately pursue different majors (either medical or dental). Use of the same admission exam procedure across different majors is based on two assumptions. First, it assumes there is no significant difference in students' scores on the admission exam. If one of the groups (either future medical students or future dental students) obtains lower scores, then less of them might pass the admission exam**.** In the end, this also affects the number of medical students or dental students who start education, ultimately graduate, and go on to the profession. Second, use of the same admission exam procedure across different majors (either medical or dental) is also based on the assumption that

the same requirements are needed for medical and dental education, which is questionable (see the aforementioned specific objectives of dental education in Europe).

The objective of this study is twofold. First, the admission exam scores of students who chose medical education are compared to those of students who chose dental education after passing the same admission exam in Flanders. We compare the scores of these two groups of students on 8 admission exams (from 2000 to 2007). A comparison is made in terms of (i) the total admission exam score, (ii) the cognitive part of the admission exam, and (iii) the non-cognitive part of the admission exam. Second, the validity of the Flemish admission exam is examined for both medical and dental students. This allows determining whether the admission exam score correlates equally well with academic grades in medical versus dental school.

METHODS

Demographic profile

Data were collected from students who passed the admission exam from 2000 to 2007 and subsequently started medical or dental studies in one of the six Flemish medical faculties (of which only two provide dental training). The total sample size was 5039. Mean age of the total group on the date of their participation in the admission exam was 18 years and 3 months. For the students who chose medical education (n=4492) the mean age was 18 years and 3 months (median=18y1m), whereas for the students choosing dental education (n=547) it was 18 years and 6 months (median=18y2m). The gender ratio amongst the participants was approximately 60% female. The percentages of males and females were equally distributed each year. The details per year are presented in Table 1.

Table 1. Sample and demographic characteristics per year

		Students choosing Medical Education				Students choosing Dental Education			
Year	Total n	n	Male (%)	Female (%)	Mean age	n	Male (%)	Female (%)	Mean age
2000	399	367	144 (39.24)	223 (60.76)	18y2m	32	13 (40.63)	19 (59.38)	18y4m
2001	416	361	133 (36.84)	228 (63.16)	18y4m	55	14 (25.45)	41 (74.55)	18y5m
2002	492	435	159 (36.55)	276 (63.45)	18y1m	57	21 (36.84)	36 (63.16)	18y5m
2003	669	601	225 (37.44)	376 (62.56)	18y4m	68	23 (33.82)	45 (66.18)	18y10m
2004	689	621	220 (35.43)	401 (64.57)	18y2m	68	18 (26.47)	50 (73.53)	18y5m
2005	832	731	270 (36.94)	461 (63.06)	18y3m	101	40 (39.60)	61 (60.40)	18y8m
2006	829	725	285 (39.31)	440 (60.69)	18y3m	104	37 (35.58)	67 (64.42)	18y3m
2007	713	651	277 (42.55)	374 (57.45)	18y5m	62	25 (40.32)	37 (59.68)	19y5m
Total	5039	4492	1713 (38.13)	2779 (61.87)	18y3m	547	191 (34.92)	356 (65.08)	18y6m

Instrument

The first part of the admission exam was designed to evaluate applicants' mastery of 4 basic science-related subjects (mathematics, physics, chemistry, and biology). Per subject, 10 multiple choice questions were asked. Every question had 4 possible answers of which only one was correct.

Next, the cognitive ability test was a reasoning test which consisted of 50 multiple choice items with 5 response alternatives per item. The problems in this test were formulated in either verbal, numerical or figural terms. Prior research demonstrated the good reliability and predictive validity of this reasoning test for medical and dental students (19,20). In particular, Minnaert (19) 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.

The remaining two tests of the admission exam were a silent reading protocol and a situational judgement test (SJT) about a physician-patient interaction. The silent reading protocol

consisted of one or more texts followed by a total of 30 multiple choice questions. The physician-patient interaction was a SJT. SJTs are measurement methods that present applicants with job-related situations and possible responses to these situations (21,22). All 30 questions of the SJT were of the multiple choice type, with four response alternatives. No medical background was needed for this SJT. For all tests of the admission exam, specific time limits were set. More information about these tests can be found in Lievens et al. (17,18).

To obtain a total admission exam score, a weighted sum of the aforementioned test scores was computed. These weights were determined by the commission overseeing the admission exam. Candidates who passed the exam (about 30%) received a certificate that guaranteed entry to either medical or dental studies in any university of the Flemish community.

Regarding the criterion measure, we retrieved students' grade point average (GPA) from the first three years of medical and dental school from archival records of all universities in Flanders. The courses in these first three years primarily deal with medical subjects but some deal with communicating with patients, internships etc. (in some universities up to 15% of courses involves dealing with patients). We gathered students' GPAs at the end of each year. Given differences across universities (different courses, teachers,…), we standardised students' GPA within university and within academic year (ie. computed z-scores). In Belgium GPA is measured on a scale from 0 to 20, with higher scores indicating better grades.

Analysis

Data were analysed using the Statistical Package for Social Sciences (SPSS) version 15.0. To examine the first objective and to compare both groups (medical and dental students), t-tests for independent groups were conducted and both significance tests and effect sizes (d) were

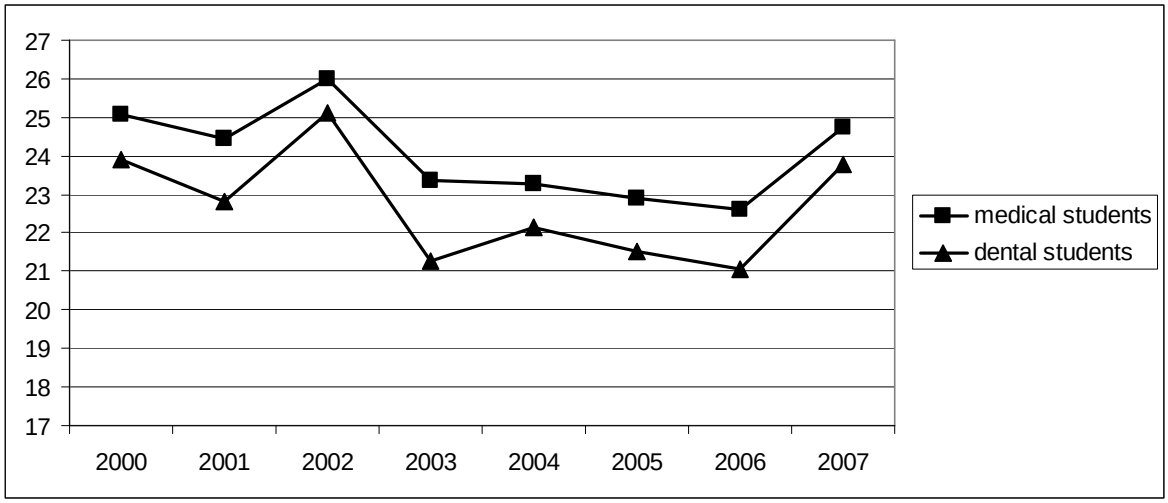
presented. The level of significance was set at $p < .05$. The effect size was defined as the difference between two means divided by the pooled standard deviation for those means. Cohen's (23) rules of thumb were used which define $d = .20$ as a small effect, $d = .50$ as a medium effect and $d = .80$ as a large effect.

To examine the second objective (validity of the admission exam score), Pearson correlations were computed between the final admission exam score (see above) and GPA (see above) in the first three years of students who passed the admission exams between 2000 and 2007. These correlations were computed separately for medical versus dental school students. As these analyses were conducted only among people who passed the exam and subsequently started in medical/dental school, these analyses are based on a smaller number of students than the mean comparisons. For instance, first year GPA of students attending the admission exam in 2007 was not yet available at the time this study was conducted.

RESULTS

Table 2 presents the descriptive statistics on the various tests and total score of the admission exam, broken down per year by chosen education. Regarding the total score (see Figure 1) on the admission exam, students who subsequently chose medicine obtained a higher score than students who chose dental education. This difference was significant in every year under study. Effect sizes of these significant differences varied from .26 to .54, showing small to medium effects. Note that in some years, the differences between both groups are quite small.

Figure 1. Total scores on Admission Exam for medical and dental students per year



A comparable consistent pattern was observed concerning the cognitive parts of the admission exam. In all years, future dental students obtained a lower score than medical students for the science knowledge tests. In 5 out of these 8 years the difference with future medical students was significant ($p < .05$, d varying from .31 to .41). In all years, future dental students had a lower score than medical students on the cognitive ability test and in 1 out of 8 years the difference was significant ($p = .043$, $d = .21$). A comparable result was found for the silent reading protocol test where future dental students always scored lower and in 5 out of 8 years this lower score was significantly different ($p < .05$, d varying from .25 to .59).

For the doctor-patient interaction results were not consistent. In 1 out of 8 years, medical students scored significantly higher than dental students ($p = .007$, $d = .28$). In 3 out of 8 years however, dental students obtained a higher score than medical students (2002, 2004 and 2007) but these differences were not statistically significant.

Table 2. Means and standard deviations of Admission Exam test scores for medical and dental students per year

		Medical students			Dental students					
		n	Mean	SD	n	Mean	SD	t	p	D
	Cognitive part									
	Science	367	13.38	2.57	32	12.52	2.20	1.82	.070	.34
	Cognitive ability test	367	32.43	4.08	32	31.28	5.40	1.18	.247	.27
2000	Silent reading protocol	367	13.93	3.63	32	13.80	3.01	0.19	.848	.04
	Non cognitive part									
	SJT	367	16.93	2.74	32	16.10	2.45	1.66	.098	.30
	Total score	367	25.06	3.10	32	23.91	2.27	2.04	.042	.38
	Cognitive part									
	Science	361	13.15	2.22	55	12.65	2.39	1.53	.128	.22
	Cognitive ability test	361	27.33	4.68	55	26.18	5.07	1.67	.095	.24
2001	Silent reading protocol	361	11.39	3.80	55	9.18	3.32	4.10	.000	.58
	Non cognitive part									
	SJT	361	17.34	2.87	55	16.91	3.45	1.01	.316	.15
	Total score	361	24.45	3.65	55	22.83	3.85	3.04	.003	.44
	Cognitive part									
	Science	435	13.61	2.13	57	12.90	1.77	2.76	.007	.34
	Cognitive ability test	435	30.58	4.95	57	30.49	5.03	0.13	.898	.02
2002	Silent reading protocol	435	19.11	3.93	57	18.15	3.60	1.76	.079	.25
	Non cognitive part									
	SJT	435	18.11	2.93	57	18.55	2.36	-1.29	.199	-.15
	Total score	435	25.98	2.72	57	25.12	2.30	2.27	.023	.32
	Cognitive part									
	Science	601	11.02	3.01	68	9.79	2.94	3.22	.001	.41
	Cognitive ability test	601	27.72	5.39	68	26.73	5.32	1.44	.151	.18
2003	Silent reading protocol	601	19.80	4.33	68	17.19	4.55	4.69	.000	.59
	Non cognitive part									
	SJT	601	19.07	2.82	68	18.60	2.76	1.31	.190	.17
	Total score	601	23.34	3.82	68	21.26	3.89	4.25	.000	.54
	Cognitive part									
	Science	621	11.85	2.86	68	10.96	2.97	2.42	.016	.31

	Cognitive ability test	621	28.21	5.42	68	27.42	5.88	1.13	.260	.14
2004	Silent reading protocol	621	16.53	3.77	68	15.95	3.73	1.21	.229	.15
	Non cognitive part									
	SJT	621	18.14	3.26	68	18.18	3.67	-0.10	.918	-.01
	Total score	621	23.28	3.64	68	22.16	3.59	2.43	.015	.31
	Cognitive part									
	Science	731	10.94	2.68	101	9.99	2.54	3.35	.001	.35
	Cognitive ability test	731	27.92	5.33	101	26.79	4.69	2.03	.043	.21
2005	Silent reading protocol	731	19.77	4.01	101	18.79	3.80	2.30	.022	.25
	Non cognitive part									
	SJT	731	17.30	2.83	101	16.89	2.67	1.36	.175	.15
	Total score	731	22.92	3.35	101	21.51	3.15	4.00	.000	.42
	Cognitive part									
	Science	725	11.62	2.37	104	10.79	2.34	3.34	.001	.35
	Cognitive ability test	725	28.27	6.13	104	27.16	6.16	1.73	.085	.18
2006	Silent reading protocol	725	16.61	3.23	104	15.19	3.25	4.19	.000	.43
	Non cognitive part									
	SJT	725	15.64	3.88	104	14.53	4.15	2.70	.007	.28
	Total score	725	22.60	3.19	104	21.06	3.14	4.62	.000	.48
	Cognitive part									
	Science	651	12.70	2.92	62	12.08	2.52	1.63	.103	.21
	Cognitive ability test	651	29.93	5.96	62	29.41	7.14	0.55	.582	.09
2007	Silent reading protocol	651	11.04	3.60	62	9.94	3.38	2.32	.021	.31
	Non cognitive part									
	SJT	651	12.44	4.03	62	12.46	4.11	-0.05	.962	.00
	Total score	651	24.76	3.82	62	23.76	3.11	2.35	.021	.26

Note: Positive effect sizes (d) reflect differences that favor medical students whereas negative effect sizes (d) reflect differences that favor dental students.

Table 3 presents the results of the validity of the total admission exam score broken down for medical and dental students. For both medical and dental students, the total admission exam score was a valid predictor of academic grades in the first three years as all correlations were

significant. However, the total admission exam score was always a better predictor of academic grades in medical school than in dental school. For instance, the total admission exam score correlated .30 with academic performance of medical students in the first year, whereas it correlated .21 with academic performance of dental students in the first year.

Table 3. Validity of the Total Admission Exam Score (2000-2007) in Predicting GPA in the First Three Academic Years Broken Down by Medical and Dental Education

	Medical education		Dental education	
	n	r	n	r
Year 1	3859	.30**	400	.21*
Year 2	2102	.23**	191	.14*
Year 3	1605	.24*	134	.20

Note. * p<.05; ** p<.01

DISCUSSION

Dental education in Europe faces enormous challenges. The skill set which used to be accepted on graduation from dental graduates will need to be broader and higher (3). Dental education must adapt to these rapidly increasing demands. The admission process is also a part of this challenge. The nature of the admission process depends not only on the number of candidates and the capacity of the educational facilities but also on the views of the school administration and the wider academic community, as well as national policy on the openness of higher education. There is a clear need for research to improve the reliability and predictive power of currently used admission methods (7). The admission procedure of a particular country determines the quality of the students selected. In addition, the consequences of actions taken in educational settings and the efficiency of these actions depend to a great extent on the admission

system used.

As noted above, admission systems to dental education vary widely across Europe. This study speaks to admission systems wherein the same method (same tests, same cut-off score) is used across students who will ultimately pursue different majors (either medical or dental). Such systems are based on the assumptions that there is no significant difference between the capacities of students choosing for either of the two majors and that the requirements for both majors are the same. This study examines these two assumptions in the case of the Flemish admission exam. The present study is unique as it uses data from a multiple year period. As the authors were unable to identify prior studies that addressed the difference between admission exam scores and validities for future medical and dental students, future studies are needed to examine these issues in other systems and other countries in Europe.

Overall, our results are both striking and robust. Across all years, dental students systematically scored lower on the cognitive tests of the admission exam. For the non-cognitive test, there is no consistent pattern, although it should be mentioned that future dental students sometimes outperformed future medical students (albeit not significantly). As the 'weakest' students with respect to the cognitive skills were those who made the choice for dental studies, one can question whether the same success criteria should apply to them. Results further showed that the final admission exam score was a valid predictor of academic grades in the first three years of medical and dental education. However, the final admission exam score was always a better predictor of academic grades in medical school than in dental school, indicating that the two majors are not comparable. These somewhat lower correlations for dental curriculum could be explained by the fact that dentistry requires specific practical skills which are not assessed by the current admission exam.

These results deserve attention in light of the fact that in Flanders, the profile of the dental curriculum seems unattractive among the general public. Therefore, fewer students probably take the admission exam with the intention to start dental education (as compared to those who want to pursue medical education). In addition, this study shows that this particular group has less chances of passing the admission exam, leading to a small group who can actually start dental education in Flanders. Taken together, this means that the admission exam does not recruit enough students to answer population oral health needs in the future. In fact, since the exam takes place, the total intake number of dental students in Flanders never reached the quorum which is allowed at the end of the studies. Moreover, 50% of all Flemish dentists is nearly +50 yrs of age. So a shortage of practitioners is expected by the year 2015. Therefore, attempts to make dental studies and the dental profession more attractive in the eye of the public should be undertaken to increase the number of students in this field.

Several limitations of this study should be acknowledged and therefore, some caution in the interpretation of the results is warranted. First, in the present study the preferred career choice of the students was not measured before they took the admission exam. The present data relate to those students who passed the entrance examination; unknown are the passing rates among those who had a medical/dental curriculum in mind before participating. Such information became available only in 2008. Results (unpublished data) showed different passing rates for students who aspire to medical studies (20.7%) as compared to students who want to study dentistry (11.8%). The difference in total admission exam score was again significant (M=17.95, SD= 4.97) for students who want to pursue medical education vs. M=16.6, SD= 4.88 for students who want to pursue dental education (t=4.81, p=.000). These data corroborate our main conclusions.

As the admission exam is only developed for the Flemish part of Belgium, restrictions to the generalisability of the results must be acknowledged. The perception and prestige of a certain profession may vary from country to country. It would be worthwhile to determine if the results could be generalised to other countries (e.g., by comparing grades in high school, high school leaving exam scores or matriculation scores of both medical and dental students).

CONCLUSIONS

This study took a closer look at admission to dental education in Flanders. Students who passed the Flemish admission examination for medicine and dentistry and started the dental curriculum scored significantly lower with respect to sciences and cognitive ability compared to those who started medicine. The key findings that students who have an average lower score choose to enter dental school in Flanders and that the validity of the exam is lower for dental education raise questions about using the same admission exam for two obviously different majors.

ACKNOWLEDGEMENTS

This research was funded by the Flemish Ministry of Education and Training. The authors would like to thank Professor emeritus Daniel van Steenberghe, president of the entrance examination, for his insightful comments on an earlier version of this manuscript.

REFERENCES

1 Hobson RS. A view of European challenges in dental education. *Br Dent J* 2009; 206**:** 65-66.

2 European Commission. The Bologna Process – towards the European higher education area. The Bologna Declaration. European Commission, 1999.

3 Hobson RS. Challenges to future dental education. Br Dent J, in press.

4 Banoczy J. Harmonisation of dental education and curricula in Europe. Internat Dent J 1999: 49: 69-72.

5 Shanley DB. Convergence towards higher standards in international dental education. N Y State Dent J 2004: 70: 35-39.

6 Sanz M, Widström E, Eaton KA. Is there a need for a common framework of dental specialties in Europe? Eur J Dent Educ 2008: 12: 138-143.

7 Gaengler P, de Vries J, Akota I, Balciuniene I, Berthold P, Gajewska M, Johnsen D, Urtâne I, Walsh L, Zijlstra, A. Student selection and the influence of their clinical and academic environment on learning. Eur J Dent Educ 2002: 6: 8-26.

8 Scott J. Dental education in Europe: the challenges of variety J Dent Educ 2003: 67: 69-78.

9 Lynch CD, McConnell RJ, Hannigan A. Dental school admissions in Ireland: can current selection criteria predict success? Eur J Dent Educ 2006, 10, 73-79.

10 Komabayashi T, Åstrom A. Dental education in Norway. Eur J Dent Educ 2007; 11: 245-250.

11 Hoad-Reddick G, MacFarlane TV. An analysis of an admissions system: can performance in the first year of the dental course be predicted? Br Dent J 1999; 186: 138-142.

12 Röding K. Professional competence in final-year dental undergraduates: assessement of students admitted by individualized selection and through traditional modes. Eur J Dent Educ 2001: 5: 12-16.

13 Heintze U, Radeborg, K, Bergtsson H, Stenlåås A. Assessement and evaluation of individual prerequisites for dental education. Eur J Dent Educ 2004: 8: 152-160.

14 Coebergh J. Dutch medical schools abandon selection for lottery systems for places. *StudentBMJ* 2003;11:138.

15 Lievens F, Coetsier P. Situational tests in student selection: An examination of predictive validity, adverse impact and construct validity. International journal of selection and assessment 2002; 10: 245-257.

16 Lievens F. Longitudinal study of the validity of different cognitive ability tests in a student admission context. Applied HRM research 2004; 9(1): 27-30.

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

18 Lievens F, Buyse T, Sackett P. Retest effects in operational selection settings: development and test of a framework. Personnel Psychology 2005; 58: 981-1007.

19 Minnaert A. 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 1996.

20 Minnaert A, Janssen PJ. The additive effect of regulatory activities on top of intelligence in relation to academic performance in higher education. Learning and Instruction 1998; 9: 77-91.

21 McDaniel MA, Hartman NS, Whetzel DL, Grubb WL. Situational judgment tests, response instructions, and validity: A meta-analysis. Personnel Psychology 2007; 60: 63-91.

22 Lievens F, Peeters H, Schollaert E. Situational Judgement tests: a review of recent research. Personnel Review 2008; 37:426-441.

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

CHAPTER 3

THE VALIDITY OF SITUATIONAL JUDGMENT TESTS IN DENTAL STUDENT SELECTION[1]

Usually cognitive tests are used to select students into dental education. Yet, cognitive predictors explain only part of the variance in academic performance. Therefore, interviews and personality tests are often used to measure non-cognitive characteristics. Recently, situational judgment tests (SJTs) have drawn the attention. There is evidence that SJTs can be valid predictors in medical admission contexts. This study examines the validity of an SJT measuring interpersonal skills for predicting academic performance of dental students. Incremental validity over cognitive tests is also examined.

This study included 796 dental students who passed the admission exam for medical and dental studies in Flanders and enrolled in the two Flemish dental schools. Academic performance (GPA) in the five years of dental studies served as criterion.

Corrected correlation between the cognitive tests of the admission exam and GPA equaled .38. Their validity dropped from .45 (year 1) to .18 (year 5). However, the validity of the SJT increased from .05 (year 1) to .20 (year 5). The SJT had incremental validity in year 5.

Dental admission committees who envision assessing a broad set of capabilities, might consider using an SJT as a valuable supplement to cognitive tests. Future research needs to confirm our findings with job performance as criterion.

[1] Buyse, T., & Lievens, F. (accepted). The Validity of Situational Judgment Tests in Dental Student Selection. *Journal of Dental Education.*

INTRODUCTION

Admission committees responsible for selecting candidates for higher education programs face an important and challenging task. Especially for health professions programs such as medicine and dentistry where the admission process is typically very competitive it is incumbent upon the committee to select candidates from the total applicant pool who are most likely to succeed as students in the education program not only in the first years but also in the last years, as these years have more resemblance to real job performance. Hence, there is a clear need to use reliable and valid selection tools and to evaluate the admission process afterwards.[1]

This study aims to examine the validity of a new format of tests, namely situational judgment tests in the context of dental student selection. SJTs present applicants with written or video-based descriptions of hypothetical scenarios and ask them to indicate the appropriate response from a list of alternatives.[2,3] The context of this study is admission to dental school in the Flemish part of Belgium.

Cognitive and Non-Cognitive Predictors of Academic Performance

In many countries, pre-admission academic grades (Grade Point Average, GPA) and/or cognitive-oriented tests are used to select students for medical and dental education. Research evidence shows that pre-admission academic grades predict subsequent course-academic performance in health disciplines.[4,5] These results obtained in medical and dental education mirror meta-analytic findings of the validity of cognitive factors (GPA and standardized ability tests) for predicting a variety of academic performance outcomes in higher education in general.[6,7] For example, Sackett and his colleagues[8] examined various large data sets and found strong relationships between standardized cognitive tests and academic performance (r=.44).

However, in dental education research, the relationship between grade point average (GPA) and academic performance was stronger in the earlier years of the education program.[1] For example, one study showed that the Dental Aptitude test was a good predictor of preclinical academic success, with prediction declining when clinical components of the program were introduced into the curriculum.[9]

This highlights that cognitive factors explain only part of the variance in academic performance. Hence, admission procedures should include assessment of both cognitive and non-cognitive characteristics of applicants. The need to incorporate more than just cognitive factors has led to a growing interest in exploring possible supplemental predictors of academic performance, particularly those outside the cognitive domain.[10] For instance, in some countries (e.g., the UK) interviews are used in the admission process whereby each individual is scored on five criteria: professionalism, communication skills, manual skill, leadership/team experience and non-academic interest. Results of Hoad-Reddick and McFarlane revealed that dental applicants with high interview scores on the criterion leadership experience, performed better.[11] Smithers, Catano, and Cunningham further suggested that an interview may be useful in identifying specific behavioral characteristics deemed important for success in dental training.[9]

Besides interviews, the use of personality inventories in selecting students for dental education has also been explored. Results from a personality measure used by Chamberlain, Catano and Cunningham indicated that Conscientiousness and Neuroticism, and to a lesser extent Agreeableness were significant predictors of both first-year academic performance of dental students as well as professional behavior of dental practitioners.[12] Cariago-Lo and his colleagues concluded that the California Psychological Inventory could discriminate among medical

students who performed well and those who did not.[13] Smithers, Catano and Cunningham found that Openness to Experience was significantly related to aspects of clinical education, although, contrary to expectations, this relationship was negative.[9] A facet of Openness, Ideas, together with Positive Emotions, a facet of Extroversion, improved prediction of performance in clinical studies beyond that provided by the Dental Aptitude Test and the interview. Poole, Catano and Cunningham suggested that a combination of scores from the Dental Admission Test (DAT), a valid measure of personality, and a well-designed structured interview provided the best prediction of those applicants who will do well in both the academic and clinical aspects of dental school.[14]

In recent years, there has been a surge of research in another non-cognitive test namely, namely the Situational Judgment Test (SJT). In employment settings, three meta-analyses indicate that SJTs are related to important job performance criteria. McDaniel, Morgeson, Finnegan, Campion and Braverman report a mean corrected correlation between SJTs and job performance of .34.[15] The second meta-analysis by McDaniel, Hartman, Whetzel, and Grubb reports a mean corrected validity of .26.[16] In terms of incremental validity, SJTs accounted for additional variance (varying from 1% to 2%) over both cognitive ability and personality. Third, Christian, Edwards and Bradley found validity coefficients ranging from .19 to .43.[17]

In light of these promising results for SJTs in employment selection settings, it is understandable that there is also increasing interest to use SJTs in educational admission settings. Evidence that SJTs are valid in medical admission settings was provided by Lievens, Buyse and Sackett.[18] They explored the use of an interpersonal SJT in the Belgian medical college admission context. This SJT predicted GPA in interpersonal skills courses and had incremental

validity over cognitive tests for predicting such interpersonal GPA. Patterson also studied the use of an SJT for selection into postgraduate general practitioners training in the UK.[19] This SJT focused on three non-clinical selection criteria: empathy, integrity and coping with pressure. The SJT was the best single predictor of performance in a selection centre that used work-relevant simulations to target both clinical and non-clinical domains.

Situational Judgment Tests and Admission to Dental Studies

As discussed, SJTs can be valid predictors of non-cognitive skills in medical education. To our knowledge, research on the validity of SJTs in dental education is non-existing. On the one hand, arguments can be made that the good results regarding validity of SJTs that were found in medical selection will translate to dental selection. One can assume that candidates who get selected for medical and dental education should be capable learners, open-minded and communicative, and socially competent. Doctors and dentists, of whatever specialty, need specialist medical knowledge and a complementary palette of skills and personality traits if they are to be professionally competent.[20] Hence, using an interpersonal SJT in a dental selection context is worth considering.

On the other hand, there are also arguments that the good results of SJTs in medical settings will not extrapolate to dental settings. In fact, medical and dental students have been found to differ on various characteristics. For example, Lindemann noted differences between dental and medical students with regard to learning approaches, especially upon entrance to professional school, which suggests that students enter with different academic studying experience and strategies.[21] Other researchers found that dental students were significantly more likely to be motivated by factors relating to status and security and the nature of their occupation

(e.g., regular working hours, self employment and independence). By contrast, medical students were significantly more likely to be motivated by factors relating to career opportunities, patient care, working with people, use of personal skills, and interest in science.[22]

Research Objectives

This study has two main research objectives. First, we examine the validity of an SJT measuring interpersonal skills for predicting academic performance of dental students. In most medical/dental schools (as in the ones in this study), earlier courses focus on the acquisition of knowledge, whereas later courses place more emphasis on communication with patients and internships, thus activities that involve significant interpersonal interactions. Hence, grades in clinical years of dental school may be better predicted by interpersonal skills as measured by SJTs than grades in the first years. Second, as SJTs claim to measure skills other than cognitive abilities, we examine whether an SJT will explain incremental variance over cognitive tests for predicting academic performance.

METHOD

Procedure and Sample

This study was situated in the context of admission to medical and dental studies in Belgium. The admission exam was institutionalized in 1997. Each year, this admission exam lasts for a whole day and it is centrally administered in a large hall in Brussels.

One difference from admission practices in the U.S. is that the process in Belgium is centralized and government-run. All students interested in medical and dental studies take an examination battery. Those who pass receive a certificate that permits entry into any of the six

medical schools in Belgium. Thus, individual medical schools are not involved in the screening of candidates. This also means that the level of selectivity in Belgium is generally less strict than the level of selectivity in some U.S. medical schools. A second difference is that students enter medical and dental studies at a younger age (e.g., about 19 years of age), rather than upon completion of an undergraduate degree, as is more typical in the U.S.

This study included 12 entering cohorts of dental students in Belgium. The total applicant pool consisted of 22.498 students (36.7% male, 63.3% female; average age= 18 years and 9 months; 99.5% Caucasian) who completed the Medical and Dental Studies Admission Exam in Belgium between 1997 and 2008. On average, the passing rate of the admission exam was about 30%. Note that both medical and dental students were selected with the same admission exam.[18] Students had to indicate their choice of education (medicine or dentistry) only after passing the exam. While the total applicant pool was used for purposes of range restriction corrections to estimate validity in the applicant pool, the study focused on all 796 candidates who passed the exam and undertook dental studies at one of the two dental schools in Flanders.

Predictor Measures

The Flemish admission exam assesses various characteristics that contribute to learning or performance in medical and dental school. In particular, the exam measures knowledge in sciences and general cognitive ability. Besides these cognitive predictors, the admission exam also consisted of two additional tests, namely a silent reading protocol and a situational judgment test. These two tests are work samples because they present candidates with tasks they will encounter in their study (reading and understanding texts with a medical subject) and in the

profession (patient interactions). The following describes the development and content of the tests used in this study.

The cognitive part of the admission exam consisted of two main tests. The first part was designed to evaluate applicants' mastery of 4 basic science-related subjects (mathematics, physics, chemistry, and biology). Per subject, 10 multiple choice questions were asked. Every question had 4 possible answers of which only one was correct. Second, there is a cognitive ability test which consisted of 50 multiple choice items with 5 response alternatives per item. The problems in this general mental ability test were formulated in either verbal, numerical or figural terms. Prior research demonstrated the good reliability and predictive validity of this reasoning test for medical and dental students.[23] In particular, this study 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. In light of test security, the source of this cognitive ability test cannot be mentioned. For the same reason, sample items are not presented. Interested researchers may contact the authors to obtain more information.

The silent reading protocol was a written text that was specifically developed for the admission exam each year. The underlying rationale was to ask candidates to read and understand an article with a medical content (e.g., diabetes, lower back pain,…). Each text was about 10 pages long and included tables and figures, but no statistics. All difficult medical words were explained in an endnote. Candidates had 50 min to read the text and answer 30 questions. All questions were multiple-choice with four possible answers. Each year, the same procedure was used to develop the text and accompanying questions. An existing medical text in a popular journal or handbook served as starting point. Next, a professor in medicine developed a more elaborate version of the original. Finally, two professors in medicine assisted us in developing a

list of relevant questions and response options. Due to test security reasons, pilot testing was not possible and dropping questions after receiving applicant data was forbidden. Across the exams, the average internal consistency coefficient of this test was .74.

The SJT. In the context of the admission exam, an SJT with situations about interactions with patients was developed. The general aim of the SJT used in the admission exam was to measure interpersonal and communication skills. We used an approach analogous to other studies for developing a video-based SJT.[24] First, we collected realistic critical incidents regarding interactions between physicians/dentists and patients from experienced physicians/dentists and professors in general medicine. Second, vignettes that nested the critical interpersonal incidents were written. Two professors teaching consulting practices tested these vignettes for realism. Similarly, questions and response options were derived. Third, semiprofessional actors were hired and videotaped in a recording studio. Finally, a panel of experts (experienced physicians/dentists and professors) developed a scoring key. Agreement among the experts was generally satisfactory (Cohen's kappa's > .70) and discrepancies were resolved upon discussion, leading to the scoring rule. The scoring key indicated which response alternative was correct for a given item (+ 1 points). It was forbidden by law to use different scoring rules (e.g., penalizing for choosing an incorrect answer by assigning -1 points).

In its final form, the SJT consisted of short videotaped vignettes of key interpersonal situations that physicians/dentists are likely to encounter with patients. A narrator introduced each vignette. After each critical incident, the scene froze, and candidates received 25 seconds to answer the question ("What is the most effective response?") related to the scene. No prior medical or dental knowledge was required as the items dealt with basic interpersonal situations. In total, the SJT consisted of 30 questions of the multiple-choice type, with four response

alternatives each. The alternate form reliability of the SJTs was .66 [18], which is in line with prior studies.[25]

Total decision score. To make the actual admission decision, a weighted sum of all predictors was computed. Next, a minimal cut-off was determined on this operational composite. Weights and cut-off scores were determined by law, with the cognitive tests receiving the most weight.

Criterion Measure

The criterion consisted of Grade Point Average (GPA) in each of the five years of dental training at the only two dental schools in Flanders. This GPA was a composite (average) measure derived from course grades. These courses covered topics such as preventive dentistry, chemistry, preclinical exercises, manual dexterity, internships, dermatology, etc. In the last year of the curriculum (year 5) there was an internship. Only overall GPA was made available to us.

As this study is longitudinal, students will have contributed data for several years. Not all students contributed data for their entire academic career as some students have only recently entered dental school. Hence, the performance of student cohorts was tracked over a one-, two-, three-, four- or five-year period, depending upon their year in the dental program, and correlated with their admission exam scores. As can be seen in table 1, first year data were available for 781 students, dropping to 489 for the second year, 411 for the third year, 343 for the fourth year and 274 for the fifth year.

Note too that analyses were also conducted only for cohorts for which criterion data for the full academic curriculum (5 years) were available. As those results were identical to the ones

presented in the tables, we present results for all available cohorts because the sample sizes are then larger.

Study participants are a more homogeneous group than the pool of applicants from which they were selected. The increase in homogeneity has the effect of underestimating the true size of a correlation coefficient in the applicant population. Therefore, we corrected the correlations for multivariate range restriction. To this end, we applied the multivariate range restriction formulas of Ree and his colleagues to the uncorrected correlation matrix.[26] As suggested by Sackett and Yang, statistical significance was determined prior to correcting the correlations.[27]

RESULTS

Validity of Cognitive and Non-Cognitive Tests

Table 1 shows that the validity of the SJT increased from year 1 (uncorrected r=-.01, corrected r=.05) to year 5 (uncorrected r=.17, corrected r=.20). The uncorrected correlation between the SJT and overall GPA was .04 (corrected .14).

The corrected correlation between the cognitive composite and overall GPA was .38. The validity of the cognitive composite was significant in the first three years of dental education but it dropped from .45 (year 1) to .18 (year 5). In the last two years, the correlation of the cognitive composite with GPA was not significant. This is possibly due to the fact that other components of the program are introduced into the curriculum in these last two years (e.g., Clinical internships). Results in table 1 also show that the total admission exam is a good predictor of preclinical and clinical academic success. The silent reading protocol is not a significant predictor in any of the five years of dental education.

Table 1. Correlations among Predictors and Overall Criteria

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Predictors (N=796)										
1. Cognitive part		-.01	-.03	.77**	.17**	.12**	.10*	.09	.04	.16**
2. Silent reading protocol	.23		.03	.25**	.02	.01	.06	-.02	.05	-.00
3. SJT	.08	.18		.17**	-.01	.04	.09	.10	.17**	.04
4. Total decision score	.85	.42	.20		.18**	.16**	.16**	.13*	.16*	.19**
Criteria										
5. GPA year 1 (781)	.45	.18	.05	.47		.70**	.59**	.51**	.36**	.92**
6. GPA year 2 (489)	.39	.11	.08	.45	.78		.68**	.59**	.38**	.88**
7. GPA year 3 (411)	.33	.10	.15	.39	.61	.69		.74**	.47**	.87**
8. GPA year 4 (343)	.25	.04	.10	.28	.53	.60	.75		.63**	.86**
9. GPA year 5 (274)	.18	.20	.20	.26	.41	.42	.52	.64		.72**
10. GPA overall (781)	.38	.13	.14	.45	.79	.85	.87	.86	.74	

Note. Uncorrected correlations are above the diagonal, corrected correlations below the diagonal. Correlations were corrected for multivariate range restriction. * $p<.05$; ** $p<.01$

Next, we examined whether the SJT had incremental validity over cognitive tests for predicting GPA in dental education. To this end, we conducted hierarchical regression analyses. The cognitive composite was entered as a first block. Next, we entered the silent reading text. Finally, the SJT was entered. The results of these hierarchical regression analyses are presented in table 2. The SJT had incremental validity over the cognitive composite and the reading text, only in year 5 of dental education. Again, the inclusion of internships in that particular year, might explain this finding.

Table 2. Summary of Hierarchical Regression Analyses of Predictors on GPA

	Criteria														
	GPA year 1			GPA year 2			GPA year 3			GPA year 4			GPA year 5		
Model predictors	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2
1. Cognitive part	.17**	.03	.03**	.12**	.01	.01**	.10*	.01	.01*	.09	.01	.01	.04	.00	.00
2. Silent reading protocol	.02	.03	.00	.01	.01	.00	.05	.01	.00	-.03	.01	.00	.03	.00	.00
3. SJT	-.00	.03	.00	.04	.02	.00	.09	.02	.01	.10	.02	.01	.16**	.03	.03**

* $p<.05$; ** $p<.01$

DISCUSSION

The task of selecting the best medical and dental applicants out of an extremely competitive applicant pool is a problem faced annually by medical and dental faculties all over the world. Furthermore, there is a responsibility on admissions committees to seek evidence that the selection instruments used deliver appropriate outcomes. Therefore, this study examined the validity of the dental admission procedure in Flanders for predicting GPA along the dental curriculum. A unique aspect of this procedure is the use of an SJT in the selection of dental students.

First, the results of this study confirm the finding that cognitive predictors are valuable and necessary tools in the selection of students for dental education. The cognitive composite was a significantly valid predictor of GPA in three of the five years of dental education. Note that the validity decreased in the clinical years. This result was expected, as the later years of dental education focus on internships and practice, and no longer purely on the acquisition of new knowledge.

Second, this study extends the positive predictive validity results of SJTs found in medical education to dental education. That is, an SJT that measures interpersonal capacities has incremental validity over cognitive tests. This result applies to year 5 only, which is explained by the fact that most courses in the curriculum in year 5 involve interaction with real life patients as compared to earlier years where mostly manual dexterity is taught. Note that we are not positing that alternative measures such as SJTs should be used to replace cognitive measures. Instead, we suggest that they can be valuable additions to extant cognitive measures. Future research should examine whether our results can be confirmed when actual job performance as a dentist serves as criterion.

This study describes a single selection procedure in a specific setting. Hence, no claims concerning generalizability can be made. However, we do believe that our results are interesting fro admission systems in other countries. In any country, dentists of the future face many challenges. They should be good and fast at acquiring manual skills. They should also be open-minded and tolerant, communicative, and socially competent. To reach these objectives in the future, committees conceptualizing admission procedures for dental education should design selection procedures that include both cognitive and non-cognitive skills. Along these lines, the SJT might be a useful supplement to cognitive tests.

REFERENCES

1 Salvatori P. Reliability and validity of admissions tools used to select students for the health profession. Adv Health Sci Educ 2001; 6: 159-75.

2 Motowidlo SJ, Dunnette MD, Carter GW. An alternative selection procedure: The low-fidelity simulation. J Appl Psychol 1990; 75: 640-7.

3 Weekley JA, Jones C. Further studies on situational tests. Personnel Psychology 1999; 52: 679-99.

4 Katz GM, Mosey AC. Fieldwork performance, academic grades, and pre-selection criteria of occupational therapy students. Am J Occup Ther 1980; 34: 794-800.

5 Levine SB, Knecht HG, Eisen RG. Selection of physical therapy students: interview methods and academic predictors. J Allied Health 1886; 15: 143-51.

6 Kuncel NR, Credé M, Thomas LL. A comprehensive meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA). Academy of Management Learning and Education 2007; 6: 51-68.

7 Kuncel NR, Hezlett SA. Standardized tests predict graduate students' success. Science 2007; 315: 1080-1.

8 Sackett PR, Kuncel NR, Arneson JJ, Coopers SR, Waters SD. Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? Psychol Bull 2009; 135: 1-22.

9 Smither S, Catano VM, Cunningham DP. What predicts performance in Canadian Dental School? J Dent Educ 2004; 68: 598-613.

10 Schmitt N, Keeney J, Oswald FL, Pleskac TJ, Bilington AQ, Sinha R, Zorzie, M. Prediciton of 4-year college student performance using cognitive and non-cognitive predictors and the impact on demographic status of admitted students. J Appl Psychol 2009; 94: 1479-97.

11 Hoad-Reddick G, MacFarlane TV. An analysis of an admissions system: can performance in the first year of the dental course be predicted? Br Dent J 1999; 186: 138-42.

12 Chamberlain TC, Catano VM, Cunningham DP. Personality as a predictor of professional behavior in dental school: comparisons with dental practitioners. J Dent Educ 2005; 69: 1222-37.

13 Cariago-Lo LD, Enarson CE, Crandall, SJ, Zaccaro DJ, Richards, BF. Cognitive and noncognitive predictors of academic difficulty and attrition. Acad Med 1997; 60-71.

14 Poole A, Catano VM, Cunningham DP. Predicting performance in Canadian Dental Schools: The new CDA structured interview, a new personality assessment, and the DAT. J Dent Educ 2007; 71, 664-76.

15 McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. Use of situational judgment tests to predict job performance: A clarification of the literature. J Appl Psychol 2001; 86: 730-40.

16 McDaniel MA, Hartman NS, Whetzel DL, Grubb WL. Situational judgment tests, response instructions, and validity: A meta-analysis. Personnel Psychology 2007; 60: 63-91.

17 Christian MS, Edwards BD, Bradley JC. Situational judgment tests: Construct assessed and a meta-analysis of their criterion-related validities. Personnel Psychology 2010; 63: 83-117.

18 Lievens F, Buyse T, Sackett P. The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. J Appl Psychol 2005; 90: 442-52.

19 Patterson F, Baron H, Carr V, Plint S, Lane P. Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. Med Educ 2009; 43: 50-7.

20 Powis D. Improving the selection of medical students. BMJ 2010; 340: 708

21 Lindemann R, Duek J, Wilkerson L. A comparison of changes in dental students' and medical students' approaches to learning during professional training. Eur J Dent Educ 2001; 5: 162-7.

22 Crossley ML, Mubarik A. A comparative investigation of dental and medical student's motivation towards career choice. Br Dent J 2002; 193: 471-3.

23 Minnaert A. 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 1996.

24 Weekley JA, Jones C. Video-based situational testing. Personnel Psychology 1997; 50: 25-49.

25 Clause CS, Mullins ME, Nee MT, Pulakos E, Schmitt N. Parallel test form development: A procedure for alternate predictors and an example. Personnel Psychology 1998, 193-208.

26 Ree MJ, Carretta TR, Earles JA, Albert W. Sign changes when correcting for restriction of range: A note on Pearson's and Lawley's selection formulas. J Appl Psychol 1994; 79: 298-301.

27 Sackett PR, Yang H. Correction for range restriction: An expanded typology. J Appl Psychol 2000; 85: 112-8.

CHAPTER 4

THE LONG-TERM PREDICTIVE AND INCREMENTAL VALIDITY OF OPERATIONAL SJTS IN HIGH-STAKES SELECTION

Whether situational judgment tests (SJTs) used in high-stakes settings with actual applicants are able to predict performance in the long run is an under-examined question. This study fills this key gap in the SJT domain by examining the long-term predictive and incremental validity of an SJT used in academic admissions. This study included four cohorts of medical students (4,538 applicants, 724 entering students, 519 graduates) in Belgium. Criterion data for the full academic curriculum (seven years) were available as well as later job performance ratings. Over time (from year 1 through year 7) the validities of the SJT for predicting academic performance (GPA) slightly increased and there was evidence of incremental validity of the SJT over cognitive ability. When domain-relevant academic performance (interpersonal GPA) served as criterion, the validity of the SJT remained constant. Finally, the SJT was a predictor of supervisory-rated job performance nine years later. The implications of these findings for research on the long-term validity of selection procedures are discussed.

INTRODUCTION

"We suggest that the most pressing need in future SJT research is to determine the extent to which conclusions, largely based on concurrent samples, will generalize to applicant samples. Applicants complete SJTs under high-stakes situations that likely have an impact on their motivation" (Whetzel & McDaniel, 2009, p. 199).

"Concurrent, cross-sectional studies are suggestive … Therefore, we would recommend the use of longitudinal, predictive criterion-validation designs" (Christian, Edwards, & Bradley, 2010, p. 108).

The quotes above come from the two most recent quantitative and qualitative reviews of situational judgment tests (SJTs). Although these reviews showed that SJTs have become established alternative predictor instruments in the personnel selection domain, they also revealed key gaps in our SJT knowledge. As noted above, one key gap is that SJT criterion-related and incremental validities have been mostly based on concurrent designs instead of on predictive designs with actual applicants in high-stakes settings. In addition, little is known about whether SJTs in such settings are able to predict performance in the long run.

This study aims to fill these two critical gaps in the SJT domain. Therefore, we examine the long-term predictive and incremental validity of an SJT that was used in an actual high-stakes setting (i.e., medical school admission) for predicting performance. Criterion data including both academic performance and job performance upon completion of medical school up to nine years after admission were gathered.

STUDY BACKGROUND

The Criterion-Related Validity of SJTs

To date, three meta-analyses of the criterion-related validity of SJTs have been conducted. McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001) conducted the first meta-analysis of the validity of SJTs in employment settings. They reported a mean corrected correlation between SJTs and job performance of .34. Inspection of the studies included revealed that only 6 out of 102 studies were predictive validity studies. Moreover, there was a marked difference between the mean validity coefficient for predictive validity studies (corrected *r* of .18) and that for concurrent validity studies (corrected *r* of .35).

The second meta-analysis, by McDaniel, Hartman, Whetzel, and Grubb (2007) made a distinction between SJTs with a knowledge-based format and SJTs with a behavioral tendency format. Both formats produced similar validities, with a mean corrected validity of .26. In terms of incremental validity, SJTs accounted for additional variance (varying from 1% to 2%) over both cognitive ability and personality. Again, it was striking that the number of predictive studies was very scarce as only 4 of the 118 studies included were predictive validity studies.

Third, Christian et al. (2010) conducted a meta-analysis of the validity of SJTs for predicting specific criterion constructs (e.g., leadership, interpersonal skills, teamwork). Results showed that the validity of SJTs was higher for predicting conceptually-related performance dimensions (e.g., a teamwork SJT showed higher relationships with teamwork criteria than with leadership criteria), underscoring the importance of predictor-criterion matching. Again, it should be noted that only 6 out of the 84 studies included in this most recent meta-analysis were predictive validity studies.

Thus, these three meta-analyses indicate that in employment settings SJTs are related to important job performance criteria. In addition, the incremental validity of SJTs over cognitive ability and personality indicates that SJTs permit measuring other constructs. In addition, other meta-analytic research shows that SJTs have less adverse impact against minorities than cognitive ability tests (especially if the cognitive loading of the SJT is low, Whetzel, McDaniel, & Nguyen, 2008).

In light of these advantages it comes as no surprise that there is also increasing interest in using SJTs in high-stakes admission settings (Lievens, Buyse, & Sackett, 2005; Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Schmitt et al., 2009). Oswald et al. (2004) found that an SJT had incremental validity over college-entrance tests and personality for predicting first-year GPA and self/peer ratings on a broad range of performance dimensions (e.g., leadership). Recently, Schmitt et al. (2009) extended these findings to the prediction of four-year GPA. Although the students in those two studies completed the SJT for research purposes, there is also evidence that speaks to the validity of SJTs in actual admission contexts. In particular, Lievens et al. (2005) explored the use of an interpersonal SJT in an actual medical college admission context. The SJT predicted GPA in interpersonal skills courses and had incremental validity over cognitive tests for predicting such interpersonal GPA.

Are SJTs Valid in High-Stakes Operational Use?

Although our review above shows promise for the use and validity of SJTs, an important limitation is that almost all conclusions about SJT validity are based on concurrent validation designs. Recently, Whetzel and McDaniel (2009) cogently summarized the key drawbacks of sole reliance on concurrent validity studies as follows:

"As with most personnel selection validity literatures, most SJT validity studies rely on concurrent designs. In such designs, respondents are incumbents who typically have little motivation to distort their responses… However, operationally, tests are given to job applicants who, on average, may be motivated to distort their responses (i.e., fake to look good) because the test scores are used in determining whether they get hired. Thus, because SJT research primarily uses concurrent studies, it is possible that some of the conclusions drawn in this review may not hold for SJTs used to screen job applicants." (p. 190).

Indeed, when one considers the use of SJTs in high-stakes testing contexts among applicants, a unique set of issues arise, including the possibility of faking and seeking coaching. Although extensive research exists on the faking and coaching of personality, biodata, and integrity tests, these issues have received less attention in the SJT field, with most studies being laboratory studies with extreme groups (e.g., fake vs. honest; coached vs. uncoached).With respect to faking effects on SJTs Hooper, Cullen, and Sackett (2006) summarized the available research evidence and concluded respondents can improve their scores by faking if instructed to do so, with *d* varying from .08 to .89 *SD*. Hooper et al. (2006) emphasized that few studies have investigated the effects of faking good on the criterion-related validity of SJTs. One lab study showed that faking reduced criterion-related validity from $r = .33$ to $r = .09$ (Peeters and Lievens, 2005). Regarding coaching effects, SJT research is even scarcer. Cullen, Sackett, and Lievens (2006) conducted a lab study to examine the coachability of SJTs. They focused on SJTs developed for use in college admissions, and found that some of these SJTs were susceptible to coaching.

So, these results of faking and coaching effects of SJTs show that caution should be exerted with respect to generalizing SJT findings obtained in low-stakes contexts to high-stakes

contexts (see also MacKenzie, Ployhart, Weekley, & Ehlers, 2010). The finding of lower mean validity in the small number of existing predictive studies (i.e., mean r = .18, vs. .35 for concurrent studies) in the McDaniel et al. (2001) review suggests that the faking and coaching issues associated with a high-stakes environment do not negate the validity of the SJTs in question. However, more studies in operational settings are needed to bolster our understanding of the level of predictive validity that one might anticipate in operational use.

Do SJTs Used in High-Stakes Settings Have Long-term Validity?

Apart from the lack of predictive validation designs with actual candidates, a second drawback is that the long-term validity of SJTs has not been scrutinized. This is a key concern as five decades ago Humphreys (1960) stated that "in selection research one should not be satisfied with validation of predictors against the earliest possible criteria" (p. 318). Although there exists a large literature that is directed at whether or not cognitive ability tests retain their predictive value in the long run (Barrett, Phillips, & Alexander, 1981; Campbell & Knapp, 2001; Deadrick & Madigan, 1990; Schmidt, Hunter, Outerbridge, & Goff, 1988), few studies have focused on the long-term predictive validity of non-cognitive predictors such as assessment centers (Howard & Bray, 1988; Hinrichs, 1978; Jansen and Stoop, 2001) or personality (Lievens, Ones, & Dilchert, 2009; Thoresen, Bradley, Bliese, & Thoresen, 2004; Stewart, 1999; Stewart & Nandleolyar, 2006).

Our review above illustrates that SJTs are no exception to the general validation practice of using concurrent or short-term predictive designs for examining the validity of noncognitive predictors. In concurrent studies, criterion scores have been typically obtained from both newly selected individuals as well as individuals of varying tenure levels. In addition, in the scarce

predictive validation studies the time spans over which criteria have been gathered rarely exceeded a year or two, in most cases they are merely a few months. One exception is Lievens et al. (2005) wherein one entering cohort had proceeded as far as year four of a seven-year medical curriculum. The present study follows four entering cohorts to completion of a seven-year curriculum, and then follows a subset of these through two years of post degree job performance.

Expectations regarding the long-term validity of SJTs are usefully informed by the literature on validity change over time for other predictors and by the literature on dynamic criteria (Alvares & Hulin, 1972; Barrett et al., 1981; Campbell & Knapp, 2001; Deadrick & Madigan, 1990; Ghiselli, 1956; Schmidt et al., 1988). For longitudinal changes in predictor validity, two primary explanations have been proposed. According to the "*changing person*" model, individuals change over time which would mean that their behavior would change to reflect this change. According to the "*changing task*" model, tasks and work being performed change (Alvares & Hulin, 1972).

This changing ability/person explanation has now been largely rejected in the ability domain. Postdictive validities appear to follow the same patterns of changes as predictive validities (Humphreys & Taber, 1973; Lunneborg & Lunneborg, 1970). Similar arguments of stability can be made for personality traits which conceptually reflect stable individual differences. Recent meta-analytic evidence (Fraley & Roberts, 2005; Roberts & DelVecchio, 2000) suggests that rank-order stability is remarkably high (Caspi, Roberts, & Shiner, 2005).

At first glance one might expect a different pattern of findings in the interpersonal skills domain which is the subject of the SJT in the present study, as training programs aim to change these skills, and are successful at doing so. Arthur, Bennett, Edens, and Bell's (2003) meta-analysis of training program effectiveness reports mean *d*s for interpersonal skills of .68 for

learning criteria and .54 for behavioral criteria. However, in considering the implications of this for the changing ability/person model it is important to consider the implications for SJT validity of different types of change. Here we consider four possible ways for interpersonal skills to be changed by intervention. First, an intervention might improve the skills of all individuals by a comparable amount, in which case the validity of a predictor of interpersonal skills would be unaffected. Second, an intervention might improve the skills of those with severe deficits, but have little impact on those with good skills. In this case, it is possible that rank order is unchanged; all that is seen is a tightening of the distribution, and the validity of a predictor of interpersonal skills is also unaffected. Third, the intervention might train all individuals to a common level of interpersonal skill, in which variance would be reduced to zero, and therefore validity of a predictor would also go to zero. Fourth, the intervention might be differentially effective, resulting in substantial change in the rank ordering of individuals in terms of their interpersonal skills, and thus in substantial reduction in validity. Thus, the first two possible forms of "changing abilities" pose no threat to validity, while the last two forms do pose a threat. However, we note that if either of these latter two forms were the true state of affairs, one would observe very low pretest- posttest correlations between measures of interpersonal skills. In contrast, a high pretest-posttest correlation would be strong evidence against these latter two forms. We find such evidence in a meta-analysis by Taylor, Russ-Eft, and Chan (2005) of behavioral modeling training programs aimed at interpersonal skills. They reported a mean pretest-posttest correlation of .84 across 21 studies for the effects of training on job behaviors, which is inconsistent with either the "training eliminates variance" or the "training radically alters rank order" perspectives on change. Thus, we believe that the forms of a "changing

persons" argument that would lead to an expectation of reduced validity can also be rejected in the interpersonal skills domain.

As opposed to the changing person model, the changing task model has been successfully adopted as explanation for performance change across time (e.g., Alvares & Hulin, 1972). Several variants of the changing task model have also been incorporated into more recent theories of skill acquisition in the cognitive domain (e.g., Ackerman, 1987). This literature suggests that the temporal stability of predictor-criterion relationships for cognitive variables differs across types of abilities (general mental ability, psychomotor ability, perceptual ability), settings (educational, work), and types of work (consistent/inconsistent task performance, academic performance, job performance) (Keil & Cortina, 2001).

In this study, the changing task model can be used for formulating hypotheses about the validity of an SJT measuring interpersonal skills for predicting academic performance. In fact, in academic settings (e.g., medical school), earlier courses typically focus on the acquisition of declarative and procedural knowledge in medical sciences, mostly a cognitive exercise, whereas later courses place also more emphasis on contact with patients, applied practice, and internship performance, activities that involve significant interpersonal interactions. Due to this changing content of medical courses over time one might expect the importance of cognitive factors to eventually reduce, leaving room for other sorts of attributes. Hence, later grades in medical school may be better predicted by interpersonal skills as measured by SJTs than earlier grades. Apart from this conceptual argument, there is also empirical evidence of increasing criterion-related validities for noncognitive predictors such as personality traits and assessment center ratings. For example, Hinrichs (1978) found that assessment center ratings predicted organizational level better after 8 years post assessment than 1 year post assessment (see also

Bray & Howard, 1983; Jansen & Stoop, 2001). Recently, similar increases for specific personality traits for predicting academic performance in medical school over time have been found (Lievens et al., 2009).

In sum, our general hypothesis is that the validity of an SJT measuring interpersonal skills for predicting overall academic performance over the full curriculum (i.e., seven academic years) will increase over time (and, conversely, the validity of cognitive measures will decrease). As one of the main arguments behind the use of SJTs in high-stakes settings is that they enable the measurement of KSAOs other than cognitive tests, we also expect that an SJT will explain incremental variance over cognitive tests for predicting GPA over time. Thus,

Hypothesis 1a: The validity of an SJT used in a high-stakes context will increase for predicting GPA throughout medical school.

Hypothesis 1b: The validity of cognitive tests used in a high-stakes context will decrease for predicting GPA throughout medical school.

Hypothesis 1c: An SJT used in a high-stakes context will have incremental validity over cognitive tests for predicting GPA throughout medical school.

Our hypothesis above about the increasing validities of SJTs for predicting overall performance is grounded by the notion that the content of the criterion changes over time. Specifically, the changing task model posits that if the makeup of this study's criterion changes (i.e., over time, becoming more interpersonally loaded), the predictive power of the SJT is expected to change. One way of testing this more closely consists of investigating the validity of the SJT for predicting separate performance components. Hereby we make a distinction between

medical and interpersonal domains in medical academic performance (see also Lievens et al., 2005). A limitation of the scarce number of prior longitudinal studies is that only overall performance served as criterion so that it was difficult to make inferences about the substantive reasons why validity changed across time. In this study, we extend prior longitudinal research designs by investigating the validity of SJTs for predicting a specific performance domain over time.

In sum, the changing task model is important if the criterion of interest was an overall performance measure, as the contribution of specific domains (e.g., increasing importance of the interpersonal component) to overall performance may change across years in the academic curriculum. When performance is assessed separately in different domains (instead of using an overall performance measure), the issue of the possibility of changing importance of the domains for overall performance is held constant. Thus, in the present study, the changing task issue is then no longer a likely contributor to changes in validity over time for predicting separate performance domains. As noted above, the changing ability/person explanation can be rejected in the ability, personality, and interpersonal domain. Thus, given the above arguments against *both* a changing task and a changing person model, our expectation is that the SJT will remain a valid predictor of interpersonal performance over time. Thus we hypothesize:

Hypothesis 2: The validity of an SJT used in a high-stakes context will remain constant for predicting interpersonal GPA throughout medical school.

In this study, overall academic performance and its components as measured over the full curriculum is not the only criterion. Additionally, we examine the ability of an SJT used in high-

stakes settings to predict job performance gathered nine years after the administration of the SJT. We note, though, that the job performance measures available in the present study are overall measures, and thus we are unable to separate the job performance measures into separate technical and interpersonal components.

Few studies have examined whether selection procedures are able to predict *both* academic performance and job performance. Kuncel, Hezlett, and Ones (2004) meta-analytically examined the relationship between the Miller Analogies Tests (MAT) and both academic and job performance, as the MAT is one of the few tests that is operationally used for both educational admissions and personnel selection. They reported that the MAT predicts performance in both domains. However, that the same test can be used for both admissions and personnel selection purposes is a slightly different issue than whether a test administered at the time of application for educational admission retains its validity many years later as a predictor of job performance. The present study is a rare example of examining the latter issue. As we see job overall physician job performance as involving a combination of technical and interpersonal knowledge/skills we expect that an SJT measuring interpersonal skills will also be a good predictor of physicians' job performance gathered nine years after the administration of the SJT. The same reasoning applies to the incremental validity of the SJT over cognitive tests in predicting job performance. This leads to the following hypotheses.

Hypothesis 3a: An SJT used in a high-stakes context will show predictive validity for predicting job performance measured nine years after admission.

Hypothesis 3b: An SJT used in a high-stakes context will have incremental validity over cognitive tests for predicting job performance measured nine years after admission.

METHOD

Sample and Procedure

This study included four entering cohorts of medical students in Belgium. These cohorts were included because criterion data for the full academic curriculum (seven years) were available from these cohorts. The total applicant pool consisted of all 4,538 students (37% men and 63% women; average age = 18 years and 10 months; 99.5% Caucasian) who completed the Medical Studies Admission Exam in Belgium between 1999 and 2002. On average, the passing rate of the admission exam was about 30%. Candidates who passed the exam received a certificate that warranted entry in any medical university. Thus, there was no further selection on the part of the universities. However, not all students who passed the exam eventually chose to study medicine. While this total applicant pool was used for purposes of range restriction corrections to estimate validity in the applicant pool, the study focused on the 724 students who passed the exam and undertook medical studies at one of two large medical schools. We studied entrants at these two schools because we had access to detailed performance information at the level of the individual course, as well as information about the content of each course, thus permitting us to identify courses with an interpersonal component. These two medical schools did not differ in terms of medical curriculum from the other schools.

Criterion data (internship and job performance ratings) were obtained from archival records of those two universities. *N* for year 1 was 724. By the end of year 7 *N* dropped to 519. Student attrition due to failure (especially in the first academic year) was the most important reason for the reduction in sample size in the seven academic years. We report analyses based on the number of students present in a given year; all our analyses were also run with the group completing all seven years (*N*=519), with no substantive change in findings. Job performance

ratings were available only for the students who after their seven years of study decided to become physicians. This was about 10% of the students. Therefore, these analyses were based on *N*=64.

Predictor Measures

Predictors were gathered during the actual admission exam. Each year, the admission exam lasted for a whole day and was centrally administered in a large hall. The administration of the exam was highly standardized because it was guided by a minute-to-minute script. In the morning session, students completed the knowledge test. In the afternoon, they completed the cognitive ability test, the medical text, and the video-based SJT (physician-patient interaction). The following describes the development and content of each of the predictors used.

Knowledge test. Each year, an extensive panel of professors developed items to test knowledge related to four sciences (biology, chemistry, physics, and mathematics). Per science, there were 10 items with four possible answers. The candidates had three hours to solve these items. Across the exams included in this study, the average internal consistency coefficient of the knowledge test was .78.

Cognitive ability test. This test consisted of 50 items, each with five possible response alternatives. The items were formulated in verbal, numeric or figural terms and selected each year from a larger item pool. Hence, this was a broad cognitive ability test that aimed to measure general mental ability. The time limit was 50 minutes. In light of test security, the source of this cognitive ability test cannot be mentioned. For the same reason, sample items are not presented. Interested researchers may contact the authors to obtain more information. Prior research attested to the good reliability and predictive validity of this test for a medical student population. In

particular, Minnaert (1996) reported an internal consistency coefficient of .84 and a validity coefficient of .36 for predicting first-year GPA in medical studies.

In their meta-analysis, Kuncel, Hezlett, and Ones (2001) showed that a composite of general measures (e.g., Graduate Record Exam [GRE] verbal and numerical) combined with specific GRE subject-matter tests provided the highest validity in predicting academic performance. To provide the strongest test of the incremental validity of interpersonal skills, we used a cognitive composite that consisted of the four knowledge test score and the cognitive ability test score. Prior research demonstrated the satisfactory reliability and predictive validity of this cognitive composite for a medical student population (Lievens et al., 2005).

Written medical text. This test was specifically developed for the admission exam. The underlying rationale was to ask candidate medical students to read and understand an article with a medical subject matter. Therefore, this test can be considered as a miniaturized sample of tasks that students will encounter in their medical education. The text was about 10 pages long and it was conceived as a regular scientific article with tables and figures. No statistics were included, and all difficult medical terms were explained in an endnote. Students had 50 minutes to read the text and answer 30 questions (multiple-choice questions with four possible answers).

Each year, professors developed the text and the accompanying questions using the same procedure. An existing medical text in a popular medical journal or handbook served as starting point. Next, a professor in medicine developed a more elaborate version of the original text. Finally, two professors in medicine assisted in developing a list of relevant questions and response options. Across the exams, the average internal consistency coefficient of this test equaled .71.

Video-based SJT. There is an emerging consensus that SJTs are essentially measurement methods that can be designed to measure a variety of constructs, Chan & Schmitt, 2002; Christian et al., 2010; Whetzel & McDaniel, 2009). The general aim of the SJT used in the admission exam was to measure interpersonal and communication skills. Like the written medical text, this test was specifically developed for the admission exam.

An approach analogous to other studies (see e.g., Weekley & Jones, 1997) was used for developing the SJT. First, we collected realistic critical incidents regarding interactions between physicians and patients from experienced physicians and professors in general medicine. Second, vignettes that nested the critical interpersonal incidents were written. Two professors teaching physicians' consulting practices tested these vignettes for realism. Using a similar approach, questions and response options were derived. Third, semiprofessional actors were hired and videotaped in a recording studio. To guarantee realism, an experienced physician attended the set. Finally, a panel of experts (experienced physicians and professors in general medicine) developed a scoring key. Agreement among the experts was generally satisfactory (Cohen's kappa's > .70) and discrepancies were resolved upon discussion, leading to the scoring rule. The scoring key indicated which response alternative was correct for a given item (+1 point). It was forbidden by law to use different scoring rules (e.g., penalizing for choosing an incorrect answer by assigning -1 points). In its final version, the SJT consisted of short videotaped vignettes of key interpersonal situations that physicians are likely to encounter with patients. A narrator introduced each vignette. After each critical incident, the scene froze, and candidates received 25 seconds to answer the question ("What is the most effective response?") related to the scene. In total, the SJT consisted of 30 multiple-choice questions with four possible answers.

Across the exams included in this study, the average internal consistency coefficient for the SJT was .40. SJTs typically demonstrate low internal consistency because SJTs are construct heterogeneous at the item level (Whetzel & McDaniel, 2009).

Operational composite. To make actual admission decisions, a weighted sum of the aforementioned predictors (cognitively oriented tests, work sample, and SJT) was computed. Next, a minimal cutoff was determined on this operational composite. The weights and cutoff scores were determined by law.

Criterion Measures

Academic performance. As a first broad criterion, we gathered students' grade point average (GPA) at the end of each year. In Belgium, GPA is measured on a scale from 0 to 20, with higher scores indicating better grades. GPA correlated strongly across years, with the average corrected (for unreliability and indirect range restriction) correlation between GPA across years equaling .84. This value is similar to the values found in a recent meta-analysis about the temporal stability of GPA (Vey et al., 2003).

Similar to advancements into understanding the criterion space of job performance (Campbell, McCloy, Oppler, & Sager, 1993; Rotundo & Sackett, 2002), the multidimensionality of academic performance has recently been scrutinized (Oswald et al., 2004; Schmitt et al., 2009). Research has revealed that academic institutions consider student performance to be broader than traditional intellectual achievement. In line with this recent multidimensional conceptualization of academic performance, we differentiate the criterion of academic performance, assessed using grade point average (GPA), into two areas: medical GPA and interpersonal GPA (see also Lievens et al., 2005). To this end, two of the authors inspected

course descriptions of curricula and independently identified courses with a medical versus interpersonal component. The key inclusion criterion for the latter was that the course had to deal with communication with actual patients in the form of an internship (either short-term or long-term). Inter-rater agreement (ICC 2,1) among the authors was > .90. Discrepancies among the authors were resolved upon discussion. Next, the archival student grades on these courses were retrieved. In four of the seven academic years (i.e., in the first, fourth, sixth, and seventh year) of these universities, courses involving internships were identified. In the first year, these courses included introductory courses on patient interviewing and internships with a focus on observation. In the fourth year, multidisciplinary and communication skills courses and short-term internships were given to prepare students for clinical and professional practice. In the sixth and seventh year, several hospital-based clinical clerkships were included. This clerkship program was divided into various rotations (e.g., Children and Youth, Surgery, Primary Care), with two to four months spent in each unit. A composite score for each of these four years (called interpersonal GPA in the first, fourth, sixth, and seventh year, respectively) was obtained by averaging scores on interpersonal courses per year. Given differences across universities, we standardized students' interpersonal course grades within university and academic year. A composite interpersonal GPA measure (average interpersonal GPA across these four years) was also computed.

Apart from interpersonal GPA, we retrieved archival data on students' medical GPA in these same four years. This was a cognitively-oriented criterion measure as it consisted of grades on science and medical-related subjects. A composite score for each of the same four years as the interpersonal GPA (the first, fourth, sixth, and seventh year, respectively) was obtained by averaging scores on these medical courses per year. Again we standardized students' medical

course grades within university and academic year. A composite medical GPA measure (average medical GPA across these four years) was also computed.

As noted above, GPA data were obtained from archival records of two universities. *N* for year 1 was 724. By the end of year 7 *N* had dropped to 519. Student attrition due to failure (especially in the first academic year) was the most important reason for the reduction in sample size in the seven academic years. We report analyses based on the number of students present in a given year; all our analyses were also run with the group completing all seven years (*N*=519), with no substantive change in findings.

Job performance. A supervisory rating of job performance was included. Some of the medical students of these two medical universities (about 10%, *N* = 64) who ended their seven years of education, chose a career in general medicine, and entered a General Practitioner training program of up to two years duration. During that program, they worked under supervision (of a registered general practitioner) in a number of general practice placements. Hereby they were fully responsible for patients. All trainees were rated on a scale from 0 to 20 in practice at the end of the General Practitioner training program. The evaluations were completed by the trainee's General Practitioner supervisor, who had met regularly with them to discuss their progress. All supervisors were certified General Practitioners who had been approved as General Practitioner trainers with responsibility for supervising trainees. None of the supervisors had access to the trainees' admission exam scores when making their assessments.

As the above description refers to participants as "trainees", a question arises as to whether this should be viewed as a measure of "training performance" rather than "job performance". We view this as "job performance" in that these medical school graduates are engaged in full-time practice of medicine. They are working under supervision of a senior

General Practitioner charged with monitoring and evaluating their work, thus creating the opportunity to access these evaluations of performance for purposes of this study.

RESULTS

Preliminary Analyses

As we will test our hypotheses on data accumulated over four cohorts (four admission years, i.e., from 1999 to 2002) we began by examining whether the measurement structure underlying the admission exam was invariant across these years. A model with three factors, namely a cognitively-oriented factor (including the cognitive ability test and the four knowledge tests, see Kuncel et al., 2001), a factor on which the medical text loaded, and a factor related to the SJT, provided a good fit to the data. In particular, we tested a sequence of increasingly more restrictive tests of measurement invariance. As can be seen in Table 1, there was evidence of full measurement invariance across the four examinations because factor form, factor loadings, error variances, and factor variances/co-variances were found to be invariant across the examinations. In addition, the fit of the fully constrained model was still very good, *RNI* = .955, *CFI* = .973, and *RMSEA* = .050. Therefore, the remaining analyses will report the results for these three factors: cognitive test composite, medical text, and SJT.

Although the measurement model was found to be invariant across years, candidate mean scores per test might still differ across years. One potential reason is that the items of the admission exam were not identical across years. To preserve the integrity and the security of the tests, alternate forms per test were developed each year. Thus, we standardized candidates' test scores within each exam.

Table 1. Tests of Measurement Invariance for Multi-Group Three Factor Model of Admission Test Scores across Four Exam Years (N = 4,538)

	X^2	*df*	ΔX^2	*RNI*	*CFI*	Δ*CFI*	*AGFI*	*RMSEA*	90% CI of *RMSEA*
Equal number of factors	228.89**	110	--	.936	.971	--	.962	.061	[.053 - .069]
Equal factor loadings	230.17**	113	1.28	.941	.971	.000	.965	.059	[.051 - .066]
Equal error variances	231.97**	117	1.80	.944	.971	.000	.967	.056	[.049 -.063]
Equal factor variances/covariances	233.31**	127	1.34	.955	.973	-.002	.972	.050	[.043 -.057]

Note. *RNI* = Relative Noncentrality Index; *CFI* = Comparative Fit Index; *AGFI* = Adjusted Goodness Of Fit Index; *RMSEA* = Root Mean Square Error of Approximation.

Descriptive Statistics

Table 2 presents the means, standard deviations, and correlations among the predictors. One part of the table is based on all applicants who completed the admission tests between 1999 and 2002. As can be seen, the correlations among the three types of tests were small to moderate. The correlation between the cognitive ability test and the SJT was .20, indicating that the SJT was not heavily cognitively-loaded.

Table 2. Means, Standard Deviations, and Correlations among Predictors in the Sample

	Applicants (N = 4,538)					*Selectees (N = 723)*		*General practitioners (N = 64)*	
	M	*SD*	1.	2.	3.	*M*	*SD*	*M*	*SD*
1. Cognitive composite	11.68	2.65	--			14.08	1.67	13.49	1.47
2. Written text	15.17	4.74	.36	--		16.81	4.47	17.57	4.00
3. SJT	18.35	3.08	.20	.24	--	19.30	2.84	20.66	2.80
4. Operational composite	20.66	5.29	.91	.45	.28	24.90	3.89	25.98	1.98

Note. Although all analyses were conducted on standardized scores, this table presents the raw scores across exams. The maximum score on each test was 30, with the exception of the operational composite (maximum score = 40). Both the selectees (i.e., medical students) and general practitioners are subsamples of the applicant sample. Correlations between the predictors in the applicant group are presented. All correlations are significant at $p < .01$.

In the last four columns of Table 2, the means and standard deviations of the predictors in the selected group and the group who ultimately chose to work as general practitioners are displayed. So, this part of the table is based only on the subset of applicants that were selected (i.e., scored higher than the cut-off determined on the operational composite) and subsequently undertook medical studies in one of the two universities. A comparison of the descriptive

statistics related to the predictors in Table 2 reveals the degree of indirect range restriction (Thorndike's case 3) in each predictor due to the fact that the admission decision was made on the basis of a third variable (the operational composite). As noted, each predictor was weighted differently in the operational composite, resulting in differing degrees of indirect range restriction. Relative to the applicant pool, those selected scored 1.44 *SD* higher on the cognitive composite, .37 *SD* higher on the written text, and .33 *SD* higher on the SJT. So, as expected, there was more range restriction on the cognitive composite.

Validity of SJT for Predicting Academic Performance in the Long Run

Hypothesis 1a dealt with the long-term validity of the SJT for predicting GPA. As indirect range restriction is a special case of multivariate range restriction, we applied the multivariate range restriction formulas of Ree, Carretta, Earles, and Albert (1994) to the uncorrected correlation matrix. Statistical significance was determined prior to correcting the correlations (Sackett & Yang, 2000). The values below the diagonal of Table 3 represent the corrected correlations between the predictors and performance. The values above the diagonal are the uncorrected correlations.

Table 3 shows that the validity of the SJT slightly increased from year 1 (.10) to year 5 (.18). The last two years it dropped again but that might be due to the lower reliability of GPA in these last years (i.e., GPA was based on fewer courses). The correlation between the SJT and overall GPA was .13. Thus, there is partial support for Hypothesis 1a. While the validity of the cognitive tests was significant in all years, it decreased across the different academic years. The corrected correlation between the cognitive composite and GPA equaled .42 in year 1 and dropped to .25 by year 7.

Table 3. Correlations among Predictors and Overall Criteria in Selected Sample

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
Predictors (N = 724)															
1. Cognitive composite		-.01	-.06	.75	.32	.29	.30	.23	.17	.13	.18	.27	.29	.09	.14
2. Written text	.13		.13	.19	.02	.01	.06	.04	.07	.11	.07	.05	.06	.01	-.13
3. SJT	.03	.15		.11	.07	.07	.12	.16	.16	.08	.09	.10	.07	.21	.27
4. Operational composite	.86	.28	.16		.27	.26	.28	.18	.19	.12	.16	.23	.26	.09	.06
Criteria															
5. GPA (year 1, *N = 724*)	.42	.08	.10	.39		.74	.71	.61	.60	.49	.49	.88	.86	.45	.05
6. GPA (year 2, *N = 625*)	.39	.07	.10	.37	.77		.76	.72	.65	.52	.50	.86	.70	.47	.39
7. GPA (year 3, *N = 596*)	.41	.12	.15	.40	.73	.78		.81	.73	.60	.61	.90	.78	.48	.45
8. GPA (year 4, *N = 583*)	.30	.08	.18	.27	.63	.73	.82		.77	.59	.57	.88	.80	.56	.46
9. GPA (year 5, *N* = 576)	.26	.11	.18	.28	.62	.67	.74	.78		.66	.61	.87	.77	.58	.44
10. GPA (year 6, *N* = 520)	.18	.13	.10	.18	.51	.53	.60	.60	.67		.54	.77	.80	.63	.36
11. GPA (year 7, *N* = 519)	.25	.11	.11	.24	.51	.53	.62	.59	.62	.55		.75	.76	.60	.41
12. GPA (overall, *N = 724*)	.36	.10	.13	.34	.89	.87	.91	.89	.88	.77	.76		.93	.60	.33
13. Medical GPA (overall, *N* = 724)	.40	.12	.11	.38	.87	.73	.80	.81	.78	.80	.77	.93		.57	.25
14. Interpersonal GPA (overall, *N* = 724)	.13	.03	.22	.14	.46	.48	.49	.56	.58	.64	.61	.61	.58		.43
15. Job performance (supervisor, year 9, *N* = 64)	.15	-.11	.27	.10	.07	.39	.45	.47	.44	.37	.41	.33	.26	.43	

Note. Uncorrected correlations are above the diagonal, corrected correlations are below the diagonal. Correlations were corrected for multivariate range restriction. Apart from the last row, correlations higher than .09 are significant at .05 level; correlations higher than .12 are significant at .01 level. For the last row, correlations higher than .25 are significant at .05 level; correlations higher than .31 are significant at .01 level.

This decrease is in line with prior findings (Humphreys, 1968; Humphreys & Taber, 1973; Lin & Humphreys, 1977) and supports Hypothesis 1b.

Next, we examined whether SJTs used in a high-stakes context have incremental validity over cognitive tests for predicting GPA in the long run. To shed light on this hypothesis, we conducted hierarchical regression analyses. The matrices corrected for multivariate range restriction served as input for the hierarchical regression analyses. Statistical significance was determined prior to applying the corrections (by conducting hierarchical regressions on the uncorrected matrix of correlations). The cognitive test composite was entered as a first block because such tests have been traditionally used in medical admission exams. Next, we entered the medical text in the regression equation. Finally, we entered the SJT. The results are presented in Table 4. In all years (with the exception of the sixth one), the SJT explained incremental variance in GPA, thereby supporting Hypothesis 1c.

Table 4. Summary of Hierarchical Regression Analyses of Predictors on GPA

		Criterion																		
		GPA (year 1, *N*=724)			GPA (year 2, *N*=625)			GPA (year 3, *N*=596)			GPA (year 4, *N*=583)			GPA (year 5, *N*=576)			GPA (year 6, *N*=520)			GPA (year 7, *N*=519)		
Model	Predictors	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2	Beta	R^2	ΔR^2
1.	Cognitive composite	.32**	.10	.10**	.29**	.08	.08**	.31**	.09	.09**	.24**	.05	.05**	.18**	.03	.03**	.13**	.02	.02**	.18**	.03	.03**
2.	Reading text	.01	.10	.00	.00	.08	.00	.04	.09	.00	.02	.05	.00	.05	.03	.00	.10	.03	.01	.06	.04	.01
3.	SJT	.08	.11	.01*	.08	.09	.01*	.13	.11	.02**	.17**	.08	.03**	.17**	.06	.03**	.08	.03	.01	.09*	.04	.01*

*p < .05; **p < .01. The corrected matrices served as input for the regression analysis. Parameter estimates are for final step, not entry. Due to rounding, ΔR^2 differs .01 from the Cumulative R^2.

Table 5. Correlations among Predictors and Facet Criteria in Selected Sample

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
Predictors (N = 724)															
1. Cognitive composite		-.01	-.06	.75	.33	.23	.14	.13	.29	-.02	.11	.03	.06	.09	.14
2. Written text	.13		.13	.19	.03	.05	.14	.08	.06	.02	.01	.03	.03	.01	-.13
3. SJT	.03	.15		.11	.07	.12	.05	.07	.07	.17	.20	.16	.14	.21	.27
4. Operational composite	.86	.28	.16		.30	.19	.13	.12	.26	.02	.13	.01	.09	.09	.06
Criteria															
5. Medical GPA (year 1, *N = 724*)	.45	.09	.10	.43		.57	.42	.45	.89	.27	.43	.36	.28	.46	.01
6. Medical GPA (year 4, *N = 529*)	.31	.10	.14	.28	.60		.50	.51	.83	.23	.46	.33	.37	.52	.37
7. Medical GPA (year 6, *N = 521*)	.20	.16	.06	.20	.44	.51		.45	.77	.24	.35	.35	.35	.47	.23
8. Medical GPA (year 7, *N = 510*)	.18	.10	.08	.18	.47	.52	.46		.78	.15	.42	.41	.42	.50	.38
9. Medical GPA (overall, *N* = 724)	.40	.12	.11	.38	.90	.84	.77	.78		.31	.54	.47	.46	.57	.25
10. Interpersonal GPA (year 1, *N* = 724)	.00	.02	.17	.03	.26	.23	.24	.16	.31		.24	.24	.24	.72	.25
11. Interpersonal GPA (year 4, *N* = 529)	.18	.04	.21	.19	.45	.48	.37	.43	.55	.24		.37	.27	.70	.38
12. Interpersonal GPA (year 6, *N* = 521)	.03	.04	.16	.01	.34	.33	.35	.40	.45	.24	.37		.41	.74	.21
13. Interpersonal GPA (year 7, *N* = 510)	.11	.05	.15	.13	.30	.38	.36	.42	.47	.24	.28	.41		.71	.34
14. Interpersonal GPA (overall, *N* = 724)	.13	.03	.22	.14	.47	.52	.48	.50	.58	.72	.71	.74	.72		.43
15. Job performance (supervisor, year 9, *N* = 64)	.15	-.11	.27	.10	.03	.38	.24	.38	.26	.25	.38	.21	.34	.43	

Note. Uncorrected correlations are above the diagonal, corrected correlations are below the diagonal. Correlations were corrected for multivariate range restriction. Apart from the last row, correlations higher than .09 are significant at .05 level; correlations higher than .12 are significant at .01 level. For the last row, correlations higher than .25 are significant at .05 level; correlations higher than .31 are significant at .01 level.

Validity of SJT for Predicting Different Academic Performance Domains in the Long Run

Table 5 takes the multidimensionality of performance into account as it presents the relationship between the predictors and the different academic performance domains (medical versus interpersonal) across the years. Again, the values below the diagonal represent the corrected correlations between the predictors and performance. The values above the diagonal are the uncorrected correlations. Table 5 shows that the SJT was a significant and consistent predictor of interpersonal GPA in each year, whereas it was not a significant predictor of medical GPA (with the exception of the fourth year). The corrected validity of the SJT for predicting overall interpersonal GPA was .22. These results confirm that SJTs used in a high-stakes context show predictive validity for predicting interpersonal GPA in the long run. No significant increases or decreases were apparent, supporting Hypothesis 2. Results for medical GPA mirrored the declining trend of overall GPA, which is to be expected given the high correlation between medical and overall GPA (>.80). In the first year, the corrected correlation between the cognitive composite and GPA equaled .45. In the last year, this dropped to .18.

Note that care should be taken when comparing the validities of the SJT for predicting interpersonal GPA (r = .22) to those of the cognitive composite for predicting medical GPA (r = .40). The reason is that the medical GPA composite is based on a much larger number of courses per year (up to ten courses) than the interpersonal GPA composite (one or two courses). So, the medical GPA criterion is more reliable than the interpersonal GPA criterion. As it is also important to report analyses that correct for unreliability in the criterion, we computed the validity of the SJT for predicting a single interpersonal course and compared it to the validity of the cognitive composite for predicting a single medical course. To this end, we followed the procedure of Berry and Sackett (2009). Regarding the SJT, we computed its mean single-course

validity across interpersonal courses, obtaining a value of .16. To obtain an estimate of the reliability of the interpersonal course ratings, we computed the mean intercorrelation among the interpersonal courses. Next, we used this reliability estimate to correct the mean single-course validity of the SJT for unreliability in the criterion. A similar procedure was adopted for applying the attenuation correction to the mean validity of the cognitive composite for predicting a single medical course.

Results showed that there was indeed a difference in the reliability of the criteria. The mean intercorrelation among medical courses equaled .35, whereas the mean intercorrelation among interpersonal courses was .27. Using these reliability estimates, the mean unattenuated validity of the SJT for predicting a single interpersonal course equaled .31 and the mean unattenuated validity of the cognitive composite for predicting a single medical course was .44. Thus, when unreliability in the criterion was taken into account, the validity of the SJT for predicting interpersonal GPA (from .22 to .31) increased more than the validity of the cognitive composite for predicting medical GPA (from .40 to .44). Nonetheless, while correcting for unreliability reduces the difference between the cognitive composite-medical course correlation and the SJT-interpersonal course correlation, the cognitive composite-medical course correlation remains the stronger of the two.

Validity of SJT for Predicting Job Performance

The last set of hypotheses dealt with the predictive validity as well as the incremental validity of the SJT for predicting job performance. Hypothesis 3a stated that SJTs used in a high-stakes context will show validity for predicting job performance. As shown in Table 3, the corrected validity of the SJT was .27 for predicting supervisory-rated job performance. These

results support Hypothesis 3a. The cognitive composite correlated .15 with supervisor-rated job performance.

Table 3 also shows that the validity of the SJT for predicting job performance was generally higher than the validity of the SJT for predicting interpersonal GPA. However, that finding is based on samples that are not comparable (i.e., the 724 students entering medical school vs. the 64 students entering the General Practice program upon completing medical school). When we compute correlations between the SJT and interpersonal GPA for the sample of candidates (N = 64) who chose to start General Practice training and from whom job performance ratings were available, results showed that the SJT had comparable validities for predicting interpersonal GPA and job performance.

Hypothesis 3b posited the SJT used in a high-stakes context to have incremental validity over cognitive tests for predicting job performance. Results of the hierarchical regression analysis are presented in Table 6. The SJT explained 8% incremental variance in supervisory-rated job performance, thus supporting Hypothesis 3b.

Table 6. Summary of Hierarchical Regression Analyses of Predictors on Job Performance

		Job Performance (supervisor) (year 9, N = 64)		
Model	Predictors	Beta	R^2	ΔR^2
1.	Cognitive composite	.17	.02	.02
2.	Reading text	-.18	.04	.02
3.	SJT	.29*	.13	.08*

*p < .05; **p < .01. The corrected matrix served as input for the regression analysis.

Parameter estimates are for final step, not entry. Due to rounding, ΔR^2 differs .01 from the Cumulative R^2.

DISCUSSION

This investigation of the long-term validity of SJTs for predicting both academic and job performance has important applied conclusions for the use of SJTs in high-stakes selection. In addition, there are several theoretical implications for longitudinal research on selection procedures in general.

SJTs in High-stakes Selection Practice

This study contributes to filling a number of key gaps in the current literature on SJTs. First, it provides evidence of the predictive validity of an operational SJT, against a backdrop of a large literature made up of predominantly concurrent studies. This is an important result because lab research has shown that SJTs can be vulnerable to faking and coaching effects. Our study of the use of an operational SJT in a high-stakes context shows this SJT to be a valid predictor of both interpersonal academic performance (r = .22) and subsequent job performance ratings (r = .27). It should be noted that this study's validity coefficients were smaller than the meta-analytic mean r of .35 reported for concurrent studies, and larger than the meta-analytic mean of r = .18 for predictive studies.

Second, this study provides evidence that this predictive relationship applies when considering incremental validity over and above cognitive measures. In other words, this study provides confirmation of one of the primary assumptions underlying the exploration of SJTs as "alternative" predictors in high-stakes testing, namely SJTs enable prediction beyond that provided by cognitive ability. Clearly, alternative measures such as SJTs are not designed to replace the traditional cognitive predictors. Instead, they are meant to increase the coverage of skills not measured by traditional predictors.

Third, this study reinforces the importance of conceptually matching predictor and criterion constructs, in showing that the SJT predicts interpersonal performance, but not medical knowledge acquisition (i.e., medical course GPA). Similarly, the cognitive composite predicts medical knowledge acquisition, but not interpersonal performance.

Fourth, this study provides a rare look at the prediction of long-term criteria, as (interpersonal) performance in medical school was predictable from the SJT from year 1 through year 7. Accordingly, important knowledge is added to what we already know about the long-term validity of other selection procedures such as cognitive ability and personality. SJT research and practice has only begun to bourgeon in the last decade and so far it was unknown whether the validities would stand the test of time, especially in a high-stakes context.

Fifth, this study provides an even rarer look at the use of an SJT administered in the context of academic admissions as a predictor not only of academic performance, but also of both supervisor-rated job performance nine years later. Clearly, we need more studies that integrate both education and work criteria as they provide a much more comprehensive and robust view of the validity of admission/selection procedures. Such research might provide important evidence to all relevant stakeholders (e.g., students, admission systems, schools, organizations, general public) that the selection procedures used are valid for predicting both academic and job performance.

Long-term Validation of Selection Procedures

Apart from the aforementioned implications for SJTs, several broader theoretical conclusions for longitudinal selection research can also be drawn. This study shows that in assessing the validity of selection procedures such as SJTs for predicting academic performance,

relying on early grades in validation is likely to provide only a partial picture of the predictive value of the given selection procedures. Our results highlight the importance of examining validity longitudinally in educational contexts. Similarly, criteria used in validating selection procedures in work settings should capture contributions of workers not just during the initial months they spend on the job (i.e., the so-called honeymoon period) but during a longer time span (e.g., Thoresen et al., 2004) or even their entire tenure with the organization. Only then it can be expected that we will obtain a full understanding of the predictive value of selection procedures for job performance.

Next, this study was the first we are aware of to scrutinize the long-term validity of selection procedures (in this case SJTs) using both composite (overall academic and job performance) and specific criteria (different facets of academic performance). Prior longitudinal studies did not take different criterion domains into consideration. So, in this study we distinguished between what we expected in terms of longitudinal validity when we predicted an overall criterion (where we expected some components of the criterion to change over time) versus a specific component of the criterion (where, at least under some circumstances, we expected constant validity). Results generally supported change in validity of the SJT measuring interpersonal skills for overall criteria (except for the last two years) and consistent validity for a separate facet of performance. The results for the overall criterion can be explained by the "changing task" model that posits that if the makeup of the criterion changes (e.g., over time, becoming more interpersonally loaded), the predictive power of different predictors (in this case SJTs) is expected to change.

On a more general level, these results illustrate that absolute statements (e.g., “the validity of personality increases over time”) regarding the longitudinal validities of predictors

should not be made. In domains where the predictor construct is expected to be stable over time (i.e., the “changing persons” model does not hold) predicting performance over time seems to be another example of predicting performance across performance domains (e.g., task vs. contextual performance). Similar to how validities of a given predictor might change depending on the criterion construct, validities of a given predictor might increase, stay the same, or decrease depending on how the nature of the criterion changes over time. For instance, ancillary analyses showed that overall GPA correlated .26 with interpersonal GPA in the first year, whereas it correlated .61 with interpersonal GPA in the seventh and last year. Thus, our results demonstrate that one should take the criterion construct being targeted into account in longitudinal validation efforts. It makes little sense to posit in an absolute way that the validity of a given predictor will increase, stay constant, or decrease. Instead, it is better to state that the validity of a given predictor will increase, stay constant, or decrease “for predicting a given criterion construct" in the long term.

Limitations

The study has the following limitations. Like virtually all studies in the selection literature, it reflects an examination of a single testing program in a single setting. We make no grand claims of generalizability; rather we believe that it is useful to illustrate that an SJT *can* be valid when administered in a high-stakes setting (i.e., the motivational differences between an applicant setting and an incumbent setting do not per se render SJTs invalid), that an SJT can retain validity over an extended period of time, and that an SJT can predict performance both within an academic setting and in a subsequent work setting.

We note that there is broad agreement that SJTs are a measurement method that can be used to assess a variety of different constructs (Christian et al., 2010). The SJT used here focuses on the interpersonal domain. While this is a common usage of SJTs (i.e., it is the second most frequently assessed construct, after leadership, in Christian et al.'s classification of the SJT literature), similar predictive and longitudinal work in other construct domains is warranted.

Another limitation is the small sample size (N=64) for the analysis of validity against job performance criteria. We also wish N were larger, but note that we are studying the entire population of these medical school graduates moving into general practice. The rarity of studies following individuals from school entry to subsequent job performance nine years after administration of the predictor measure makes this a useful study to report, in our opinion, despite this limitation. Additional studies using this strategy are certainly needed before strong conclusions can be reached.

An important contextual feature worthy of note is that to the best of our knowledge there was no commercial test coaching industry in Belgium focusing on the SJT at the time of these cohorts (1999-2002). At that time, coaching was mostly done in high schools, and focused on the academic content of the admissions test (i.e., the knowledge tests). In more recent years, commercial coaching programs have arisen, and it will be useful to examine SJT validity under this changed context. We note that academic admissions testing is typically much more open to public scrutiny than employment testing. In most settings, those considering higher education all know well in advance that they will be asked to take a particular test as part of the application process, and a combination of this public knowledge and relatively high testing volumes makes commercial coaching viable. In contrast, job applicants may encounter an enormous array of

differing tests as they apply for various jobs, this limiting the viability of a coaching enterprise in many settings.

In sum, the study bolsters the continually growing case that SJTs can be a useful supplement to selection systems. It also provides important insights into research on the longitudinal validity of selection procedures in general. In the future, these insights should be enhanced further with additional predictive and longitudinal studies in other contexts and with SJTs focused on other constructs (e.g., leadership, knowledge and skill, personality).

REFERENCES

Ackerman, P.L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*, 3-27.

Alvares, K. M., & Hulin, C. L. (1972). Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. *Human Factors, 14*, 295-308.

Arthur Jr., W., Bennett Jr., W., Edens, P. S., and Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*, 234-245.

Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology, 66*, 1-6.

Berry, C. M., & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of college admissions system validity. *Psychological Science, 20*, 822-830.

Bray, D.W., & Howard, A. (1983). Personality and the assessment center method. In C.D. Spielberger, & J.N. Butcher (Eds.), *Advances in personality assessment*, Vol 3 (pp. 1-34), Hillsdale, NJ: Erlbaum.

Campbell, J. P., & Knapp, D. J. (Eds.). (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell, J.P., McCloy, R.A., Oppler, S.H., & Sager, C.E. (1993). A theory of performance. In N. Schmitt, & W.C. Borman (Eds.), *Personnel Selection in Organizations* (pp. 35-70). San Francisco, CA: Jossey Bass.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: stability and change. *Annual Review of Psychology, 56*, 453-484.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15*, 233-254.

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

Cullen, M.J., Sackett, P.R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14*, 142-155.

Deadrick, D. L., & Madigan, R. M. (1990). Dynamic criteria revisited: A longitudinal study of performance stability and predictive validity. *Personnel Psychology, 43*, 717-744.

Fraley, R., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review, 112*, 60-74.

Ghiselli, E. E. (1956). Dimensional problems of criteria. *Journal of Applied Psychology, 40*, 1-4.

Hinrichs, J. R. (1978). An eight-year follow-up of a management assessment center. *Journal of Applied Psychology, 63*, 596-601.

Hooper, A.C., Cullen, M.J. and Sackett, P.R. (2006). Operational threats to the use of SJTs: Faking, coaching, and retesting issues (pp. 205-232). In Weekley, J.A. and Ployhart, R.E. (Eds.), *Situational judgment tests: Theory, measurement and application*, Lawrence Erlbaum Associates, Mahwah, New Jersey.

Howard, A., & Bray, D.W. (1988). *Managerial lives in transition: Advancing and changing times*. New York: Guilford Press.

Humphreys, L. G. (1968). The fleeting nature of the prediction of college academic success. *Journal of Educational Psychology, 59*, 375-380.

Humphreys, L. G., & Taber, T. (1973). Postdiction study of the Graduate Record Examination and eight semesters of college grades. *Journal of Educational Measurement, 10,* 179-184.

Humphreys, L. G. (1960). Investigations of the simplex. *Psychometrika, 25*, 313-323.

Jansen, P.G.W., & Stoop, B.A.M. (2001). The dynamics of assessment center validity, Results of a 7-year study. *Journal of Applied Psychology, 86*, 741-753.

Keil, C.T., & Cortina, J.M. (2001). Degradation of validity over time: a test and extension of Ackerman's model. *Psychological Bulletin, 127*, 673-690.

Kuncel, N.R., Hezlett, S.A., & Ones, D.S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*, 162-181.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology* [*Special Section, Cognitive Abilities: 100 Years after Spearman (1904)], 86*, 148-161

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

Lievens, F., Ones, D.S., & Dilchert, S. (2009). Personality scale validities increase throughout medical school. *Journal of Applied Psychology, 94*, 1514-1535.

Lin, P.C., & Humphreys, L. G. (1977). Predictions of academic performance in graduate and professional school. *Applied Psychological Measurement, 1*, 249-257.

Lunneborg, C. E., & Lunneborg, P. W. (1970). Relations between aptitude changes and academic success during college. *Journal of Educational Psychology, 61*, 169-173.

MacKenzie, W.I., Ployhart, R.E., Weekley, J.A., Ehlers, C. (2010). Contextual effects on SJT responses: An examination of construct validity and mean differences across applicant and incumbent contexts. *Human Performance, 23*, 1-21.

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

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

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.

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

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement, 65*, 70-89.

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

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126*, 3-25.

Rotundo, M., & Sackett, P.R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy capturing approach. *Journal of Applied Psychology, 87*, 66-80.

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

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Goff, S. (1988). Joint relation of experience and ability with job performance: Test of three hypotheses. *Journal of Applied Psychology, 73*, 46-57.

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T., Quinn, A., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact of demographic status on admitted students. *Journal of Applied Psychology, 94*, 1479-1497.

Stewart, G. L. (1999). Trait bandwidth and stages of job performance: Assessing differential effects for conscientiousness and its subtraits. *Journal of Applied Psychology, 84*, 959-968.

Stewart, G. L., & Nandkeolyar, A. K. (2006). Adaptation and intraindividual variation in sales outcomes: Exploring the interactive effects of personality and environmental opportunity.

Personnel Psychology, 59, 307-332.

Taylor, P. J., Russ-Eft, D. F., & Chan, D. W. L. (2005) A meta-analytic review of behavior modeling training. *Journal of Applied Psychology, 90,* 692-709.

Thoresen, C. J., Bradley, J. C., Bliese, P. D., & Thoresen, J. D. (2004). The Big Five personality traits and individual job performance growth trajectories in maintenance and transitional job stages. *Journal of Applied Psychology, 89,* 835-853.

Vey, M.A., Ones, D.S., Hezlett, S.A., Kuncel, N.R., Vannelli, J.R., Briggs, K.H., & Campbell, J.P. (2003). *Relationships among college grade indices: A meta-analysis examining temporal influences.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

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

Whetzel, D.L., McDaniel, M.A., & Nguyen, N.T. (2008). Subgroup differences in situational judgement test performance: A meta-analysis. *Human Performance, 21,* 291-309.

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

CHAPTER 5

A CLOSER LOOK AT THE EFFECTS OF COMMERCIAL TEST COACHING ON COGNITIVE AND NON-COGNITIVE TESTS

In this study, we use propensity scoring to study the coaching effects associated with three types of tests (cognitive ability, knowledge tests and situational judgment tests) in a high-stakes context. In operational settings, pre-existing differences can result in non-equivalent groups. By using propensity scores, treatment-control comparisons can be made among individuals with approximately equal probabilities of having received the treatment.

All participants of the admission exams for medical and dental studies in Flanders (2008 and 2009) received a questionnaire on preparation activities. We focused on subsamples of examinees who (a) failed the initial examination in July, (b) chose to retake in August, and (c) if they participated in paid coaching, they did so between the July and August examinations. The result is a sample of 823 individuals who met these conditions for the knowledge test, 196 of whom received paid coaching. For the GMA test, 369 subjects met the criteria, 72 received paid coaching. Of the 894 individuals who met the criteria for the SJT, 218 received paid coaching.

Results show that the coached and non-coached groups differ substantially in terms of their pretest scores. People who seek paid coaching after July score lower than people who do not seek out commercial coaching after July. Second, while the coached and non-coached groups differed on a set of variables other than the pretest (i.e., the variables making up the propensity score), matching on these other variables does not substitute for also controlling for pretest differences. One might posit that using propensity scoring could replace a pretest score. In a high-stakes setting, this does not seem the case. Coaching effects are largest for the SJT (d=.50), followed by the knowledge test (d=.45) and GMA test (d=.34).

INTRODUCTION

There is longstanding interest in the question of the amenability of various types of tests used for high-stakes decisions to score increase via coaching. Given the rise of a substantial commercial test preparation industry, understanding the effects of coaching is of considerable practical interest. Our focus in this study is on the effects of participation in a commercial coaching program, in contrast to freely available preparation activities. The focus on commercial coaching reflects the concerns that coaching activities that prove to affect test scores may be differentially accessible based on candidate social status and financial resources.

Coaching proves a difficult area to study. In laboratory settings, one can readily assign examinees to coaching and non-coaching conditions; however, there are strong concerns about examinee motivation in such non-consequential lab settings. The perplexing problem is how to study coaching in settings where some are highly motivated to seek it and others are not. Thus, one methodological gap in the coaching literature is that it is difficult to make sense of the size of the coaching effects obtained in field settings. In operational settings, due to self selection there is no random assignment to treatment and control group. Pre-existing differences can thus result in non-equivalent groups. So far, current analytical approaches have not conclusively dealt with self-selection as a major obstacle to obtain accurate estimates of coaching effects in field settings.

A second substantive concern in extant test coaching research is that we do not know the size of coaching effects for non-cognitive tests such as situational judgment tests (SJTs). In recent years, SJTs have gained substantial interest in both educational and employment domains as potential supplements to traditional cognitively-oriented tests (McDaniel, Hartman, Whetzel, & Grubb, 2007; Lievens, Buyse, & Sackett, 2005; Oswald, Schmitt, Kim,

Ramsay & Gillespie, 2004; Schmitt et al., 2009). So far, coaching effects associated with SJTs as used in actual high-stakes settings have not been examined.

In this study, we extend the research on coaching effects in field settings by examining the size of coaching effects across a variety of cognitive (cognitive ability tests and knowledge tests) and non-cognitive tests (SJTs) used in an actual high-stakes setting (i.e., admission to medical college). We also extend the existing literature by using propensity scoring (Rosenbaum & Rubin, 1983, 1984) to help address the self-selection issue.

The next sections delve deeper into these two gaps in extant coaching research. Beforehand, however, we define test coaching and distinguish it from related terms such as test practice.

Prior Test Coaching Research

In a seminal paper, Messick and Jungeblut (1981) conceptualized different types of coaching interventions in terms of a continuum, ranging from practice on sample items at one extreme to intensive instruction aimed at developing ability and knowledge at the other extreme. They defined coaching as any test preparation to improve test scores falling between these two extremes, including interventions such as test familiarization, drill-and-practice with feedback, training in strategies for specific item formats and for general test taking, subject-matter-review, or skill-development exercises. Thus, for coaching effects, there has to be learning through instruction (in the form of an external intervention such as feedback from others, information sharing, tutoring, and test preparation). These definitions are in line with conceptualizations outlined by various authors (Kulik, Bangert-Drowns, & Kulik, 1984; Sackett, Burris, & Ryan, 1989).

In the past, the effects of coaching were primarily studied in relation to cognitively-oriented tests in educational settings. As an overall conclusion, large-scale reviews and meta-

analyses in educational settings (Bangert-Drowns, Kulik, & Kulik, 1983; DerSimonian & Laird, 1983; Kulik et al., 1984; Messick & Jungeblut, 1981; Slack & Porter, 1980) found that coaching produced small but practically meaningful increases in scores on cognitively-oriented tests. For instance, the meta-analysis of Becker (1990) revealed that coaching interventions raised SAT-Verbal scores by .09 *SD*s and SAT-Math scores by .16 *SD*s.

Similar results have been reported about the impact of commercial test coaching on test performance in medical education. McGaghie, Downing and Kubilius (2004) concluded in their qualitative review of 10 field studies that the utility and value of commercial test preparation courses in medicine on test performance, if any, is small. The found that five studies report small test score improvements that can be attributed directly to the commercial courses, whereas the other five studies did not reveal any test score differences between coached and uncoached individuals.

More recently, Hausknecht, Halpert, Di Paolo and Moriarty Gerrard (2007) conducted a meta-analysis and found that effects were larger when coaching was delivered between tests. While pre-test and post-test scores differed by .64 *SD* in groups receiving coaching, that figure does not separate practice effects from coaching effects, differentiate between lab and field studies, or differentiate between studies retesting with the same vs. alternate test forms. The estimate of coaching effects in operational settings using alternate forms with a coaching program of average length was .06 *SDs*, a value far more consistent with prior research.

So far, little research on coaching effects has been conducted in relation to more recent non-cognitive predictors such as SJTs. However, as SJTs become more popular in a student admission context, test preparation firms may be expected to attempt to coach people how to respond to them most effectively. Two laboratory studies of SJT coaching have been reported. Cullen, Sackett and Lievens (2006) examined two situational judgment tests. Strategies for raising scores on each test were generated, and undergraduates were trained in

the use of these strategies using a video-based training program. Results indicated that one SJT was susceptible to coaching (d = .24), while the other was not. Ramsay et al. (2003) found that a brief 10-minute coaching intervention explaining the dimensions on which the SJT would be scored produced a positive effect (d = .34).

In sum, although coaching effects have a rich research tradition in the educational and employment area, prior studies have typically focused on cognitively-oriented tests such as cognitive ability and knowledge tests. Alternative test formats such as SJTs that have recently grown in popularity have received virtually no attention. Given the interest in using SJTs in high-stakes testing it is important to extend our knowledge of coaching effects in field settings from cognitively-oriented tests to non-cognitive tests such as SJTs. In terms of substantive hypotheses, the prior literature supports coaching effects for all three types of tests examined here, and thus we hypothesize coaching effects for the knowledge test (H1), the cognitive ability test (H2), and the SJT (H3). Beyond the hypothesis of significant coaching effects, in light of the limited research in operational testing settings, we also view the estimation of the magnitude of the commercial coaching effects as an applied issue of great interest.

Approaches For Dealing with Self-Selection in Test Coaching Research

As noted above, in field settings the coachability of tests has typically been examined using a quasi-experimental design because although some individuals receive the coaching intervention while others do not, individuals have not been randomly assigned to groups (treatment vs. control) as in a true experiment. In quasi-experimental coaching designs, there are typically extraneous factors (i.e., self-selection of participants into coaching programs) that determine whether individuals receive the treatment. Prior research has revealed empirical evidence for such pre-existing individual difference correlates in self-selection

between control and coached groups. Ryan, Ployhart, Greguras, and Schmit (2006) found that self-selection was related to demographic variables (i.e., attendees of coaching programs were more likely to be female and African American) and trait-related variables (i.e., attendees tended to be lower in stress tolerance). To the extent that the assignment mechanism also correlates with the potential outcome, interpretation of treatment-control differences in quasi-experiments is confounded (Rubin, 1974).

Over the years, several approaches have been proposed to this problem of pre-existing differences (non equivalent groups) in field settings (Connelly, Sackett, & Waters, 2010). In a first approach, researchers may use an ANCOVA strategy, where one or more covariates are selected and the treatment effect is estimated after controlling for variance in the dependent variable associated with these covariates. However, if not all relevant covariates are included, this approach can over- or underestimate true treatment effects (that would be found in a true experiment).

As a second approach for resolving the problem of quasi-experimental design, researchers can select a subsample of individuals such that each individual in the treatment condition is paired with a very similar individual in the control condition (e.g., using pairs of individuals with same gender and age). Data for the control subjects that are not used are discarded and analysis is conducted with only the selected individuals. So, treatment effects are estimated among individuals who are comparable in some way. Unfortunately, such *matching* procedures become complicated as the number of variables on which subjects are matched increases.

Third, using a *pre-test in quasi-experiments* is to be commended but even in this case two threats to internal validity make the design weaker than a true experiment. First, as a pre-test is not a perfectly veridical indicator of the latent construct, some pre-existing differences between treatment and control group on the dependent variable may go unmeasured and

therefore uncontrolled. Second, pre-test post-test change comparisons do not control for potential interactions between treatment effects and aptitudes correlated with treatment assignment. For example, individuals choosing to attend a test coaching program might be more motivated in the course than would someone not otherwise attending. If course motivation is a component of coaching effectiveness, pre-/post-test change comparisons will overestimate the coaching treatment effect that would be observed in a true experimental design (where course motivation is expected to be equal in treatment and control groups). So, it is desirable to at least examine and potentially control for other covariates, even is a pre-test is available.

Propensity Scoring and Test Coaching Effects

Recently, Harder, Stuart and Anthony (2010) and Connelly et al. (2010) introduced the approach of propensity scoring to the I/O psychology community to improve the internal validity of quasi-experiments. Propensity scoring was developed as a method to model the assignment mechanism operating in quasi-experiments (Rosenbaum & Rubin, 1983, 1984). In propensity scoring, treatment assignment is predicted in a logistic regression by a selected set of covariates knowable prior to treatment assignment. For each individual in the sample, this logistic regression estimates the probability that (s)he would have received the treatment, given his/her standing on a number of covariate predictors. These probabilities are called ‘propensity scores’. By using propensity scores, treatment-control comparisons can be made among individuals with approximately equal probabilities of having received the treatment condition. For all treatment cases in the sample, a matched subset of control participants are selected for comparison based on the correspondence of their propensity score. Thus, propensity scoring is used to select statistically equated experimental and control subjects, thereby improving the internal validity of quasi-experimental research designs.

Central to propensity scoring is the process through which covariates are selected to create the propensity score. First, when covariates that relate to the treatment condition and treatment outcome are omitted, propensity score matching will produce biased estimates of treatment effects (Austin, Grootendorst, Normand, & Anderson, 2007). Second, all covariates must be "knowable" prior to receiving the treatment intervention (Rosenbaum & Rubin, 1983, 1984). These constraints ensure that any association between the covariate and the treatment assignment is not an outcome of the treatment, as such a relationship would bias treatment estimates toward zero.

In this study, we use propensity scoring to study the coaching effects associated with three types of tests (cognitive ability tests, knowledge tests, and SJTs) in a high-stakes context. To determine the effects of commercial coaching (paid coaching) on the different kind of tests used in the admission exam, propensity scores are computed, using a wide range of variables as covariates. Only variables that are not affected by the coaching activities are selected in computing the propensity score (see below).

METHOD

Sample and Procedure

This study was situated in the context of admission to medical and dental studies in Belgium. Each year, this admission exam lasts for a whole day and it is centrally administered in a large hall in Brussels. Per year, candidates have two opportunities (July and August) to take the exam. Students who do not succeed in July and who choose to retest typically do so in August. In 2008 and 2009, 67.9% of examinees failed the initial examination; of these, roughly 65% chose to retest.

All 6,773 students attending the admission exams in 2008 and 2009 received an email with a link to a web-based questionnaire. This email was sent to them approximately five

months after attending the examination. Two reminder emails were sent. A total of 3,585 candidates returned a usable questionnaire (52.9% response rate). The demographic makeup of this group was: 33.7% male and 66.3% female; 82.4% Belgians and 17.6% foreigners; 99.3% White; mean age = 18 years and 7 months. The percentage of candidates who reported attending any kind of paid coaching in the full sample was 33.6%, 29.8%, and 27.6% for the knowledge test, cognitive ability test, and the SJT respectively.

In light of the objectives of this study, we focused on subsample of examinees who (a) failed the initial examination in July, (b) chose to retake in August, and (c) if they participated in paid coaching, they did so between the July and August examinations (rather than prior to the initial July examination). This ensures that a pre-coaching and a post-coaching score are available for each examinee. The result is a sample of 823 individuals who met these conditions for the knowledge test, 196 of whom received paid coaching. For the GMA test, 369 subjects met the criteria, 72 of whom participated in paid coaching for this test. Of the 894 individuals who met the criteria for the SJT, 218 received paid coaching.

We conducted analyses to compare these three subsamples to the testing population. Results showed that percentages of passers were smaller in the three subsamples. In the testing population, 32.1% passed the admission exam. In the knowledge test subsample, 25% passed the admission exam. The percentages were 22.5% and 28.4% in the cognitive ability test sample and SJT subsamples, respectively. In addition, the three subsamples contained more Belgians as compared to their percentage in the total group that attended the admission exam. On average, 75.9% Belgians attended the exam but the three subsamples contained about 90% Belgians (90.2% in knowledge sample, 87% in GMA sample and 89.9% in SJT sample). As for gender, 63% women attended the exam, whereas the subsamples contained approximately 70% women (70.2% for knowledge and SJT subsample and 68.3% in GMA sample). As the subsamples consisted only of test-takers who took the test two times (as

compared to the population wherein some participants attended for the third of fourth time), the range in age in the subsamples is smaller than in the population. These differences found between our subsamples and the population should be taken into consideration when generalizing our results to the full population of candidates.

Measures

Background variables. The questionnaire included questions on demographic variables (sex, age, country of birth), years in high school (6 or more), high school rank (first, second, third or fourth quartile), years/hours of study in particular subject areas (sciences), parents' education level (no high school, high school, university), parents' profession (employed/unemployed; medical/dental profession or not), financial burden to pursue higher education (no burden, small burden, high burden), and medical career aspirations (general practitioner, dentist, other specialist, don't know yet).

Test coaching activities. Students indicated whether they engaged in various test coaching activities. On the basis of prior research (Messick & Jungeblut 1981; Powers & Rock, 1999; Becker, 1990), a list of thirteen possible coaching and practice activities was compiled. These activities were information/coaching sessions at high school/universities, training courses with a friend or relative, on-site training course, making homework after training, reading books, looking at websites, asking information from medical or dental students, reading official brochures/websites, completing practice tests freely provided by a third party, engaging in web-based discussion groups, and attending web-based coaching courses. Two of these coaching activities (i.e., on-site training course and web-based coaching course) were commercial (paid) coaching activities.

Students indicated their involvement in each of these thirteen activities for each of the three tests of the admission exam: knowledge test, General Mental Ability test (GMA) and

Situational Judgment Test (SJT). Specifically, students indicated whether they attended this coaching activity or not. Students also mentioned when they attended the coaching (prior to the July session or prior to the August session).

Treatment. As we wanted to examine the effect of paid coaching activities, the treatment condition was whether or not each individual had paid to attend a test coaching program. As already noted, it was also crucial that participants had attended such paid coaching programs only prior to the August session. Accordingly, there was a pre-coaching score available for these candidates (i.e., the score on the July exam). Candidates, who indicated that their coaching activities took place *prior* to the July exam, were excluded from our analyses because these candidates had logically no pre-coaching score.

Given that we were interested in the effects of the coaching activities for each test type, this study has three treatment conditions: (1) attended paid coaching for knowledge test after July; (2) attended paid coaching for GMA test after July, and (3) attended paid coaching for SJT after July. Per test type, candidates who indicated they followed paid coaching after July were labeled as the "coached group". All other candidates were labeled as the "uncoached group".

Dependent Variables

The dependent variables were candidates' scores on the knowledge test, GMA test, and SJT of the admission exam collected both at pre-coaching (test scores in July) *and* post-coaching (test scores in August). For each test, different test forms were used for each test administration. Possible differences in difficulty across forms do not confound the assessment of coaching effects, as pre-post differences among those attending coaching are compared with pre-post differences among those not attending coaching.

Knowledge test scores. The first part of the admission exam evaluated applicants' mastery of 4 basic science-related subjects (mathematics, physics, chemistry, and biology). Per subject, 10 multiple-choice questions were asked. Every question had 4 possible answers of which only one was correct.

GMA test scores. The cognitive ability test was a reasoning test that consisted of 50 multiple-choice items with 5 response alternatives per item. The problems in this test were formulated in either verbal, numerical or figural terms. Prior research demonstrated the good reliability and predictive validity of this reasoning test for medical and dental students (Minnaert, 1996). In particular, Minnaert 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.

As the GMA test in 2008 was prone to test security breaches, results for GMA are only reported for 2009. Therefore, the sample of candidates taking the GMA test in this study is smaller than the samples related to the knowledge test and SJT.

SJT scores. The third part of the admission exam was an SJT about a physician-patient interaction. The general aim of the SJT used in the admission exam was to measure skills other than cognitive ability (i.e., interpersonal and communication skills). Prior research shows the good validity of this SJT in predicting interpersonal GPA in the medical curriculum (Lievens, et al., 2005). All 30 questions of the SJT were of the multiple-choice type, with four response alternatives.

ANALYSES

Propensity Score Covariates

The background variables and other (non-paying) test coaching activities were used in creating the propensity score (i.e., predictors of treatment condition assignment). Thus, all

variables that were knowable prior to treatment assignment and theoretically relevant were included as "covariates" to be used in creating the propensity score. This resulted in 46 covariates being used to create the propensity score; these variables are listed in the appendix. Note that this set of covariates includes not only basic demographic variables but also important variables that are theoretically linked to treatment assignment (e.g., parents' profession and financial situation) or potential treatment effect moderators (e.g., following other prep activities).

Although using 46 covariates to create the propensity score represents a substantially larger set of predictor variables than typically used in regression equations, such use is less problematic in the context of creating a propensity score. Specifically, the purpose of the logistic regression creating the propensity score is *not* to make accurate estimates of population parameters of regression weights. Instead, the goal is simply to accurately model the treatment assignment mechanism within the present sample. Though some of the predictive power of the logistic regression may indeed capitalize on chance within the present sample rather than reflecting the "true" population relationship of covariates with treatment assignment, those "true" population relationships in the logistic regression are not the focus in propensity scoring. Therefore, parsimony of the regression model is less important than improving the predictive accuracy of the logistic regression.

Missing Data Treatment

In examining the dataset, many covariates to be used to create the propensity score had missing data. Such missing data present difficulties in creating the propensity score because predicted probabilities cannot be calculated for individuals with missing data on any covariate. D'Agostino and Rubin (2000) note that non-response may be a relevant variable

itself in creating the propensity score and recommend including indicators of missingness in creating the propensity scores.

Therefore, we followed a two- step process for dealing with such missing data. First, non-response indicators were created for each variable specifying whether or not a response was observed for each individual. These non-response indicators were added to the list of covariates used to create the propensity scores. Second, we imputed missing values from observed values on other variables using maximum likelihood estimation with the estimation maximization (EM) algorithm. This two step process both models any relationship of variable missingness to receiving the treatment condition by including non-response indicators in the propensity score and provides estimation of a complete dataset to use in creating the propensity score[1].

Creating the Propensity Scores

Traditionally, propensity scores have been used in either matching or stratification approaches (D'Agostino, 1998). In matching approaches, a subset of control participants are selected for comparison to treatment participants based on the correspondence of control subjects' propensity scores. Nearest-neighbor is the most straightforward matching procedure. In stratification approaches to using propensity scores, treatment-control comparisons are made within multiple groups of approximately equivalent propensity scores. Since stratification approaches result in somewhat more distant matches between treatment and control subjects (Austin, 2009) the matching approach is used in this study.

Since the effect of coaching is examined for the three parts of the admission exam, we conducted three separate analyses. The same procedure is used for each of these three analyses. The covariates listed in the Appendix, along with the missing covariate response

[1] Treatment assignment and post-treatment scores on knowledge tests, GMA and SJT were not used in imputing missing covariate data.

indicators, were entered in a logistic regression to predict whether individuals did or did not receive paid coaching.

Results of these logistic regressions provide the opportunity to examine the quality of the subsequent matching process by checking the Cox and Snell R^2 coefficient. Generally, a Cox and Snell R^2 coefficient of 0 means that there is no need to use propensity scores, as this indicates that the variables examined prove not to differ between the treated and non-treated groups. Conversely, a coefficient of 1 is indicative of a complete confound, precluding the use of propensity scores, as it is not possible to identify individuals with equal propensity for self-selection into the treated group, such that individuals who did receive the treatment could be matched with equally propensed individuals who did not receive the treatment. The logistic regressions produced a Cox and Snell R^2 of .24, .31, and .23 for the treatments concerning knowledge test, GMA test, and SJT, respectively. These results suggest that in each of the three cases, the coached and uncoached group differed substantially on the covariates included in the logistic regression. From these logistic regressions, each individual's predicted probability of receiving the coaching (for knowledge test, GMA or SJT) was retained as the propensity score.

Next, we used an SPSS macro developed by Painter (2004) to create matched pairs of control participants and treatment participants. That is, control participants were selected for comparison to treatment participants based on the correspondence of their propensity scores. The basic (nearest-neighbor) matching procedure ensures that control individuals selected are the closest possible match to the treatment individuals. However, all matches may not be close. A matching procedure may exhaust all possible control individuals with high propensity scores, forcing treatment individuals with high propensity scores to be matched with control individuals without particularly high propensity scores (though they are the closest match remaining). An adequate approach to dealing with these potentially poor

matches is to only include treatment-control pairs with closely matching propensity scores (so-called caliper matching). We applied a .20 caliper to matching on the propensity score (i.e., only treatment-control pairs with absolute difference in propensity scores less than .20 were matched). Although in caliper matching the selected sample more closely matches treatment and control, it comes with a trade-off, namely it results in a further reduction of the sample. Therefore, we present both basic matching and caliper (.20) matching results[2].

RESULTS

Reductions in Treatment-Control Differences Using Propensity Scores

We first present information on the degree to which coached and non-coached groups differed, as indexed by their propensity scores. We then show the degree to which creating samples matched on propensity scores reduces these differences. We present this information separately for the SJT, knowledge test, and GMA.

For the SJT, basic nearest neighbor matching yielded a sample of 218 coached and 218 uncoached individuals. The .20 caliper matching approach selected smaller samples of 178 each in the coached and uncoached groups.

Consistent with conventions in studies using propensity scores, we first contrasted raw versus matched coached-uncoached group differences on covariates. Such comparisons indicate how matching individuals on propensity scores reduces potentially biasing factors associated with pre-existing differences on these covariates. The left portion of Table 1 shows the ten covariates with the greatest raw coached-uncoached differences for the SJT and compares these raw differences with differences in the matched sample, as well as differences

[2] We also evaluated matching with calipers that were narrower than .20 (e.g., .10), as simulations have indicated that narrower calipers reduce treatment-control differences on covariates as well as providing more accurate treatment effect estimates (Austin, 2009). In our sample, the reduction in bias for the covariates with tighter covariates was minimal, however, and treatment effects estimated with these tighter calipers closely corresponded to those with the .20 caliper. Thus, to save space, we report and describe only those results observed with the .20 caliper.

in the groups' average propensity score. Table 1 shows a raw propensity score difference between coached and uncoached individuals of *d*=1.50. The average propensity score of individuals in the coached group is larger than the average propensity score of individuals in the uncoached group. This finding shows that the propensity score effectively discriminates between those who receive coaching for the SJT and those who do not. The matching procedure reduced coached-uncoached differences on the propensity score to *d*=.40 and .20 caliper matching further reduced the difference to *d*=-.03.

Table 1. Bias Reductions in Matching Approach to Propensity Scoring for SJT

Variable	Raw (*N*=894)	Matched Groups: Basic (*N*=436)	Matched Groups: .20 Caliper (*N*=356)
Propensity Score	1.50	.40	-.03
Covariates with Greatest Differences			
1. Website discussions about SJT	.54	.03	-.04
2. Information sessions outside school/university about SJT	.40	.17	-.05
3. Information sessions outside school/university about knowledge tests	.36	.12	-.03
4. Website discussions about knowledge tests	.32	.00	.02
5. Complete exercises about the GMA test at home	.32	.14	.08
6. Education level father	.31	.14	.09
7. Information sessions outside school/university about GMA test	.31	.11	-.09
8. Website discussion about GMA test	.30	-.03	-.09
9. Financial burden of education	-.25	-.08	.02
10. Read official brochure and website and complete exercises on SJT	.25	.01	-.05

For the knowledge test, basic nearest neighbor matching yielded a sample of 196 uncoached individuals who were selected with propensity nearest to the 196 coached

individuals. Logically, the .20 caliper matching approach selected smaller samples of 149 in the coached group and 149 in the uncoached group. Table 2 shows a raw propensity score difference between coached and uncoached individuals of *d*=1.30. Matching reduced this to .48 and .20 caliper matching further reduced the difference to -.05.

Table 2. Bias Reductions in Matching Approach to Propensity Scoring for Knowledge Test

Variable	Raw (*N*=823)	Matched Groups: Basic (*N*=392)	Matched Groups: .20 Caliper (*N*=298)
Propensity Score	1.30	.48	-.05
Covariates with Greatest Differences			
1. Information sessions outside school/university about knowledge tests	.51	.15	-.02
2. Website discussions about SJT	.40	.10	.03
3. Education level father	.34	.04	.06
4. Close relative is doctor or dentist	.34	.19	.12
5. Information sessions outside school/university about SJT	.30	.12	-.06
6. Information sessions outside school/university about GMA test	.29	.09	-.08
7. Complete exercises about the GMA test at home	.29	.12	-.07
8. Education level mother	.26	.01	-.07
9. Read books on GMA tests	.13	.05	.01
10. Financial burden of education	-.25	-.05	-.03

For the GMA test (table 3), basic nearest neighbor matching yielded a sample of 72 coached and 72 uncoached individuals. The .20 caliper matching approach selected smaller samples of 45 each in the coached and uncoached groups. Similar bias reductions were found for this test. The raw difference in propensity scores between the coached and uncoached group was *d*=1.72. Matching procedure reduced this difference to .73 and .20 caliper matching reduced the difference to *d*=-.05.

Table 3. Bias Reductions in Matching Approach to Propensity Scoring for GMA Test

Variable	Raw	Matched Groups	
		Basic	.20 Caliper
	(*N*=369)	(*N*=144)	(*N*=90)
Propensity Score	1.72	.73	-.05
Covariates with Greatest Differences			
1. Complete exercises about the GMA test at home	.39	.19	-.13
2. Website discussions about SJT	.38	.24	.19
3. Information sessions outside school/university about SJT	.34	.23	-.07
4. Website discussions about GMA test	.31	.17	.10
5. Complete exercises about knowledge tests at home	.31	.11	-.14
6. Hours of mathematics in high school	.29	-.04	.00
7. Information sessions outside school/university about knowledge tests	.29	.22	-.12
8. Information sessions outside school/university about GMA test	.29	.26	-.16
9. Read official brochure and website and complete exercises on SJT	.28	.04	-.06
10. Informal training (friend or family) about knowledge tests	.27	.10	.05

Thus for each test, coached and uncoached groups differ on a number of the variables used to construct the propensity score. Basic matching reduces these differences substantially; .20 caliper matching essentially eliminates differences between the groups. Tables 1 to 3 also show considerable reductions on the covariates with the largest coached-uncoached differences as a result of the nearest neighbor matching procedure. Across these covariates, propensity score matching yields a reduction in coached-uncoached differences.

Estimation of Test Coaching Effects

Tables 4, 5, and 6 present pre and post-test means and *SD*s for the full sample, the matched sample, and the .20 caliper matched sample for the SJT, the knowledge test and the GMA test, respectively.

Table 4. Treatment Effect Estimates Associated with SJT from Matching Approaches to Using Propensity Scores

	Post Test (August)					Pre Test (July)					Effect Estimates
	No coaching		Coaching			No coaching		Coaching			
	M	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*	Trtmnt *d*
Raw (*N*=894)	14.74	5.13	16.20	4.15	.30	12.52	5.26	11.11	4.65	-.29	.59
Matched (*N*=436)	14.77	4.91	16.20	4.15	.29	12.11	5.41	11.11	4.65	-.20	.50
.20 caliper (*N*=356)	14.72	4.90	16.10	4.29	.28	12.31	5.10	11.14	4.75	-.24	.53

Note. Diff= Treatment Mean – Control Mean; *d*'s calculated by dividing raw mean differences by σ. For the SJT σ= 4.83.

On the basis of this information, we computed six separate estimates of the coaching effect for each test. The first three are based on post-test information only. The first is a simple comparison of post test scores for the coached and uncoached groups, the second compares these groups in the propensity matched sample, and the third compares these groups in the .20 caliper sample. These first three are presented to illustrate the consequences of attempting to estimate coaching effects in the absence of pretest information. In such post-test only designs, representing the treatment effect is a standardized mean difference such as Cohen's *d*. Cohen's *d* is defined as the difference between these two raw means divided by the standard deviation of the population (in this case all attendants of the admission exam since it was implemented).

Table 5. Treatment Effect Estimates Associated with Knowledge Test from Matching Approaches to Using Propensity Scores

	Post Test (August)					Pre Test (July)					Effect Estimates
	No coaching		Coaching			No coaching		Coaching			
	M	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*	Trtmnt *d*
Raw (*N*=823)	7.81	3.41	8.66	3.21	.24	8.29	2.63	7.54	2.25	-.21	.45
Matched (*N*=392)	7.96	3.34	8.66	3.21	.19	8.46	2.68	7.54	2.25	-.26	.45
.20 caliper (*N*=298)	7.90	3.44	8.83	3.30	.26	8.42	2.81	7.69	2.18	-.20	.47

Note. Diff= Treatment Mean – Control Mean; *d*'s calculated by dividing raw mean differences by σ. For the knowledge test, σ = 3.53.

In many field settings (such as this study), researchers have access to pre-coaching scores on the dependent variable. Such designs have the advantage that they allow researchers to control for pre-existing differences between coached and uncoached groups. Hence, the second three estimates parallel the first three (i.e., comparing full sample, matched, and .20 caliper matched samples), but also incorporate pre-test information. Here coaching effects are computed as coached uncoached *d* for the post-test minus the coached-uncoached *d* for the pretest. That is, Effect = (Post-test coached – Post-test control) – (Pre-test coached – Pre-test control).

Table 6. Treatment Effect Estimates Associated with GMA Test from Matching Approaches to Using Propensity Scores

	Post Test (August)					Pre Test (July)					Effect Estimates
	No coaching		Coaching			No coaching		Coaching			
	M	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*	Trtmnt *d*
Raw (*N*=369)	27.97	5.19	28.19	5.25	.03	29.79	5.26	28.10	5.52	-.24	.27
Matched (*N*=144)	29.21	4.98	28.19	5.25	-.14	31.58	4.39	28.10	5.52	-.49	.34
.20 caliper (*N*=90)	29.00	5.41	28.40	5.45	-.08	31.34	4.37	27.82	5.91	-.50	.41

Note. Diff= Treatment Mean – Control Mean; *d*'s calculated by dividing raw mean differences by σ. For GMA test σ = 7.04.

As a summary, Table 7 presents each of the six estimates of the coaching effect for each of the three tests. A consistent pattern emerges for all three tests, namely (a) relatively similar coaching effect estimates for raw, matched, and .20 caliper matched samples within the post-test only and within the pre-post estimates, and (b) substantial differences between estimates obtained for using a post-test only strategy vs. a pre-post strategy.

Table 7. Summary table for coaching effects (d) using post test only and pre-post test design

	SJT	Knowledge test	GMA test
Post test only design			
Raw	.30	.24	.03
Matched	.29	.19	-.14
.20 caliper	.28	.26	-.08
Pre-post test design			
Raw	.59	.45	.27
Matched	.50	.45	.34
.20 caliper	.53	.47	.41

These findings in Table 7 are driven by two things. The first is that the coached and non-coached groups differ substantially in terms of their pretest scores, as shown in Tables 4 to 6. People who seek out commercial coaching after the first administration score lower than people who do not seek out commercial coaching after the first administration. This implies that dramatically different coaching estimates are obtained if one does not correct for pretest scores. For example, without correcting for the pretest score, the *d* of the GMA test is .03, whereas it is .27 when one corrects for the pretest score.

The second is that while the coached and uncoached groups differed on a set of variables other than the pretest (i.e., the variables making up the propensity score), matching on these other variables does not substitute for also controlling for pretest differences. For example, while the size of coaching effects associated with SJTs nearly doubles when a pre-

test score as available (from .30 to .59), it changes only marginally when propensity scoring is applied (e.g., from. 30 to .29 or from .59 to 50). Thus, if one did not have pretest information one might posit that using the large number of variables available to compute a propensity score might be an effective substitute. In the present setting, this premise proves incorrect: controlling for differences in propensity is not an effective substitute for controlling for pretest differences.

As the two last rows in Table 7 control for pretest scores as well as propensity scores, these rows (either nearest-neighbor matching or the more stringent .20 caliper matching) might be seen as the "best available" estimates of coaching effects in this setting. These rows show that coaching effects are largest for the SJT (*d*=.50), followed by the knowledge test (*d*=.45) and GMA test (*d*=.34).

DISCUSSION

This study provides several contributions to the coaching literature. First, our study shows the necessity to deal with the self-selection problem in coaching research in operational settings. Our results exemplify that the coached – uncoached groups are not equivalent. Generally, coached and uncoached groups might not be equivalent because they differ (a) on their standing on the construct measured by the test and/or (b) on features (other than the construct) relative to score improvement. Our results are in line with these expectations. Individuals who had lower pre-test scores were more likely to seek paid coaching afterwards. In addition, propensity scores of coached and uncoached individuals differed. As choosing commercial coaching is not a random act, it is important to use analytical approaches that control for pretest scores as well as for differences on other variables (i.e., propensity). So far, current analytical approaches have not conclusively dealt

with self-selection as a major obstacle to obtain accurate estimates of coaching effects in field settings. This might have affected the coaching effects obtained, as shown by the difference in estimates presented in Table 7 when analytical approaches that control for pretest scores and propensity are and are not employed.

Second, this study has implications regarding the analytical approaches that one might use for estimating coaching effects. Generally, it is important to state that in a quasi experiment, the assignment mechanism is per definition always unknown. So, all analytical approaches used for estimating coaching effects should always be regarded as mere attempts to deal with the unmeasured variables and self-selection problem. Hereby some approaches focus on the pretest, whereas others aim to match samples on as large as possible set of potentially relevant covariates. Our results show that in this particular setting -all else equal- one want to correct for the pretest scores, whereas the use of propensity scoring is more of an add-on. However, in other settings, exactly the opposite results might be found. Therefore, it is important to state that no general conclusions about the relative superiority of the use of pretest score over propensity scores and vice versa can be drawn. That said, we recommend that practitioners use a variety of analytical approaches. Specifically, controlling for pretest scores as well as for other variables (demographics and other coaching related activities) might bring them as close as they can get in estimating coaching effects.

Third, this is the first study with an estimate of the effects of paid/commercial coaching on SJTs in high-stakes contexts. A key finding is that the SJT is more prone to coaching effects than GMA or knowledge tests. Hereby it is worthwhile to compare these results to the results of practice effects in high-stakes settings of Lievens et al. (2005) that showed that SJT, GMA, and knowledge test scored the same in terms of practice effects. Apparently, this is not the case when people are coached to perform better on SJTs. More broadly, these results cast doubt on the potential of SJTs to be included in long-standing high

stakes test contexts. When items become known and people are coached, SJT performance can be improved. Future research is needed to ascertain whether the improvement is genuine or artificial.

Some limitations should be acknowledged. First, propensity score matching generally requires exclusion of a number of participants. Hence, propensity score matching methods are most effective when researchers have a large pool of controls to select from. Another potential drawback is that the veracity of matching on propensity scores depends largely on including all covariates that predict treatment assignment and either predict the treatment outcome or moderate the treatment effect. If such key covariates are omitted, this will result in biased treatment effects. Note, however, that these potential limitations associated with using propensity scores are not unique to propensity scoring but also present when using the covariates directly in an ANCOVA approach. In this study, we believe that conceptually all important covariates were included in our propensity score. Although individual difference variables (e.g., conscientiousness) were not included, it should be noted that our propensity scores comprised of the behavioral manifestations of these underlying traits in the form of other coaching activities.

A second set of potential limitations is related to the generalizability of our results. This study was situated in Belgium in a high-stakes educational context. The high-stakes testing program had been running for ten years. In addition, all students included in this study had prior exposure to the test (see their pre-coaching scores). They also received feedback on the test (their score on the test) when they failed the first time. Future research is needed to examine whether the same results are found when there is no prior exposure to the test.

To shed light on the cross-cultural generalizability of our results, it is interesting to compare this study's differences in propensity scores (reflecting how coached and uncoached groups differ on a number of the variables) to the differences in propensity scores that Connelly et al.

(2010) reported in their re-examination of coaching effects on the SAT. As shown in Tables 1 to 3, we obtained *d*= 1.50, 1.30, and 1.72 for the three tests, respectively. These values are only somewhat smaller than the *d*=1.95 of Connelly et al. (2010).

APPENDIX

Variables Included in Propensity Score (in this case related to the SJT test score)

Background Variables
• Gender • Age in years • Country of birth • Number of years of high school • Main course in high school • High school rank • Hours of mathematics, physics, biology, chemistry, Latin, and Greek • Education Father • Education Mother • Does father work? • Does mother work? • Is father doctor or dentist? • Is mother doctor or dentist? • Is a close relative doctor or dentist? • Is anyone in family doctor or dentist? • Financial burden of higher education • Number of attendances admission exam • Anticipated career choice if pass
Non-commercial prep activities
• Information at school • Information at university • Information outside school or university • Training at school • Informal training (friend or family) • Homework after training • Books • Ask information from friends or students • Read description of test in brochure or on official website • Complete exercises in brochure or on official website • Read other websites for information • Read web-based forums and discuss

Note. Missing value variables were also included as covariates (total number of variables used in creating the propensity score is 46)

REFERENCES

Austin, P.C. (2009). Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal, 51,* 171-184.

Austin, P.C., Grootendorst, P., Normand, S.L.T., & Anderson, G.M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine, 26,* 754-768.

Bangert-Drowns, R.L., Kulik, J.A., & Kulik, C.C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research, 53,* 571-585.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60,*373-417.

Connelly, B.S., Sackett, P.R., & Waters, S.D. (2010). *Reducing bias through propensity scoring: A study of SAT coaching.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans.

Cullen, M.J., Sackett, P.R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14,* 142-155.

D'Agostino, R.B., Jr. (1998). Tutorial in biostatistics: Propensity score method for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17,* 2265-2281.

D'Agostino, R.B., Jr., & Rubin, D.B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association, 95,* 749-759.

DerSimonian, R., & Laird, N.M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review, 53,* 1-15.

Harder, V.S., Stuart, E.A. & Anthony, J.C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15(3),* 234-249.

Hausknecht, J.P., Halpert, J.A., Di Paolo, N.T., & Moriarty Gerrard, M.O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92,* 373-385.

Kulik, J.A., Bangert-Drowns, R.L., & Kulik, C.C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*, 179-188.

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

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

McGaghie, W. C., Downing, S. D., & Kubilus, R. (2004). What is the impact of commercial test preparation courses on medical examination performance? *Teaching and Learning in Medicine, 16* , 202-211.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89,* 191-216.

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.

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

Painter, J. (2004). Propensity matching via SPSS retrieved February 10, 2009 from http://www.unc.edu/~painter/SPSSsyntax/propen.txt.

Powers, D.E., & Rock, D.A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36*, 93-118.

Ramsay, L.J., Gillespie, M.A., Kim, B.H., Schmitt, N., Oswald, F.L., Drzakowski, S.M. & Friede, A.J. (2003). *Identifying and preventing score inflation on biodata and situational judgment inventory items.* Invited Presentation to the College Board, NY, NY.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rosenbaum, P.R., & Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701.

Ryan, A.M., Ployhart, R.E., Greguras, G.J., & Schmit, M.J. (2006). Test preparation programs in selection contexts: self-selection and program effectiveness. *Personnel Psychology, 51*, 599-621.

Sackett, P.R., Burris, L.R., & Ryan A.M. (1989). Coaching and practice effects in personnel selection. In Cooper, C.L. & Robertson, I.T. (Eds.). *International review of industrial and organizational psychology*. Chicester: John Wiley.

Schmitt, N., Keeney, J., Oswald, F.L., Pleskac, T.J., Billington, A.Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-Year College Student Performance Using Cognitive and Non-cognitive Predictors and the Impact on Demographic Status of Admitted Students. *Journal of Applied Psychology, 94,* 1479-1497.

Slack, W.V., & Porter, D. (1980). The scholastic aptitude test: A critical appraisal. *Harvard Educational Review, 50,* 154- 175.

CHAPTER 6

GENERAL CONCLUSIONS AND DISCUSSION

This chapter provides a summary and a critical discussion of the main findings obtained in the empirical studies, presented in chapter 2 through chapter 5. The main research objectives stated in chapter 1 will guide this integrated overview of the results. The aims of this dissertation were to examine whether a single admission exam (consisting of a cognitive part and a non-cognitive SJT part can be used for two different majors (medical and dental studies). Next, the predictive validity of the SJT was examined for the selection of (1) dental, and (2) medical students. In the latter study the long-term predictive validity of the SJT was studied as students' job performance measures were used as criterion. The last research objective concerned the susceptibility of an SJT to coaching effects. In the first part of this final chapter, the empirical findings are briefly summarized. Next, the strengths and limitations of the present dissertation are acknowledged. Finally, directions for future research are identified and the chapter ends with practical implications for selection in higher education.

RESEARCH OVERVIEW

This doctoral dissertation started with a research-based overview of the literature on SJTs and on the use of medical and dental admission procedures worldwide, eventually leading to the main research questions investigated in this dissertation (see chapter 1). More specifically, the empirical studies presented in chapter 2 to chapter 5 addressed four main research objectives, relating to (1) the use of the same admission procedure for different majors (dental and medical students), (2) the predictive validity of the SJT for dental education, (3) the long term predictive validity of this SJT for medical education, and (4) the effects of coaching activities on SJT performance. The following briefly summarizes the findings of this dissertation in terms of these four objectives.

Research Question 1: "Can the same admission exam tests be used for different academic majors?

One aim of the present dissertation is to take a critical look at the Flemish admission exam for medical and dental studies. In Flanders, as opposed to most other countries, medical and dental students are selected by the same admission exam, with the same tests, weights, and the same cut-offs. It is known that a minority of students participate in the admission exam in order to become a dentist. The Flemish system is based on the assumption that (1) there is no significant difference between the capacities of students choosing for either of the two majors and (2) that the requirements for both majors are the same. The results discussed in chapter 2 are both striking and robust. It was found that dental students systematically score lower on the cognitive parts of the admission exam. For the SJT, results were not consistent. In some of the years, dental students obtained a higher (albeit not significant) score. On the SJT, these differences were less apparent. This study shows that students

aspiring a career in dentistry have less chance to pass the admission exam. As the cognitively "weakest" students choose dental studies, one could question the cut-off of the admission exam for these students. Furthermore, one could also question the weight which is given to the SJT. Raising this weight could increase the number of passing students aspiring dentistry. As dentistry in Flanders has a negative public image, and as a shortage of dentists already exists, the results of this study have major practical implications. This finding raises questions about using the same admission exam procedure (tests, weights, cut-offs) for two related, but obviously different majors.

Research Question 2: What is the predictive validity of the SJT in dental education?

In the past, most studies regarding the criterion-related validity of SJTs were concurrent in design and did not involve the use of SJTs in operational high-stakes settings (Christian, Edwards, & Bradley, 2010; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007). In chapter 3, a predictive validity design of the use of SJTs in an operational high-stakes setting is presented. The study in chapter 3 examined the validity of the admission procedure in Flanders for predicting grade point average (GPA) during the dental curriculum (5 years). The results of this study confirm prior findings that cognitive predictors are valuable and necessary tools in the selection of students for dental education. All cognitive tests (science related tests and the cognitive ability test) were valid predictors of GPA in three of the five years of dental education. The validity of these predictors decreased in the later clinical years of the curriculum which was an expected result since these years focus on practice. Furthermore, this study extends the positive predictive validity findings of SJTs found in medical education to dental education. The SJT used in the admission exam is developed to measure interpersonal skills, though the SJT situations are mostly medical rather than dental in nature.

This SJT has incremental validity over the cognitive predictors in year 5 of dental education. As year 5 focuses on interaction with real life patients, this is a practically relevant result that adds arguments to the discussion whether one should use non-cognitive predictors in admission to higher education.

Research Question 3: What is the long-term predictive validity of an SJT (measuring interpersonal skills) in medical education?

In prior research, the predictive validity of the Flemish admission exam was confirmed for the first years of medical education (Lievens, Buyse, & Sackett, 2005a). As the ultimate goal of the selection procedures in higher education is to select the candidates who do well as professionals (rather than to select candidates who do well as students), study 3 examined the predictive validity in the later years of medical education (which are more related to the profession) and ultimately in the profession itself. The long-term predictive validity of SJTs has never been studied in an operational high-stakes context. The study in chapter 4 shows that the SJT used in the Flemish admission exam can be a valid predictor of both interpersonal academic performance and of subsequent job performance ratings. Again, incremental validity over and above cognitive measures is found. Obviously, using the same SJT on a long-term basis may be possible. However, previous lab research has shown that SJTs can be vulnerable to faking and coaching effects. Cullen, Sackett, and Lievens (2006) found that some SJTs are susceptible to coaching. It should be mentioned that the cohorts used in this study took the admission exam in 1999-2002. At that time no commercial test coaching industry focused on the SJT. In more recent years, commercial coaching programs have arisen, and long-term predictive validity of the SJT may be scrutinized. This question was the focus of chapter 5 (RQ 4).

Research Question 4: Are SJTs susceptible to coaching effects?

Coaching effects might jeopardize the goal of the admission procedure: selecting the best students and professionals. Investigating coaching effects is complicated by the fact that people who seek coaching may be different from people who do not seek this coaching. To minimize these pre-existing group differences, the method of propensity scoring was used. Propensity scoring was initially developed as a method to model the assignment mechanism operating in quasi-experiments and was recently introduced to the I/O psychology field (Rosenbaum & Rubin, 1983, 1984; Connelly, Sackett & Waters, 2010; Harder, Stuart & Anthony, 2010). In propensity scoring, assignment to treatment or non-treatment condition is predicted by a logistic regression using a selected set of covariates knowable prior to treatment assignment. Indeed, the results in chapter 5 show the necessity to deal with these pre-existing differences. In our context, the coached group is not equivalent to the uncoached group. Candidates who had lower pre-test scores were more likely to seek coaching. These findings confirm the findings by Sackett, Burris and Ryan (1989) who note that those with lower abilities are more likely to attend a coaching program. Our study controlled for pre-test scores and for differences on multiple other variables (i.e., propensity scores). Controlling for pre-test scores as well as for other variables probably brought us as close as possible to estimating coaching effects. The key finding of this study is that the SJT was more prone to coaching than knowledge tests or GMA. This result adds knowledge to the finding of Lievens, Buyse and Sackett (2005b) who showed that SJT, GMA, and knowledge tests scored the same in terms of practice effects.

LIMITATIONS

Although the studies in this dissertation cover more than 10 years of admission exam data and use longitudinal and predictive validation design of a fairly new selection method, some limitations should be acknowledged. First, like almost all studies in the selection literature, this dissertation reflects on an admission procedure in a specific setting. The admission context this dissertation describes is rather unique in the world. Flanders is the only region that uses the same admission exam for two obviously different majors. Moreover, due to historical reasons, high school grades are not used as an additional predictor in the selection process, which is the case in most other countries. To our knowledge, other countries don't use an SJT in the selection of medical or dental studies. Hence, it should be acknowledged that no great claims of generalizability can be made. However, the studies in this dissertation prove useful in showing that a relatively new method to measure interpersonal skills can possibly be used as a selection tool in a high-stakes setting, for two different majors, that an SJT can retain validity over an extended period of time, and that an SJT can measure something over and above tests that measure cognitive abilities.

A second limitation is the small sample size in some studies. For example, for the analysis of validity against job performance criteria in study 3, the sample size is 64. Note that the entire population of these medical school graduates moving into general practice since 1999 is studied. Since medical education takes 7 years, and the practice program for general practitioners takes 2 more years, only this small group could be examined. Small sample sizes are inherent to a longitudinal approach. The same limitation applies to study 4. Few students actually follow paid coaching. Since the propensity scoring method requires exclusion of a number of participants, the sample sizes on which analyses are based, is rather small despite the initially large data set. The .20 caliper matching shows that the initial data

set should have been even larger since many of the matched couples are excluded from these analyses.

The third limitation in this dissertation applies to the specific admission exam procedure which is extensively described in chapter 1. During the 13 years (1997-2010) the admission exam was administered, many changes have occurred, either due to theoretical, practical or institutional reasons. Moreover, since every year new items are developed, it is difficult to keep the difficulty index of the admission exam constant. Thus, the different admission exams are sometimes hard to compare. In most studies in this dissertation, different cohorts are studied as one group. For example in study 2, all entering cohorts since 1997 were used to study the validity of the admission exam for dental education. In study 3, the entering cohorts of 1999 (full exam with 12 tests) until 2002 were used (shortened exam, 7 tests) to study the long term predictive validity of the SJT. To meet this limitation, admission exam scores are standardized per year, and GPA scores are standardized per year and per university. Other research showed that the admission exam instruments were comparable across years (Lievens & Sackett, 2007).

IMPLICATIONS FOR FUTURE RESEARCH

The use of SJTs in higher education is rare. Previous research and the papers in this dissertation indicate that SJTs can be useful supplements in selection procedures in higher education. As SJTs are measurement methods that can be used to assess a variety of constructs in employment settings where similar predictive and longitudinal validity coefficients are found (Christian et al., 2010) it would be interesting to examine their use for the selection of students in other majors. However, an SJT measuring interpersonal capabilities is not useful in every context. The study of Lievens et al. (2005a) indicated the

importance of matching predictor and criterion measures. Therefore, it could be crucial to study the validity of SJTs that measure important work behaviors (interpersonal and communication skills, leadership skills, teamwork, etc.) in particular education and professional settings.

Chapter 4 describes the long-term validity of SJTs in the context of general practice (GP). This is only one of the many options (specialties) medical students can choose from. Unfortunately, the data on the performance and attitudes of the other specialties in practice were not available to us. It is stated by some researchers that selecting for interpersonal relationship skills is to be recommended only when selecting GPs and psychiatrists (Arnold, 2008). Hence, it is interesting to compare the long-term validity of the SJT for GP's, psychiatrists, but also for anesthetists, periodontists, etc.

A third implication for future research relates to the incremental validity of SJTs over and above cognitive ability and personality measures. In this study, we examined only the SJT's incremental validity over and above cognitive ability. Chan and Schmitt (2002) stated that SJTs have been shown to measure stable individual difference attributes that do not completely overlap with measures of job experience, cognitive ability, and the Big Five personality traits. This gives SJTs potential incremental validity above both cognitive ability and personality. McDaniel, Powell Yost, Ludwick, Hense and Hartman (2004) examined the incremental validity of an SJT over cognitive ability and the Big Five for managerial performance level ratings across 15 competencies. The addition of the SJT raised the validity from .22 to .30. In their meta-analysis of 2007, McDaniel et al. found incremental validity of the SJT over cognitive ability and the Big Five ranging from .01 to .02. It would be interesting to broaden these results, found in a concurrent validation study, to a high-stakes setting.

Although initial evidence for the use of SJTs in personnel and student selection is encouraging, and the studies in this dissertation shed some extra light on their potential in high-stakes education settings, one underresearched issue is whether applicants can be coached in responding SJT items. The study presented in chapter 5 contributes to the literature by trying to fill the gap in this research domain. This study shows that commercial coaching programs eventually arise when it becomes clear that each year thousands of students participate. However, as this SJT is used in student selection, the question could arise whether our findings are in fact generalizable to the use of SJTs in personnel selection. Future research should investigate why and under what specific conditions SJTs are most and least coachable. In this respect, other item characteristics should be compared in terms of their validity and coachability (e.g., complexity, length, or specificity of item stems and response alternatives). It might seem that coaching effects will be less problematic in typical personnel selection settings because these are generally small scale and one-off. Applicants encounter an enormous amount of different tests in their job search. This limits the economic viability of a potential coaching industry. On the other hand, some consultancy firms are interested in finding cost-effective, efficient ways to select large samples of applicants. In business and selection settings, the interest in large-scale selection procedures (like SJTs) increases. Therefore, in these settings, coaching effects need further investigation. Equally problematic is the effect of coaching on the long-term use of SJTs. It would be useful to examine the validity of the SJT of the admission exam under this changed context. Furthermore, the effects of different presentation formats of SJTs (video, 3D) on coachability and validity need further examination.

PRACTICAL IMPLICATIONS

In September 2002, the British Medical Journal published several letters answering the question “What is a good doctor and how do you make one?” Several people, including general practitioners, specialists, nurses, patients, educators and researchers expressed their opinion. In most of the answers of health care providers concepts as “compassion, understanding, empathy, honesty, competence, commitment and humanity” appeared. Patients primarily wanted doctors who listened to them (Hurwitz, & Vass, 2002). The Flemish admission exam for medical and dental studies answers this desire by administrating an SJT that is developed to measure the interpersonal and communication skills of potential doctors and dentists of the future. At the time it was decided to install the admission exam, there was widespread agreement that a selection procedure should include both cognitive and non-cognitive measures. Earlier studies of this SJT (Lievens et al., 2005a, 2005b; Lievens, Sackett, & Buyse, 2009) and the studies in this dissertation prove that a measurement method that is designed to grasp interpersonal skills, can indeed predict future interpersonal performance and has incremental validity over and above the validity that is accounted for by cognitive measures. Hence, the choice to insert a measure of interpersonal capacities, seems to have been a good one.

One major issue of interest is the presentation format of the SJT. Lievens and Sackett (2006) showed that a video-based format of an SJT has more predictive validity than a written version. Video-based SJTs are medium-fidelity simulations (Chan & Schmitt, 1997; Weekley & Jones, 1997). The study in chapter 5 indicates that at least written SJTs are coachable. It is possible that video-based SJTs are less susceptible to coaching than their written counterparts. In the future, use of other SJT formats (virtual reality, cartoons) might provide a possible solution.

As was indicated in many previous studies and in this dissertation, SJTs can be good supplements to cognitive tests. However, they should not replace the measures of cognitive ability. In all predictive validity studies in this dissertation, the correlation between the cognitive part of the admission exam and medical GPA in medical and dental studies was higher than the correlation between the SJT and these criteria. The cognitive tests accounted for a larger part of the predicted variance, while the SJT explained an additional part in some of the latter years of education and in the job performance criterion. Therefore, a test of general mental ability and other science-related tests should remain the core part of any medical and dental admission exam.

Finally, this dissertation adds to the debate in medical and dental admission research mentioned in chapter 1. That is, there is no longer agreement on the need to incorporate non-cognitive measures in the selection of medical students. Our studies indicate that academic ability and non-cognitive attributes are positively but nonetheless minimally correlated, as opposed to the arguments of Norman (2004). Therefore, we tend to agree with researchers like Powis (2008) and Bore, Munro, and Powis (2009) who argue that medical selection should incorporate more than just measures of academic achievement.

This dissertation started with a quote, we would like to end with another one.

"Psychologists will continue to debate whether qualities such as compassion and empathy are innate or can be learned. They will continue to differentiate between acting in an empathetic manner or genuinely feeling the quality of empathy. We can be sure, though, that psychologists, and for that matter philosophers as well, will generally agree that empathy forms a crucial underpinning for competence and professionalism on the part of physicians." (Barr, 2010, p. 129)

REFERENCES

Arnold, P.C. (2008). Letters to the editor. *Medical Journal of Australia, 189,* 236.

Barr, D.A. (2010). Questioning the premedical paradigm. Enhancing Diversity in the medical profession a century after the Flexner report. John Hopkins University Press, Baltimore.

Bore, M., Munro, D., & Powis, D.A. (2009). A comprehensive model for the selection of medical students. *Medical Teacher, 31,* 1066-1072.

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

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15*, 233-254.

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

Connelly, B.S., Sackett, P.R., & Waters, S.D. (2010). *Reducing bias through propensity scoring: A study of SAT coaching.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans.

Cullen, M.J., Sackett, P.R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14,* 142-155.

Harder, V.S., Stuart, E.A. & Anthony, J.C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15(3),* 234-249.

Hurwitz, B., & Vass, A. (2002). What's a good doctor and how can you make one? British Medical Journal, 325, 667-668.

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

Lievens, F., Buyse, T., & Sackett, P.R. (2005b). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, *58*, 981-1007.

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

Lievens, F., & Sackett, P.R. (2007). Situational judgment tests in high stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology, 92,*1043-1055.

Lievens, F., Sackett, P.R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology*, *94*, 1095-1101.

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

McDaniel, M.A., Powell Yost, A., Ludwick, M.H., Hense, R.L., & Hartman, N.S. (2004). *Incremental Validity of a Situational Judgment Test*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.

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

Norman, G. (2004). The morality of medical school admissions. *Advances in Health Sciences Education, 9*, 79-82.

Powis, D. A. (2008). Selecting medical students [editorial]. *Medical Journal of Australia, 188*, 323-324.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rosenbaum, P.R., & Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

Sackett, P.R., Burris, L.R., & Ryan, A.M. (1989). Coaching and practice effects in personnel selection. In Cooper CL, Robertson IT (Eds.). *International review of industrial and organizational psychology 1989*. Chichester: John Wiley & Sons.

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

DUTCH SUMMARY

Het gebruik van Situational Judgment Tests (SJT) bij selectie in de personeelscontext is de laatste decennia enorm toegenomen. Uit onderzoek in deze context blijkt de goede voorspellende kracht van SJTs, zij verklaren extra variantie bovenop cognitieve voorspellers en bovendien worden zij door kandidaten erg positief onthaald.

Bij selectie in het hoger onderwijs baseert men zich traditioneel op cognitieve voorspellers. Steeds vaker worden succesvolle prestaties van studenten breder geformuleerd en daarom gebruikt men vaak persoonlijkheidsvragenlijsten of interviews. Het Vlaams toelatingsexamen voor Arts en Tandarts bestaat eveneens uit cognitieve en niet-cognitieve voorspellers. Anders dan in andere landen, is de niet-cognitieve proef bij dit toelatingsexamen een SJT.

Het Vlaams toelatingsexamen selecteert zowel artsen als tandartsen met dezelfde toelatingsprocedure. Onderzoeksvraag 1 bekijkt het gebruik van eenzelfde selectie-instrument voor twee verschillende studierichtingen. De resultaten tonen aan dat studenten die voor geneeskunde kiezen voor alle cognitieve proeven van het toelatingsexamen een hogere score halen dan studenten die voor tandheelkunde kiezen. Voor de SJT is dit beeld niet consistent. Het is dan ook de vraag of het selecteren van studenten voor beide opleidingen wel door middel van hetzelfde toelatingsexamen mag gebeuren.

De validiteit van SJTs werd in het verleden al meermaals aangetoond in de context van personeelsselectie. Onderzoeksvraag 2 bekijkt de validiteit van de SJT voor de opleiding tandheelkunde. De validiteit van de SJT stijgt naarmate men de opleiding tandheelkunde verder doorloopt. Naarmate de opleiding vordert, komen meer stagevakken aan bod waar ook interpersoonlijke capaciteiten een rol spelen bij het quoteren. De SJT had in dit onderzoek

enkel incrementele validiteit bovenop de cognitieve voorspellers in het laatste jaar van de opleiding.

De predictieve validiteit van de SJT werd in het verleden al aangetoond voor de opleiding geneeskunde. Onderzoeksvraag 3 richt zich voornamelijk op de voorspellende kracht van de SJT wat het beroep van arts betreft. De SJT die interpersoonlijke capaciteiten meet, voorspelt zoals verwacht de score op interpersoonlijke vakken in de opleiding geneeskunde beter dan de cognitieve predictoren. Bovendien heeft de SJT ook een hoge correlatie met functieprestatie.

Een laatste onderzoeksvraag gaat over coachingeffecten. Studenten bereiden zich jaar na jaar beter voor op het toelatingsexamen. De coachingeffecten van cognitieve proeven werden in het verleden al vaak onderzocht. In studie 4 gaan we dieper in op de coachingeffecten van SJTs. Uit de resultaten valt af te leiden dat studenten die erg lage scores halen in juli, eerder geneigd zijn om betalende coaching te volgen. Als gevolg van deze coaching halen zij in augustus hogere scores. De coaching effecten voor de SJT zijn niet te verwaarlozen. Kandidaten die getraind worden via antwoordstrategieën, halen hogere scores op de SJT.

Dit doctoraat draagt bij tot het aantonen van het belang van niet-cognitieve proeven bij een selectieprocedure in het hoger onderwijs. Een SJT is een mogelijk alternatief voor persoonlijkheidstesten en interviews. Niet-cognitieve proeven dragen iets extra bij bovenop cognitieve proeven. Echter, omdat studenten makkelijk gecoacht kunnen worden op SJTs, dient men bij het gebruik van SJTs op lange termijn zorgvuldig te werk te gaan.