MEASURING INTEGRITY FOR SELECTION INTO MEDICAL SCHOOL

Development of a Situational Judgement Test

Wendy de Leng

Measuring Integrity for Selection into Medical School
Development of a Situational Judgement Test

Wendy E. de Leng

Wendy Eline de Leng
Measuring Integrity for Selection into Medical School: Development of a Situational Judgement Test
Thesis, Erasmus University Rotterdam, the Netherlands

ISBN: 978-94-6375-662-4

Cover: Studio Slick (Rick van Driel)
Printing: Ridderprint, the Netherlands

Measuring Integrity for Selection into Medical School
Development of a Situational Judgement Test

Meten van integriteit voor de selectie van geneeskundestudenten
Ontwikkeling van een situationele beoordelingstest

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

woensdag 18 december 2019 om 15.30 uur

door

Wendy Eline de Leng
geboren te Dordrecht

Erasmus University Rotterdam

Promotiecommissie:

Promotoren:	Prof.dr.ir. A.P.N. Themmen Prof.dr. M.Ph. Born
Overige leden:	Prof.dr. M.A. Frens Prof.dr. D. van der Linden Dr. J.K. Oostrom
Copromotor:	Dr. K.M. Stegers-Jager

Contents

Chapter 1

General Introduction

"When anyone has offended you and asks you to excuse him – what ought you to do?" (Binet & Simon, 1916).

The above citation presents an item derived from one of the first standardised intelligence test, developed by Binet and Simon (1916). This item of the subscale 'Reply to an Abstract Question' indicates that Binet and Simon perceived intelligence as encompassing abilities beyond factors such as quantitative reasoning and working memory. Likewise, in the medical profession, scientific knowledge and cognitive abilities are not the only attributes that matter (Duffy, 2011). Noncognitive attributes such as personality, motivation and emotional intelligence also contribute to the provision of good health care. These noncognitive attributes are essential to the medical profession since physicians are involved in the close interaction with patients, patients' family members and colleagues. Accordingly, the Canadian Medical Education Directions for Specialists (CanMEDS) advocate that physicians should not only be assessed in their role as Medical Expert, but also in the other roles as defined in the CanMEDS framework, i.e. Communicator, Collaborator, Leader, Health Advocate, Scholar and Professional (Frank, Snell, & Sherbino, 2015).

Increased attention for noncognitive attributes in the medical profession and medical education has affected medical school admissions. Medical schools worldwide receive more applications than the number of available admission spots and are therefore presented with the task to select their students. In addition to commonly-used cognitive admission tests, admission committees search for methods to effectively and efficiently measure noncognitive attributes among large groups of applicants. Noncognitive admission tools which are used in practice include personal statements, references, personality and emotional intelligence tests, interviews, multiple mini-interviews and selection centres (Patterson et al., 2016). Recently, the search for noncognitive tools in medical school admissions has led to the introduction of the Situational Judgement Test (SJT). In medical school admissions, SJTs present applicants with challenging situations that could be encountered in medical school. These dilemma-like scenarios are followed by a number of response alternatives to the situation, which applicants have to judge on their appropriateness (Weekley & Ployhart, 2006). One noncognitive attribute that has received considerable attention in the medical profession, medical education and medical school admissions is integrity. Therefore, the aim of this thesis is to develop an SJT to measure integrity in medical school applicants. Additionally, we examine how several characteristics of the SJT influence the quality of an SJT used for medical school selection.

The noncognitive attribute central to this thesis is integrity. Integrity is considered a primary component in various conceptualisations of professionalism (Hillis & Grigg, 2015). Professionalism is a noncognitive attribute that has received substantial attention, because it is considered essential for society's trust in the medical profession (Cruess, 2006). Professionalism is defined as "the ethical and humanistic skills needed to practice medicine" (Baernstein & Fryer-Edwards, 2003) and is conceptualised in many different ways, for example in the light of the physician-patient interaction or as a set of traits typical for the profession (Cruess, Johnston, & Cruess, 2004). Brown and Ferrill (2009) established a taxonomy of professionalism distinguishing competence domains (that is professional capability), connection domains (that is interpersonal compatibility) and character domains (that is personal reliability). Integrity is considered a fundamental part of the character domain and is described using terms such as being authentic, open and honest and basing decisions and behaviours on values and principles (Brown & Ferrill, 2009). Swick (2000) conceptualised professionalism as a set of behaviours and described integrity as a core humanistic value that is crucial in the treatment of patients. In addition, professionalism and integrity are not only conceptualised by what they comprise, but also by their opposing

characteristics, that is unprofessional behaviour (Van Mook et al., 2010). In this thesis, integrity is characterised by the *presence* of sincerity, fairness and modesty, which are facets of the personality trait honesty-humility (Ashton & Lee, 2005) and by the *absence* of self-centred attitudes, thoughts and beliefs which result in antisocial behaviour (Barriga & Gibbs, 1996).

The admission tool central to this thesis is the SJT. SJTs were first used for personnel selection. The earliest examples of SJTs focused on judgement skills of soldiers during World War II and on supervisory skills in occupational settings (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; Motowidlo, Dunnette, & Carter, 1990). Nowadays, SJTs are used for the selection for many types of jobs, including insurance agents (Dalessio, 1994), secondary school teachers (Elliott, Stemler, Sternberg, Grigorenko, & Hoffman, 2011), police officers (De Meijer, Born, Van Zielst, & Van der Molen, 2010) and general aviation pilots (Hunter, 2003). Additionally, SJTs have made their entrance in educational selection contexts, introducing general measures of college performance such as the College Life Questionnaire (Sternberg & Collaborators, 2006) and the Situational Judgement Inventory developed by Oswald, Schmitt, Kim, Ramsay, and Gillespie (2004) and measures of specific student skills, for example information-seeking skills (Rosman, Mayer, & Krampen, 2015). Ultimately, the favourable psychometric properties (that is sufficient levels of criterion-related validity and low adverse impact on ethnic and socioeconomic minority applicants, see below for an overview) and the standardised applicability to large groups of applicants have also increased the use of SJTs in medical school admissions.

SJTs are extremely versatile measures that come in various shapes and forms. SJTs can differ – among other things – in their content, construct, stimulus format, response instructions, response format and scoring method (Campion, Ployhart, & MacKenzie, 2014). Previous studies demonstrated that changing any of these characteristics can alter the test's psychometric qualities. As a consequence, implementing an SJT in medical school admissions requires the careful examination of which test characteristics improve the assessment of applicants' noncognitive attributes. This thesis describes five studies that examine how several SJT characteristics influence various quality criteria of an SJT in measuring integrity among medical school applicants (see Table 1 for a schematic overview of this thesis). The SJT characteristics and quality criteria are described below. This general introduction starts with a description of prior research concerning the quality criteria of SJTs and the findings of previous studies on the influence of SJT characteristics on these criteria. Next, specific applications of SJTs to medical school admissions are described. Last, the general introduction is concluded with the outline of this thesis.

Table 1
Schematic overview of the SJT characteristics (rows) and quality criteria (columns) examined in this thesis. 'Ch.' refers to the chapter number in this thesis.

	Reliability	Criterion-related validity	Construct validity	Subgroup differences	Fakability	Applicant perceptions
Development method			Ch. 3			
Response format						Ch. 5
Scoring method	Ch. 2	Ch. 6	Ch. 2	Ch. 2	Ch. 4	
Response appropriateness		Ch. 6	Ch. 3		Ch. 4	
Response instructions						Ch. 5

Quality criteria

Reliability
In the SJT literature, the test's reliability is commonly estimated using internal consistency reliability coefficients (mostly coefficients alpha). These reliability estimates of SJTs show consistent results. A review across 39 SJTs reported reliability coefficients ranging from .43 to .94 (McDaniel et al., 2001). Another review reported a relatively low mean reliability of .46 based on 56 alpha coefficients (Catano, Brochu, & Lamerson, 2012). Low internal consistency reliabilities are often attributed to the heterogeneous test content of SJTs, which cause them to be multidimensional at the item level (Chan & Schmitt, 2005). Interestingly, even though most researchers recognise that internal consistency reliability estimates are of limited use for SJTs, most papers still report coefficients alpha. Recommended alternatives for internal consistency reliability estimates are alternative form reliability and test-retest reliability (Whetzel & McDaniel, 2009). Although these reliability estimates are less affected by the multidimensional structure of SJTs, they introduce other problems, such as the difficulty to develop equivalent, alternate forms of SJTs and the possible influence of practice effects.

Construct validity
SJTs are considered measurement methods that can be designed to assess a variety of constructs (Ployhart & MacKenzie, 2011), for example personal initiative (Bledow & Frese, 2009), integrity (De Meijer et al., 2010), emotional intelligence (Libbrecht & Lievens, 2012) and several personality dimensions (Mussel, Gatzka, & Hewig, 2018; Olaru et al., 2019; Oostrom, De Vries, & De Wit, 2019). Nonetheless, most SJTs are developed with a predominant focus on matching the test content to the criterion domain and on maximising the predictive validity of the test without aiming to measure a certain construct (Olaru et al., 2019). Consequently, the SJT literature has concentrated mainly on the criterion-related validity of SJTs and has directed relatively little attention to the constructs measured by SJTs (Schmitt & Chan, 2006). In fact, a content analysis across 136 SJTs indicated that 33% of the

SJTs did not specify the constructs measured by the test (Christian, Edwards, & Bradley, 2010). Related to the lack of clarity on the constructs measured is the construct heterogeneity of SJTs (Guenole, Chernyshenko, & Weekly, 2017), illustrated by factor analyses that often reveal multidimensional internal structures consisting of a large number of uninterpretable factors (Whetzel & McDaniel, 2009). Nevertheless, studies on construct-driven SJTs have demonstrated sufficient levels of convergent and discriminant validity and promising factor analytical results (Mussel et al., 2018; Oostrom et al., 2019). A construct-oriented approach to developing and investigating SJTs may enhance the theoretical understanding of SJTs (Christian et al., 2010), increase the generalisability of SJTs across different settings and promote a cleaner measurement of the targeted constructs (Lievens, 2017).

In contrast to the notion that SJTs are measurement methods able to assess different constructs are suggestions of general constructs that are assessed by all SJTs, for example cognitive ability, job knowledge and personality traits (McDaniel, Hartman, Whetzel, & Grubb, 2007; McDaniel & Nguyen, 2001), implicit trait policies or general domain knowledge (Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010; Motowidlo, Hooper, & Jackson, 2006), contextual knowledge (Chan & Schmitt, 2005) or tacit knowledge (Sternberg, Wagner, & Okagaki, 1993). Finally, other studies have suggested that the variance in SJTs is mainly explained by situation factors (Westring et al., 2009) or by a general judgement factor (Jackson, LoPilato, Hughes, Guenole, & Shalfrooshan, 2017). Overall, it appears that the variance in SJTs is explained by construct-specific as well as method-specific variance (Schmitt & Chan, 2006).

Criterion-related validity

The popularity of SJTs in medical school admissions is driven by its useful levels of criterion-related validity found in personnel selection (Bledow & Frese, 2009; Oostrom, Born, Serlie, & Van der Molen, 2012). A meta-analysis of SJTs used in personnel selection settings indicated a criterion-related validity for job performance of .26 across 118 validity coefficients (McDaniel et al., 2007). The sufficient levels of criterion-related validity are a likely result of the close correspondence between the test content and the multidimensional criterion domain (Christian et al., 2010). In fact, Lievens, Buyse, and Sackett (2005a) demonstrated that the criterion-related validity of an SJT is stronger when the predictor construct is linked to the criterion construct. Finally, the positive findings concerning criterion-related validity in personnel selection have also been demonstrated in educational selections settings. For instance, an SJT was able to predict self-reported student performance (r = .53) and absenteeism (r = -.27) (Oswald et al., 2004).

Related to research on criterion-related validity are studies demonstrating the incremental validity of SJTs over traditional measures of cognitive ability and personality in predicting job performance (Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; O'Connell, Hartman, McDaniel, Grubb, & Lawrence, 2007). However, as demonstrated by the meta-analysis of McDaniel et al. (2007), incremental validity coefficients of SJTs (ΔR^2 ranging between .03 and .07) appear considerably lower than their average criterion-related validity coefficient (r = .26). Incremental validity of SJTs has also been demonstrated in an educational setting. For instance, Schmitt et al. (2009) found that an SJT – together with a biodata measure – used for college admissions has incremental validity over high-school grade point average (GPA) and the score on the Scholastic Assessment Test (SAT) or American College Testing Assessment (ACT) in predicting cumulative college GPA (ΔR^2 = .03). In addition, the SJT produced larger incremental validity over the traditional predictors for noncognitive outcome measures such as absenteeism (ΔR^2 = .12) and organisational

citizenship behaviour (ΔR^2 = .20), again indicating the importance of linking the predictor and criterion constructs (Lievens et al., 2005a).

Subgroup differences

Nowadays, organisations and educational institutions aspire to select a workforce or student population that is diverse with regard to gender and ethnic and socioeconomic background, for business, social and ethical reasons (Ployhart & Holtz, 2008). However, traditional, criterion-valid selection instruments (e.g. cognitive tests) often show large score differences in favour of the majority subgroup, a problem that is labelled the diversity-validity dilemma (Cook, 2016). For instance, ethnic subgroup differences on cognitive ability tests are estimated to have effect sizes of approximately d = 1.00 to 1.20 in favour of White test takers (De Soete, Lievens, Oostrom, & Westerveld, 2013). In contrast, SJTs generally have lower adverse impact on minority applicants than traditional selection instruments. A meta-analysis indicated considerably smaller effect sizes for subgroup differences on SJTs, although White respondents still obtained better SJT scores than Black (d = .38), Hispanic (d = .24) and Asian (d = .29) respondents (Whetzel, McDaniel, & Nguyen, 2008). However, the adverse impact of SJTs on ethnic minority subgroups does not seem to be lower than the adverse impact of personality measures of conscientiousness and extraversion on these subgroups (Weekley, Ployhart, & Harold, 2004). A study in an educational context showed comparable SJT scores across ethnic subgroups (d ranging between 0.05 and 0.21, favouring the White group), whereas ethnic subgroup differences in SAT/ACT scores were much larger (d ranging between 1.01 and 1.09, all in favour of White students) (Oswald et al., 2004).

With regard to gender, women tend to obtain slightly higher SJT scores than men (d = 0.11) (Whetzel et al., 2008). A study in an educational context demonstrated a rather large SJT score difference in favour of women (d = 0.70), which contradicted the higher SAT/ACT scores in favour of men (d = 0.29). Contrary to ethnicity and gender, the adverse impact of SJTs on individuals of lower socioeconomic background has been less extensively investigated. An exception is the study of Lievens, Patterson, Corstjens, Martin, and Nicholson (2016) in the medical domain, who examined the capability of an SJT to increase medical student diversity and found that the score difference between the low and high socioeconomic subgroups was smaller on an SJT (d = 0.20) than on a cognitive ability test (d = 0.35). Overall, SJTs have the potential to reduce adverse impact, but the lower adverse impact is not a guaranteed quality of SJTs.

Fakability

The high stakes in selection settings reinforce applicants' tendency to fake, that is to intentionally distort responses during the selection process in order to create a better impression (Roulin, Krings, & Binggeli, 2016). Faking is mainly a problem of personality measures that use self-reports which have no right or wrong answers (Cook, 2016). Although some researchers have argued that faking poses no real threat to the use of personality instruments in selection (Ingold, Kleinmann, König, & Melchers, 2015; Ones, Viswesvaran, & Reiss, 1996), others have shown that faking on personality measures can affect criterion-related validity (Niessen, Meijer, & Tendeiro, 2017b) and hiring decisions (Donovan, Dwight, & Schneider, 2014). With respect to SJTs, instructed faking studies have shown that respondents who received instructions to fake obtained higher SJT scores than respondents who received instructions to respond honestly (d = 0.89) (Peeters & Lievens, 2005). Higher SJT scores in a fake condition than in an honest condition were also demonstrated in two within-subjects studies (Nguyen, Biderman, & McDaniel, 2005; Oostrom, Köbis, Ronay, & Cremers, 2017). Additionally, a study comparing the SJT scores of existing groups of

incumbents and applicants found that applicants obtained higher scores than incumbents (d = 0.88), presumably because they are more motivated to fake (Ployhart, Weekley, Holtz, & Kemp, 2003). However, a different study demonstrated that incumbents obtained higher SJT scores than applicants (d = 0.60), which was explained by incumbents having more job knowledge than applicants (Weekley et al., 2004). Even though the above mentioned findings indicate that SJTs are not immune to faking, SJTs are assumed to be less susceptible to faking than personality measures, presumably because SJTs are less transparent and more complex to fake than personality measures (Hooper, Cullen, & Sackett, 2006).

Applicant perceptions

Applicant's perceptions of an SJT should be investigated, since unfavourable perceptions of a selection procedure may have negative consequences. For instance, applicants who react negatively to the selection process may dissuade other potential applicants to apply and are more likely to file complaints (Hausknecht, Day, & Thomas, 2004). Additionally, unfavourable applicant perceptions may negatively affect test performance through reduced test-taking motivation (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997). According to the organisational justice model of Gilliland (1993), applicant perceptions of a selection procedure are influenced by factors such as perceived job-relatedness, face validity and perceived predictive validity. Therefore, like other selection tools that are strongly related to the criterion domain – such as work samples and assessment centres, SJT receive positive applicant perceptions (Hausknecht et al., 2004; Lievens, 2013; Macan, Avedon, Paese, & Smith, 1994) and may be perceived more favourably than traditional knowledge tests (Luschin-Ebengreuth, Dimai, Ithaler, Neges, & Reibnegger, 2015).

SJT characteristics

Development method

As mentioned before, most SJTs are developed with a predominant focus on matching the test content to the criterion domain. The close correspondence to the criterion domain is mostly achieved by using the critical incident (CI) technique for collecting observed incidents of human behaviour (Flanagan, 1954). SJT-developers use the CI-technique to collect anecdotes that are relevant to the criterion domain (e.g. a fellow student who violates a patient's confidentiality) by conducting interviews with Subject Matter Experts (SMEs) who have experience in the criterion domain (such as staff at a medical school). These critical incidents constitute the basis for the SJT-scenarios, which are subsequently presented to a second group of SMEs to gather possible response options to the described situations (McDaniel & Nguyen, 2001). This empirical, inductive development approach contributes to the contextualised nature of the SJT, which has a positive influence on the criterion-related validity of the test (Holtrop, Born, de Vries, & de Vries, 2014) and enhances perceptions of job-relatedness (Lievens, 2013). Although the CI-technique provides a point-to-point correspondence to the relevant context, the dominant focus on the criterion domain hinders the theoretical understanding of what constructs are measured by SJTs. In contrast, SJTs can be developed using a theoretical, deductive development approach that matches the content of the test to a pre-specified theoretical framework. For instance, the response options of the SJT can be written according to the facets or subcategories of well-established constructs. The deductive development approach may provide better insight in what is exactly measured by an SJT, thereby improving the theoretical understanding of SJTs (Christian et al., 2010), but may – as a side effect – reduce the fidelity of the SJT.

Response format

Various types of response formats exist, for example single-answer multiple choice, dual-answer multiple choice (e.g. best and worst response), rank-all and rate-all (Campion et al., 2014). Arthur et al. (2014) compared the psychometric outcomes of three SJT response formats (dual-answer multiple choice, rank-all and rate-all) and found the rate-all format to have the highest internal consistency reliability, the lowest correlation to general mental ability (GMA) and the highest correlation to personality. Higher internal consistency reliability for rate-all SJT scores is likely a result of the larger score variance in comparison to dual-answer multiple-choice and rank-all SJT scores (Ployhart & Ehrhart, 2003). Further, Arthur et al. (2014) explained the lower correlation of the rate-all SJT to GMA as a result of higher cognitive and information processing demands for dual-answer multiple-choice and rank-all formats than for rate-all formats. Ployhart and Ehrhart (2003) also compared three response formats (i.e. single-answer multiple choice, dual-answer multiple choice and rate-all) and found higher internal consistency reliability for the rate-all format, albeit no differences between response formats in criterion-related validity. Additionally, Arthur et al. (2014) demonstrated that an SJT using a dual-answer multiple-choice or rank-all response format showed larger ethnic subgroup differences than an SJT using a rate-all response format. This finding was also explained by dual-answer multiple-choice and rank-all formats having a stronger cognitive loading.

Even though rate-all response formats show higher internal consistency reliability and stronger correlations to noncognitive attributes, these formats appear more susceptible to faking than multiple choice or rank-all formats (Arthur et al., 2014). Finally, response formats have been shown to influence applicants' perceptions of an SJT. For instance, medical students perceived an SJT using a short-answer or interview response format more favourably than an SJT using a multiple-choice or rank-all response format (Neal, Oram, & Bacon, 2018). Additionally, Arthur et al. (2014) indicated more positive applicant perceptions for rate-all than for rank-all response formats. A possible explanation for the more positive perceptions of rate-all formats is that these formats allow more nuanced responses that better fit the dilemma-like nature of the SJT-scenarios.

Scoring method

In contrast to traditional cognitive ability tests, SJTs have no clear-cut right or wrong answers. The scoring key of an SJT can be developed in various ways, for example using the judgements of group of experts, based on a theory, or based on the relationship with a criterion (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). However, most SJTs are scored using an expert-based scoring key, implying that the applicant's judgements are compared to a consensus judgement obtained from an external reference group. Numerous methods exist to convert the similarity between a respondent's judgement and the consensus judgement into a score. Which method to use depends first of all on the response format of the SJT. The SJT examined in this thesis uses a rating format.

McDaniel, Psotka, Legree, Yost, and Weekley (2011) examined three scoring methods for a rating SJT that differed in how they controlled for response tendencies in the use of a rating scale, for example, a respondent only using the extreme rating scale points (a more detailed explanation of these scoring methods is provided in Chapter 2). An SJT scored using a method that controlled for response tendencies had stronger concurrent validity and showed a smaller ethnic subgroup difference than the same SJT scored using a method that did not control for response tendencies (McDaniel et al., 2011). The authors explain that individual differences in response tendencies distort the concurrent validity because these differences introduce a source of systematic error that is irrelevant to the criterion. Additionally, when

controlling for response tendencies, one also controls for potential ethnic subgroup differences in response tendencies, for example African Americans tend to use more extreme rating scale points than White and Asian Americans (Bachman, O'Malley, & Freedman-Doan, 2010). Controlling for ethnic subgroup differences in response tendencies could subsequently reduce ethnic subgroup differences in the SJT score.

In addition to the way of controlling for response tendencies, expert-based scoring keys differ in who are considered as experts. Most SJT scoring keys use a group of SMEs (e.g. supervisors or managers) as a reference group. However, some studies have advocated that the respondents themselves may comprise an adequate reference group, especially if it is not clear who should serve as SMEs or if only a small group of SMEs is available (Legree, Psotka, Tremble, & Bourne, 2005). This suggestion is supported by research demonstrating a high similarity between scoring keys based on experts and scoring keys based on novices (Hedlund et al., 2003; Motowidlo & Beier, 2010).

Response appropriateness

Regarding the response options of an SJT, a distinction can be made between generally appropriate response options (i.e. describing 'what to do' in challenging situations) and generally inappropriate response options (i.e. describing 'what not to do' in challenging situations). This distinction appears to be relevant for the psychometric outcomes of an SJT. For instance, an SJT scored on the basis of the correct identification of the worst response ('what not to do') was demonstrate to have stronger predictive validity than an SJT scored on the basis of the correct identification of the best response ('what to do') (Stemler, Aggarwal, & Nithyanand, 2016). Another study showed that an SJT score based on identifying bad responses significantly differed between novices and experts, whereas no significant differences between novices and experts were found in the SJT score based on identifying good responses (Elliott et al., 2011). Stemler et al. (2016) describe the capacity to know 'what to do' and the capacity to know 'what not to do' as two different skills distinguished by the motivation to approach success and the motivation to avoid failure, respectively.

An explanation provided for the stronger validity results of an SJT score based on what one should not do is the existence of more consensus on what is considered an inappropriate response than on what is considered an appropriate response. In other words, multiple appropriate responses exist that will lead to a positive outcome, but the chosen response depends on the particular situation and the responder's personality (Stemler et al., 2016). For example, appropriately solving a conflict with a supervisor differs between vertical and horizontal organisational structures. In contrast, inappropriate responses have a high chance to result in a negative outcome, making the appropriateness of these types of response options less ambiguous (e.g. becoming aggressive in a conflict with a supervisor is an obvious incorrect solution), and less susceptible to the influence of factors such as personal preference and style.

Response instructions

Although other types of response instructions exist (e.g. Oostrom et al. (2017)), most SJTs use one of two types of response instructions, namely knowledge or behavioural tendency instructions. Knowledge instructions require respondents to judge the appropriateness of the SJT response options in terms of what should be done, whereas behavioural tendency instructions require respondents to judge the SJT response options in terms of what their most likely response would be (Ployhart & Ehrhart, 2003). Response instructions have been demonstrated to influence the construct validity of SJTs, where knowledge instructions lead

to stronger correlations to cognitive ability and behavioural tendency instructions result in stronger correlations to personality traits (McDaniel et al., 2007). Additionally, Ployhart and Ehrhart (2003) found stronger criterion-related validity for SJTs using 'would do' instructions than 'should do' instructions, which they explained by a better alignment of the criterion (i.e. study skills and behaviours) to the 'would do' SJT than to the 'should do' SJT. Further, these authors suggested that the type of response instructions must fit the construct measured by the SJT, for example 'would do' instructions might best fit SJTs measuring personality, whereas 'should do' instructions are better suited for SJTs measuring job knowledge.

Although the construct of interest in this thesis (i.e. integrity) is related to personality, and therefore seems to correspond more strongly to 'would do' instructions, the SJTs in this thesis utilise 'should do' instructions. The choice for 'should do' instructions was motivated by the high-stakes context of medical school admissions. Response instructions have been shown to influence the fakability of SJTs, with SJTs using knowledge instructions being less affected by faking than SJTs using behavioural tendency instructions (Nguyen et al., 2005; Oostrom et al., 2017). Due to their higher susceptibility to faking, SJTs using 'would do' instructions are highly impractical in medical school admissions, because it may be assumed that applicants have a strong incentive to fake. Additionally, it appears that the differences found between knowledge and behavioural tendency instructions in the construct and criterion-related validity of an SJT are either reduced or cancelled out in high-stakes situations (Lievens, Sackett, & Buyse, 2009). Therefore, in line with Lievens et al. (2009), the SJTs investigated in this thesis used 'should do' instructions, in order to avoid a moral dilemma for applicants whether they should fake or not.

In addition, although SJTs using 'should do' instructions have stronger cognitive correlates than SJTs using 'would do' instructions, they are not without a certain degree of noncognitive correlates (McDaniel et al., 2007), due to the noncognitive content of most SJTs. Therefore, despite their cognitive correlates, SJTs using 'should do' instructions could still contribute to the predictive validity of admission procedures consisting of traditional cognitive tests.

SJTs in medical school admissions

The relevance of noncognitive attributes to the medical profession, as stated earlier, has led to an increased interest in the use of noncognitive selection tools in medical school admissions. Due to the large number of applicants, medical school admission committees worldwide search for noncognitive admission instruments that can be efficiently administered to large groups in a standardised manner. Additionally, admission committees pursue methods that widen access to medical school and increase the diversity of the medical student population. For these reasons, SJTs have gained popularity in medical school admissions. SJTs have been used for selection into postgraduate clinical training (Gardner & Dunkin, 2017; Koczwara et al., 2012; Patterson, Baron, Carr, Plint, & Lane, 2009) and for admission to undergraduate medical education (Fröhlich, Kahmann, & Kadmon, 2017; Husbands, Rodgerson, Dowell, & Patterson, 2015; Lievens, 2013; Luschin-Ebengreuth et al., 2015; Schripsema, Van Trigt, Borleffs, & Cohen-Schotanus, 2017). Recent reviews have indicated the effectiveness of SJTs as selection tools for medical school admissions (Patterson, Knight, et al., 2016; Patterson, Zibarras, & Ashworth, 2016) and have demonstrated their ability to predict long-term outcomes such as national licensure examinations (Dore, Reiter, Kreuger, & Norman, 2017) and internship and job performance (Lievens & Sackett, 2012). In general, SJTs were shown to be more valid for noncognitive

than cognitive criteria in medical school (Libbrecht, Lievens, Carette, & Côté, 2014; Lievens et al., 2005a) and better predictors of noncognitive criteria than were traditional admission tests (Ahmed, Rhydderch, & Matthews, 2012). Consequently, SJTs have incremental validity over traditional admission tests in predicting medical school performance. Moreover, an SJT for medical school admissions had less adverse impact on applicants of a low socioeconomic background than cognitive tests (Lievens et al., 2016). Thus, overall, the SJT appears to be a valuable contribution to medical school admission procedures. Nevertheless, research on how the test's characteristics may influence the performance of the SJT in the context of medical school admissions is limited. An exception is the study of Lievens and Sackett (2006) who demonstrated that an SJT for medical school admissions had higher predictive and incremental validity in a video-based format in comparison to a written format.

Outline of the thesis

The increasing use of SJTs in medical school admissions requires the investigation of how the test's characteristics influence the quality of the SJT. The studies presented in this thesis examine the influence of various test characteristics on a number of quality criteria of an SJT measuring integrity for medical school admissions (see Table 1).

The study presented in **Chapter 2** examines the influence of the scoring method of an SJT on three quality criteria: internal consistency reliability, ethnic subgroup differences and correlations with personality. This study is based on an integrity SJT developed by Husbands et al. (2015), which was translated to the Dutch using a back-translation. In line with previous studies, this study examines scoring methods that differ in how they control for response tendencies (McDaniel et al., 2011; Weng, Yang, Lievens, & McDaniel, 2018) and in which individuals comprise the reference group (Hedlund et al., 2003; Motowidlo & Beier, 2010). Additionally, the scoring methods vary in the type of central tendency statistic that is used to summarise the consensus judgement (i.e. mean, median or mode) and in the type of distance that is calculated between the respondent's judgement and the consensus judgement (i.e. absolute or squared). Crossing these four aspects results in 28 different scoring methods.

The SJT characteristics central to the study presented in **Chapter 3** are the development method and the response appropriateness and the quality criterion that is examined is the construct validity of the test. This study describes the development of a Dutch language SJT for medical school admissions measuring integrity, by combining of an empirical, inductive development approach with a theoretical, deductive development approach. The inductive approach bases the test content on empirical input collected among medical students and staff, whereas the deductive approach bases the test content on two integrity-related theoretical models. One theoretical model which relates positively to integrity is used to develop response options which describe 'what to do' in a challenging situation, while the other theoretical model which relates negatively to integrity is used to develop response options which describe 'what not to do' in a challenging situation. The construct validity of the SJT is examined by correlating the SJT scores (three score-versions, namely the total score, 'what to do' and 'what not to do') to four external integrity-related questionnaires.

The study presented in **Chapter 4** examines the fakability of the SJT that was developed in the previous chapter. Using a within-subjects design, the same SJT is administered to the same applicants twice, namely in a low-stakes and high-stakes situation. Applicants are expected to be more inclined to fake when the stakes are higher. The study examines the influence of two SJT characteristics – scoring method and response appropriateness – on the fakability of the SJT. The SJT under investigation uses a rating response format, allowing the examination of faking through the endorsement of more extreme rating points on the

rating scale. This study examines if more extreme responding is related to a larger score change between the low and high-stakes situations when using a scoring method that does not control for response tendencies. Additionally, SJT fakability was compared for SJT scores based on 'what to do' or 'what not to do' response options, because prior research found differences in respondents' inclination to exaggerate positive characteristics and de-emphasise negative characteristics (Donovan, Dwight, & Hurtz, 2003).

The quality criterion central to the study presented in **Chapter 5** are the applicant perceptions of the SJT. The influence of two SJT characteristics – response instructions and response format – is examined in a between-subjects study among medical school applicants. Applicants are presented with one of four SJT versions (i.e. should do-rating, should do-pick one, would do-rating or would do-pick one) and are asked to rate seven applicant perception items based on the procedural justice dimensions described by Gilliland (1993). Additionally, this study examines demographic subgroup differences based on gender, ethnic background and socioeconomic status in applicant perceptions of the SJT.

The study presented in **Chapter 6** examines the criterion-related validity of the SJT developed in Chapter 3 by relating the SJT score to the evaluation of students' professional behaviour (that is: sufficient / insufficient) in the first year of medical school. Although the number of students displaying unprofessional behaviour is small, unprofessionalism in medical students may cause serious harm to the medical school, fellow students or patients. Unfortunately, due to the small prevalence of students receiving an insufficient evaluation of their professional behaviour, traditional statistical techniques (such as logistic regression analysis) fail to achieve adequate classification accuracy for the low-prevalent group. Therefore, in addition to traditional statistical techniques, this study uses innovative techniques from the field of machine learning to classify unprofessional first-year medical students based on several cognitive and noncognitive admission variables, including the SJT. Machine learning is often applied to the classification of low prevalence events (e.g. diagnosing rare diseases) and provides various approaches for handling the classification of rare events.

This thesis is concluded with a General Discussion (**Chapter 7**) of the findings of these studies. Methodological and practical implications and directions for future research are provided.

Chapter 2

Scoring method of a situational judgement test: Influence on internal consistency reliability, adverse impact and correlation with personality?

This chapter has been published as:
De Leng, W.E., Stegers-Jager, K.M., Husbands, A., Dowell, J.S., Born, M.Ph., & Themmen, A.P.N. (2017). Scoring method of a Situational Judgment Test: Influence on internal consistency reliability, adverse impact and correlation with personality? *Advances for Health Sciences Education, 22*, 243-265.

Abstract

Situational Judgement Tests (SJTs) are increasingly used for medical school selection. Scoring an SJT is more complicated than scoring a knowledge test, because there are no objectively correct answers. The scoring method of an SJT may influence the construct and concurrent validity and the adverse impact with respect to non-traditional students. Previous research has compared only a small number of scoring methods and has not studied the effect of scoring method on internal consistency reliability. This study compared 28 different scoring methods for a rating SJT on internal consistency reliability, adverse impact and correlation with personality. The scoring methods varied on four aspects: the way of controlling for systematic error, and the type of reference group, distance and central tendency statistic. All scoring methods were applied to a previously validated integrity-based SJT, administered to 931 medical school applicants. Internal consistency reliability varied between .33 and .73, which is likely explained by the dependence of coefficient alpha on the total score variance. All scoring methods led to significantly higher scores for the ethnic majority than for the non-Western minorities, with effect sizes ranging from 0.48 to 0.66. Eighteen scoring methods showed a significant small positive correlation with agreeableness. Four scoring methods showed a significant small positive correlation with conscientiousness. The way of controlling for systematic error was the most influential scoring method aspect. These results suggest that the increased use of SJTs for selection into medical school must be accompanied by a thorough examination of the scoring method to be used.

Introduction

Background

Selection into medical school has been dominated by cognitive-based measures which are predictive for academic performance, but are less predictive for clinical performance (Ferguson, James, & Madeley, 2002; Salvatori, 2001). Adding noncognitive-based measures to cognitive-based measures may improve the predictive quality of a selection procedure (Kulatunga-Moruzi & Norman, 2002; Lucieer, Stegers-Jager, Rikers, & Themmen, 2016; Powis, 2015). Noncognitive-based selection instruments with good validity and reliability are essential for this purpose, because selection into medical school is highly competitive, with the number of applicants greatly exceeding the number of available places.

An upcoming noncognitive-based measure for selection into medical school is the Situational Judgement Test (SJT). An SJT presents applicants with several situations that they may encounter during the job (or at medical school), followed by a number of possible responses to that situation. Respondents are instructed to judge the appropriateness of these responses by stating what they would or should do in the described situation (Motowidlo et al., 1990; Weekley & Ployhart, 2006). Administering SJTs in work-related selection procedures has several beneficial characteristics: i) good predictive validity with regard to job performance (McDaniel et al., 2001), ii) incremental validity over and above cognitive ability and personality (Clevenger et al., 2001), iii) less adverse impact than cognitive measures (McDaniel & Nguyen, 2001), iv) higher favourability ratings by candidates than in cognitive tests (Lievens, 2013) and v) more efficient administration to large groups of applicants than other noncognitive-based instruments (e.g. assessment centres) (Motowidlo et al., 1990).

Previous studies on the use of SJTs for selection into medical school have shown that these beneficial characteristics of SJTs also apply in a medical school context (Koczwara et al., 2012; Lievens, 2013; Lievens et al., 2005a; Lievens & Sackett, 2012; Patterson et al., 2009; Patterson, Zibarras, et al., 2016; Patterson, Zibarras, Carr, Irish, & Gregory, 2011).

Despite the good qualities mentioned above, some aspects of SJTs require more research. One of these aspects is the scoring method (Whetzel & McDaniel, 2009). Scoring an SJT is more complicated than scoring a traditional knowledge test because there are no objectively correct answers, since SJTs consist of dilemmas with no clear-cut solutions (Bergman et al., 2006). Different researchers and practitioners have used different methods to convert the judgements on an SJT to a score, which has led to a large variety of scoring methods. This study will investigate the effect of these various scoring methods on three psychometric qualities (i.e. internal consistency reliability, adverse impact and correlation with personality). For this purpose, we used a previously validated integrity-based SJT (Husbands et al., 2015) for the selection of medical school applicants at a Dutch medical school.

Choice of scoring method depends on the type of scoring key and response format of an SJT. This study will focus on scoring methods for SJTs that use a rational scoring key and a Likert scale response format. A rational scoring key uses the judgements of a reference group of Subject Matter Experts (SMEs) to determine the ‘‘correct’’ answer. SMEs are individuals highly experienced in the relevant domain (Bergman et al., 2006). The Likert scale response format instructs the respondents to rate the appropriateness of each response option on a rating scale (Weekley, Ployhart, & Holtz, 2006).

Scoring methods

The scoring methods in this study differ on four aspects: the way of controlling for systematic error, the type of reference group, the type of distance and the type of central tendency statistic.

Aspect 1: controlling for systematic error

SJTs with a rational scoring key and a Likert scale response format can be scored using raw, standardised, and dichotomous consensus (McDaniel et al., 2011). Raw consensus computes the distance between the applicant's rating and the mean rating of the reference group using the raw data. Standardised consensus calculates the distance after conducting a within-person z standardisation such that each applicant has a mean of zero and a standard deviation of one across the SJT items. Dichotomous consensus divides the Likert scale in the middle. Points are awarded when an applicant's position on the Likert scale is on the same side as the reference group. Some dichotomous scoring methods increase the scoring range by applying a negative correction by subtracting points when applicants are on the other side of the Likert scale.

By standardising or dichotomising the data, McDaniel et al. (2011) attempted to control for systematic error. Systematic error in an SJT score may be caused by response tendencies or coaching in strategies on how to use the Likert scale, for example only opt for the extremes or only opt for the middle of the scale (McDaniel et al., 2011). Moreover, response tendencies are influenced by ethnic differences. For example, Black and Hispanic Americans are more inclined to use the extremes of a Likert scale than White Americans (Bachman & O'Malley, 1984; Hui & Triandis, 1989). By standardising or dichotomising the data, these cultural differences in the use of a Likert scale no longer influence the SJT score. Raw consensus does not control for systematic error.

McDaniel et al. (2011) examined the effect of these three scoring methods on the concurrent validity in two studies, using scores on a biodata scale measuring quitting tendencies and supervisory ratings of job performance as criterion. Higher concurrent validity was found for the standardised consensus and dichotomous consensus scales than for the raw consensus scale, which they explained by the removal of systematic error from the SJT score. In addition, the standardised and dichotomous consensus scales resulted in substantially smaller differences between White and Black respondents than the raw consensus scale, which they attributed to the removal of ethnic differences in the use of a Likert scale. Similarly, Legree, Kilcullen, Psotka, Putka, and Ginter (2010) found a higher concurrent validity for a standardised scale than a raw scale.

Next to using raw, standardised and dichotomous consensus, a score on an SJT with a rational scoring key and Likert scale response format can also be calculated using percent agreement (Legree et al., 2005). Percent agreement uses the endorsement ratios among the SMEs to determine the score corresponding to each rating. Percent agreement, like raw consensus, does not control for systematic error.

An example of a scoring method using percent agreement assigns two points to the Likert scale point endorsed by 50% or more of the SMEs and one point to the scale point endorsed by 25-50% of the SMEs (Chan & Schmitt, 1997). Another example assigns a score to each Likert scale point depending on the proportion of the reference group that endorsed that rating point (Lievens et al., 2015).

Aspect 2: reference group

A second aspect on which scoring methods may differ is the reference group. As stated above, a rational scoring key uses the judgements of a group of SMEs to determine the ''correct''

answer on an SJT. Most SJT scoring methods use SMEs because it is expected that they have knowledge about what behaviour is effective and ineffective in their field (Motowidlo & Beier, 2010). However, a number of SJT studies have used the group of respondents itself as a reference, a procedure called Consensus Based Measurement (CBM). Legree et al. (2005) argued that this procedure may be more appropriate for constructs for which no clear SMEs can be identified. A study on an SJT used for the US Airforce found that the mean ratings of the SMEs strongly correlated with the mean ratings of the group of respondents (Legree, 1995; Legree & Grafton, 1995). Similar results were found for an SJT measuring Tacit Knowledge of Military Leadership comparing lieutenants (i.e. SMEs) with cadets (Hedlund et al., 2003). Comparison of two SJT scoring keys based on either novices' or experts' mean effectiveness ratings found a correlation of .75 between the two keys (Motowidlo & Beier, 2010). In addition, both scoring keys resulted in scores that had similar criterion-related validity coefficients. These results were explained by novices' possession of a different, more general type of knowledge outside the specific job context. Furthermore, Lineberry, Kreiter, and Bordage (2014) stated that for script concordance tests used for assessing clinical reasoning skills, having experience does not indicate that someone is an infallible expert and that residents (i.e. novices) can outperform most panellists (i.e. SMEs). We are not aware of any previous research on the effect of using a less experienced reference group in a medical selection context.

Aspect 3: distance

A third aspect on which scoring methods may differ is the type of distance that is calculated between an applicant's rating and the overall rating of the reference group (SMEs or respondents). Some SJT studies have used the squared distance (McDaniel et al., 2011), whereas others have used the absolute distance (Legree, 1995). Squaring the distance gives more weight to ratings that deviate more from the reference group (Legree et al., 2005).

Aspect 4: central tendency statistic

A fourth aspect on which SJT scoring methods may differ is the manner of how the judgements of the reference group are summarised (i.e. central tendency statistic). Most SJT scoring methods have used the mean as a central tendency statistic, whereas some studies have used the mode (De Meijer et al., 2010; Lievens et al., 2015). Scoring methods using the mode assign points to the Likert scale point that most of the people in the reference group endorse. Besides the mean and mode, another widely used central tendency statistic is the median, which reflects the number at the central point when the data are ranked in numerical order (McCluskey & Lalkhen, 2007). To our knowledge, the median has so far never been used for scoring SJTs. For the sake of completeness, this study will include all three central tendency statistics.

Present study

The first goal of this study was to investigate the effect of scoring method on the internal consistency reliability of an SJT score. The appropriateness of internal consistency as a reliability estimate for SJT scores is often called into question (Catano et al., 2012). Internal consistency reliability estimates, such as coefficient alpha, are based on the assumption that all items measure the same latent trait on the same scale, i.e. that the same latent trait equally contributes to all item scores (Yang & Green, 2011). The multidimensional nature of SJTs violates this strict assumption resulting in an inaccurate estimate of reliability (Graham, 2006). However, the integrity-based SJT used in this study was designed to measure one dimension, which might lead to a less serious violation of the assumption of

unidimensionality. This is supported by a meta-analysis of Campion et al. (2014) that reported a mean alpha of .57 across 129 coefficients (range 0–.92). In addition, it was shown that coefficient alpha was significantly higher for SJTs that had a larger focus on one dimension. The focus of the current integrity-based SJT on one dimension may support the use of internal consistency reliability. So, given the anticipated unidimensionality of the SJT used in this study and because coefficient alpha is still commonly reported in the SJT literature, we chose it as a measure of comparison between scoring methods. To the best of our knowledge, this will be the first study to investigate the effect of different scoring methods on the internal consistency reliability.

The second goal of this study was to examine the effect of scoring method on adverse impact, by analysing the differences between Dutch and non-Western minority applicants. Adverse impact will be examined because SJTs may play an important role in promoting fairness in medical school selection, since SJT scores potentially demonstrate lower ethnic subgroup differences than cognitive ability test scores. On cognitive ability tests, White test takers have been shown to score approximately one standard deviation higher than non-White test takers (De Soete et al., 2013). A meta-analysis on ethnic subgroup differences across 32 SJTs – mainly originating from the US – showed that White test takers score approximately 0.38 standard deviation higher than Black test takers, 0.24 standard deviation higher than Hispanic test takers and 0.29 standard deviation higher than Asian test takers (Whetzel et al., 2008). A Dutch study also found that the ethnic subgroup difference in an integrity SJT score ($d = 0.38$) was lower than in a cognitive ability test score ($d = 0.48$) (De Meijer et al., 2010). Selection on only cognitive ability test scores might lead to the rejection of more ethnic minority applicants than ethnic majority applicants, whereas selection on SJT scores may increase the admission rate among ethnic minorities, resulting in a more culturally diverse medical student population. To promote the expected positive influence of an SJT on fairness, it is crucial to investigate the potential influence of scoring method on adverse impact. In line with the findings of McDaniel et al. (2011), we expect that scoring methods controlling for systematic error (i.e. standardised and dichotomous consensus) will lead to smaller ethnic differences than scoring methods that do not (i.e. raw consensus and percent agreement). The other scoring method aspects (i.e. type of reference group, distance and central tendency statistic) have not been studied in combination with adverse impact before.

The third goal of this study was to investigate the effect of scoring method on the correlation between the SJT score and three of the Big Five personality traits. The Big Five describes someone's personality using five broad dimensions: neuroticism (i.e. emotional instability), extraversion (i.e. outgoing and energetic), openness to experience (i.e. intellectual curiosity), agreeableness (i.e. altruistic and compassionate) and conscientiousness (i.e. organized and persistent) (Costa & McCrae, 1992). The correlation with the Big Five was examined because three of the five dimensions (i.e. conscientiousness, emotional stability and agreeableness) have been shown to moderately and positively correlate with SJT scores (McDaniel et al., 2007) and integrity test scores (Marcus, Lee, & Ashton, 2007). Moreover, the validity and reliability of the scores on the Big Five measure used in this study (i.e. NEO-PI-R (Costa & McCrae, 1992)) has repeatedly been demonstrated (Costa & McCrae, 2008), including in samples of adolescents (De Fruyt, Mervielde, Hoekstra, & Rolland, 2000). It is therefore expected that the integrity-based SJT will be correlated to these three Big Five dimensions and that the resulting correlation coefficients will provide a good measure of comparison between the scoring methods. We hypothesise that scoring methods that control for systematic error will lead to higher correlation coefficients, because the influence of response tendencies regarding the use of Likert scales is removed from the SJT score (Legree et al., 2010; McDaniel et al., 2011). We are unaware

of any previous studies that have investigated the effect of type reference group, distance and central tendency statistic on the correlation of an SJT score with personality.

Methods

Procedure
The SJT was administered during the selection procedure for the Erasmus MC Medical School in 2014 and 2015 (N = 1025). The administration was solely for research purposes and participation was voluntarily. The Erasmus MC Medical School selects students on their participation in extracurricular activities, their performance on five cognitive tests during three on-site testing days (Urlings-Strop, Stijnen, Themmen, & Splinter, 2009) and their pre-university Grade Point Average (GPA). The administration of the SJT was conducted during the on-site testing days, using paper-and-pencil. An additional questionnaire was administered regarding applicants' demographic characteristics. A personality questionnaire was administered online when applicants registered for the selection procedure. The applicants were informed that the SJT and questionnaires were administered solely for research purposes and that their answers would not influence the outcome of the selection procedure. Participation was voluntarily.

Measures

Integrity-based Situational Judgement Test
The integrity-based SJT used in this study was developed in the United Kingdom (UK) (Husbands et al., 2015). The authors translated this SJT to Dutch. This translation was validated using the back translation procedure described by Brislin (1970). The back translation was conducted by an independent commercial translation office. The authors discussed and made appropriate changes to the translated version.

The SJT consisted of ten scenarios describing problematic situations that could occur during medical school. Each scenario was followed by five response options. The respondents had to judge the appropriateness of each response option on a four-point Likert scale (1: *Very inappropriate* - 4: *Very appropriate*) in terms of what should be done given the situation (i.e. knowledge-based instructions (Ployhart & Ehrhart, 2003). An example of an SJT item is presented in Appendix 2A.

A rational scoring key for this SJT was developed based on the judgements of 16 SMEs (75 % female). The mean age of this group was 40.8 years (*SD* = 11.1). The SMEs were individuals involved in teaching professionalism in the medical curriculum. Two of the SMEs were medical doctors. The mean number of years of experience with professionalism in the medical curriculum of this group was 6.4 (SD = 5.9). All SMEs were native Dutch. The intraclass correlation coefficient (ICC) among the SMEs was .65, indicating a moderate agreement (two-way mixed model, absolute agreement).

Demographics
An applicant was considered a non-Western minority when one of his/her parents was born outside Europe or North-America (Statistics Netherlands; www.cbs.nl).

The socioeconomic status of an applicant was determined by the level of education of his/her parents. A division was made between first-generation and non-first-generation university students. First-generation university students were defined as students whose

parents did not attend university (either a research university or a university of applied science).

Personality questionnaire

In 2014, the Dutch version of the NEO-PI-R was administered to assess the applicants' standing on the Big Five personality traits (Costa & McCrae, 1992; Hoekstra, Ormel, & De Fruyt, 1996). The questionnaire consisted of 240 statements that applicants had to judge on a five-point Likert scale (1: *Strongly disagree* - 5: *Strongly agree*). The five personality subscales demonstrated good internal consistency reliabilities (coefficient alpha): .92 for neuroticism, .87 for extraversion, .85 for openness, .87 for agreeableness and .88 for conscientiousness. Due to the length of the questionnaire, the NEO-PI-R was not administered in 2015.

Scoring methods

In preparation for this study we combined the four aspects on which scoring methods can differ; this yielded 28 scoring methods to be tested (Figure 1). These scoring methods followed the categorisation into raw, standardised and dichotomous consensus scoring methods as proposed by McDaniel et al. (2011).

Within each of the raw and standardised scoring methods, the distance (absolute or squared) was calculated between the applicant's rating and the overall rating of the reference group on the Likert scale. The reference group was either made up of the 16 SMEs or of the group of respondents itself. The overall rating of this reference group was reflected by either the mean, median or mode.

In addition to the raw and standardised consensus scoring methods, the dichotomous consensus scoring method was applied. The reference group consisted of either the SMEs or the group of respondents itself. Another variation was applied by either assigning zero points to or subtracting one point from applicants whose rating was located on the opposite side of the Likert scale than the reference group.

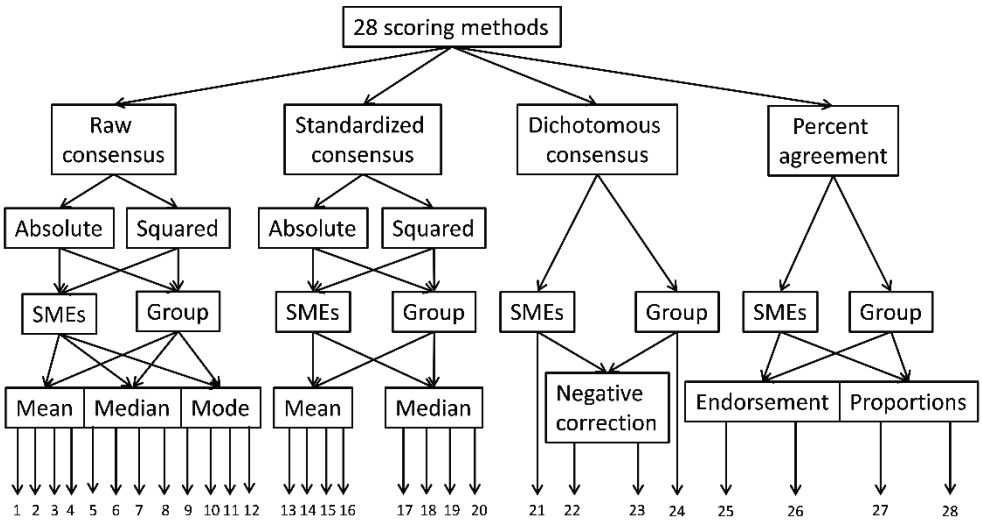

Figure 1. Schematic representation of the 28 scoring methods. SMEs = Subject Matter Experts

The 24 scoring methods based on either raw, standardised or dichotomous consensus were complemented with four scoring methods based on percent agreement (Legree et al., 2005). These scoring methods used either the 25-50% endorsement rule used by Chan and Schmitt (1997) or assigned a score to each Likert scale point corresponding to the proportion of subjects in the reference group who endorsed that point (Lievens et al., 2015). The reference group consisted of either the SMEs or the respondents.

The correlations between the 28 scoring methods are presented in Appendix 2B. Although some correlation coefficients indicated a large overlap between the scoring methods (i.e. within the raw consensus scoring method set), other scoring methods showed less overlap (i.e. between the raw and dichotomous scoring method sets).

To our knowledge, of half of these scoring methods no results have been published in the context of application to an SJT (i.e. scoring methods using the median, scoring methods calculating the distance from the group mode, dichotomous scoring methods using the SMEs, percent agreement scoring methods using the endorsement rate of the group and the proportions of the SMEs).

Statistical analysis

Both SPSS (IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.) and R (Version 3.1.0) were used to convert the judgements on the SJT to a score, using the different scoring methods. The raw and standardised consensus scoring methods that used the group of respondents itself as a reference were conducted using a leave-one-out method (Hastie, Tibshirani, & Friedman, 2009). This method removes the applicant whose score needs to be calculated from the dataset, and calculates the summary statistic across the remaining group members. The distance between the applicant and the remaining group members composes the applicant's score.

Coefficient alpha was used as an estimate of internal consistency reliability (Cronbach, 1951). Independent t-tests were used to examine the 28 different SJT scores on disparities between first-generation and non-first-generation university applicants and between Dutch and non-Western minority applicants. The effect sizes of the social and ethnic disparities were reflected by Cohen's *d* (Cohen, 1988). A stricter alpha level ($\alpha = .001$) was used because of the large number of comparisons.

For each scoring method, Pearson correlations were used to determine the correlation between the SJT score and the three Big Five personality traits for which we expected a correlation.

General linear models were used to examine which scoring method aspects significantly influenced the outcome measures (i.e. coefficient alpha, effect size and correlation coefficient). For each outcome measure, four general linear models were tested, namely one model for each scoring method aspect. The four aspects were tested in separate models because the small number of data points (i.e. 28) did not allow entering all four aspects in one model. The effect sizes were corrected for the reliability of the scoring method by dividing Cohen's *d* by coefficient alpha, since low reliability may obscure subgroup differences (Lievens, Peeters, & Schollaert, 2008).

Results

Participants

Nine-hundred thirty-one medical school applicants responded (response rate = 90.8 %). The demographic characteristics of this sample are depicted in Table 1. The two cohorts (2014

and 2015) were similar with regard to gender, age and ethnicity. Cohort 2015 consisted of significantly more first-generation students than cohort 2014, but the size of this effect was small ($X^2(1) = 6.02$, $p = .014$, $\phi = .08$). Personality data were obtained from 73.3 % of the participants from cohort 2014. SJT scores did not significantly differ between respondents and non-respondents to the personality questionnaire.

Table 1
Demographic characteristics of the participants in this study for each cohort.

	2014 (N = 521)	2015 (N = 410)
Gender (% female)	64.1	62.7
Age (Mean (SD))	19.1 (1.9)	19.2 (1.9)
Ethnicity		
% Dutch	58.3	57.2
% non-Western minority	31.3	32.2
% Western minority	10.4	10.6
SES (% First-generation university students)	24.0	31.6

Note. SD = Standard deviation SES = Socioeconomic status

Internal consistency reliability
Coefficient alpha varied from .33 to .73 depending on the scoring method (Table 2). The lowest coefficient alpha was found for the scoring method that calculated the absolute distance from the mean of the group of respondents itself using standardised consensus. The highest coefficient alpha was found for the scoring method that calculated the absolute distance from the mean of the group of respondents itself using raw consensus.

For the general linear models with coefficient alpha as dependent variable, the way of controlling for systematic error was the only significant factor with a very large effect size, $F(3, 24) = 40.05$, $p < .001$, $\eta^2 = .83$. Raw consensus led to a significantly higher coefficient alpha than the other three methods of controlling for systematic error. In addition, standardised consensus and percent agreement yielded a significantly higher coefficient alpha than dichotomous consensus.

Adverse impact
All scoring methods led to significantly higher scores for the Dutch majority than for the non-Western minorities (Table 3). The effect sizes (d) of these differences ranged from 0.48 to 0.66 (medium effect). The largest differences were found for the scoring methods that calculated the absolute distance from the SME median using standardised consensus. The smallest ethnic difference was observed for all scoring methods that used dichotomous consensus.

For the general linear models with the corrected effect size as dependent variable, the way of controlling for systematic error was again the only significant factor with a very large effect size, $F(3, 24) = 15.54$, $p < .001$, $\eta^2 = .66$. Raw consensus led to smaller corrected effect sizes than standardised and dichotomous consensus, but not percent agreement.

None of the scoring methods led to significant differences between first-generation university applicants and non-first-generation university applicants (data available upon request). Due to the lack of significant differences, no general linear models were tested.

Table 2
Descriptive statistics and internal consistency reliability (alpha coefficient, α) for the 28 rate-SJT scoring methods.

Scoring method	M (SD)	Min. - Max.	α
Raw consensus			
1. Abs. dist. - SME mean	34.32 (6.02)	20.01 - 64.99	.67
2. Abs. dist. SME median	33.11 (6.61)	13.50 - 66.50	.56
3. Abs. dist. SME mode	32.95 (6.52)	14.50 - 65.50	.55
4. Sqr. dist. SME mean	36.25 (12.50)	11.48 - 107.72	.67
5. Sqr. dist. SME median	42.44 (13.27)	12.75 - 122.75	.61
6. Sqr. dist. - SME mode	41.81 (13.18)	13.25 - 121.25	.60
7. Abs. dist. - Group mean	31.26 (6.31)	16.32 - 63.09	.73
8. Abs. dist. - Group median	28.93 (7.00)	11 - 63	.61
9. Abs. dist. - Group mode	29.07 (6.99)	11 - 63	.59
10. Sqr. dist. - Group mean	30.35 (11.56)	8.47 - 100.35	.73
11. Sqr. dist. - Group median	35.67 (12.85)	11 - 113	.65
12. Sqr. dist. - Group mode	36.28 (13.01)	11 - 115	.63
Standardised consensus			
13. Abs. dist. - SME mean	32.86 (4.63)	21.24 - 51.67	.44
14. Abs. dist. - SME median	33.52 (4.68)	19.09 - 51.54	.41
15. Sqr. dist. - SME mean	34.46 (9.57)	14.31 - 34.46	.49
16. Sqr. dist. - SME median	36.29 (9.61)	13.47 - 79.99	.45
17. Abs. dist. - Group mean	30.42 (3.91)	20.99 - 50.67	.33
18. Abs. dist. - Group median	29.91 (4.57)	18.27 - 51.00	.43
19. Sqr. dist. - Group mean	29.11 (7.77)	13.58 - 74.24	.45
20. Sqr. dist. - Group median	30.08 (8.89)	12.63 - 79.44	.51
Dichotomous consensus			
21. SME as ref.	34.34 (3.55)	21 - 44	.34
22. SME as ref. – neg. corr.	18.78 (7.04)	-8 - 38	.34
23. Group as ref.	37.56 (3.59)	22 - 47	.34
24. Group as ref. – neg. corr.	25.21 (7.11)	-6 - 44	.34
Percent Agreement			
25. Endorsement rate - SME	54.23 (7.32)	29 - 74	.49
26. Endorsement rate - Group	47.53 (5.17)	26 - 60	.46
27. Proportions - SME	19.39 (2.16)	11.18 - 25.59	.54
28. Proportions - Group	18.84 (1.55)	11.34 - 22.63	.58

Note. M = Mean SD = Standard deviation SME = Subject Matter Expert Min. = Minimum Max. = Maximum Abs. = Absolute Sqr. = Squared dist. = distance ref. = reference neg. = negative corr. = correction

Table 3
Results of independent t-tests for Dutch (N = 490) vs. non-Western differences (N = 269) in SJT scores generated by 28 scoring methods.

Scoring method	Dutch	Non-Western	*d*
Raw consensus			
1. Abs. dist. - SME mean	32.88 (5.38)	36.51 (6.50)	0.61
2. Abs. dist. SME median	31.47 (5.92)	35.58 (7.05)	0.63
3. Abs. dist. SME mode	31.34 (5.82)	35.37 (6.95)	0.63
4. Sqr. dist. SME mean	33.28 (10.86)	40.82 (13.98)	0.60
5. Sqr. dist. SME median	39.24 (11.50)	47.39 (14.84)	0.61
6. Sqr. dist. - SME mode	38.70 (11.37)	46.67 (14.74)	0.61
7. Abs. dist. - Group mean	29.95 (5.61)	33.16 (7.03)	0.50
8. Abs. dist. - Group median	27.29 (6.27)	31.31 (7.49)	0.58
9. Abs. dist. - Group mode	27.37 (6.21)	31.51 (7.48)	0.60
10. Sqr. dist. - Group mean	27.94 (9.88)	33.96 (13.37)	0.51
11. Sqr. dist. - Group median	32.66 (11.06)	40.02 (14.35)	0.57
12. Sqr. dist. - Group mode	33.13 (11.10)	40.86 (14.59)	0.60
Standardised consensus			
13. Abs. dist. - SME mean	31.69 (4.23)	34.52 (4.43)	0.65
14. Abs. dist. - SME median	32.30 (4.25)	35.22 (4.60)	0.66
15. Sqr. dist. - SME mean	32.07 (8.52)	37.80 (9.36)	0.64
16. Sqr. dist. - SME median	33.88 (8.51)	39.69 (9.56)	0.64
17. Abs. dist. - Group mean	29.53 (3.63)	31.55 (3.72)	0.55
18. Abs. dist. - Group median	28.83 (4.25)	31.30 (4.34)	0.58
19. Sqr. dist. - Group mean	27.47 (7.14)	31.13 (7.40)	0.50
20. Sqr. dist. - Group median	28.10 (8.11)	32.52 (8.58)	0.53
Dichotomous consensus			
21. SME as ref.	35.07 (3.32)	33.43 (3.46)	0.48
22. SME as ref. – neg. corr.	20.22 (6.59)	16.98 (6.86)	0.48
23. Group as ref.	38.31 (3.37)	36.69 (3.44)	0.48
24. Group as ref. – neg. corr.	26.70 (6.66)	23.49 (6.79)	0.48
Percent Agreement			
25. Endorsement rate - SME	56.04 (6.76)	51.72 (7.35)	0.61
26. Endorsement rate - Group	48.74 (4.71)	45.78 (5.11)	0.60
27. Proportions - SME	19.93 (1.99)	18.66 (2.16)	0.61
28. Proportions - Group	19.20 (1.37)	18.32 (1.63)	0.58

Note. SME = Subject Matter Expert Abs. = Absolute Sqr. = Squared dist. = distance ref. = reference neg. = negative corr. = correction *d* = Cohen's *d* (effect size) All differences were significant ($p < .001$)

Correlation with personality

Eighteen scoring methods resulted in an SJT score that had a significant but small positive correlation with agreeableness (Table 4). The largest correlation coefficients were found for scoring methods calculating the distance from the SME mean using standardised consensus. In addition, four scoring methods resulted in an SJT score that had a significant but small positive correlation with conscientiousness. The largest correlation coefficients were found for scoring methods calculating the absolute distance from the SME mean and median both using standardised consensus. Due to the low effect sizes and the small range of significant correlation coefficients, no general linear models were tested.

Table 4
Pearson correlation coefficients between the SJT score and the three Big Five personality dimensions in cohort 2014 only (N = 382).

Scoring method	N	A	C
Raw consensus			
1. Abs. dist. - SME mean	-.03	**-.11**	-.04
2. Abs. dist. SME median	.01	**-.11**	-.07
3. Abs. dist. SME mode	0	**-.11**	-.06
4. Sqr. dist. SME mean	-.03	**-.12**	-.04
5. Sqr. dist. SME median	-.01	**-.12**	-.06
6. Sqr. dist. - SME mode	-.01	**-.12**	-.05
7. Abs. dist. - Group mean	-.06	-.07	.02
8. Abs. dist. - Group median	-.06	-.08	0
9. Abs. dist. - Group mode	-.03	-.08	0
10. Sqr. dist. - Group mean	-.06	-.09	0
11. Sqr. dist. - Group median	-.06	**-.11**	0
12. Sqr. dist. - Group mode	-.05	**-.11**	-.01
Standardised consensus			
13. Abs. dist. - SME mean	0	**-.15**	**-.12**
14. Abs. dist. - SME median	-.01	**-.12**	**-.12**
15. Sqr. dist. - SME mean	0	**-.15**	**-.10**
16. Sqr. dist. - SME median	0	**-.13**	**-.11**
17. Abs. dist. - Group mean	.01	-.10	-.07
18. Abs. dist. - Group median	0	**-.11**	-.06
19. Sqr. dist. - Group mean	.02	-.10	-.06
20. Sqr. dist. - Group median	.01	**-.11**	-.06
Dichotomous consensus			
21. SME as ref.	-.07	.07	.10
22. SME as ref. – neg. corr.	-.07	.07	.10
23. Group as ref.	.02	**.14**	.05
24. Group as ref. – neg. corr.	.02	**.14**	.05
Percent Agreement			
25. Endorsement rate - SME	0	**.10**	.05
26. Endorsement rate - Group	.04	.06	.01
27. Proportions - SME	.03	**.11**	.05
28. Proportions - Group	.04	.08	.01

Note. N = Neuroticism A = Agreeableness C= Conscientiousness SME = Subject Matter Expert Abs. = Absolute Sqr. = Squared dist. = distance ref. = reference neg. = negative corr. = correction Bold coefficients reflect a significant relationship. For the scoring methods using distance metrics (number 1 to 20), a negative correlation coefficient reflects a positive relationship and vice-versa.

Discussion

This study shows that the psychometric quality of an SJT greatly depends on the choice of scoring method, specifically in the way the scoring method controls for systematic error. Firstly, the way of controlling for systematic error strongly affects the internal consistency reliability of an SJT score, with higher reliability estimates for scoring methods that use raw consensus. Secondly, the way of controlling for systematic error influences the adverse

impact of the SJT score, with a lower adverse impact for scoring methods that use raw consensus compared to dichotomous and standardised consensus. Lastly, the different scoring methods had a minor influence on the correlation with agreeableness and conscientiousness, but the practical significance of these correlations was negligible.

Internal consistency reliability

Our first finding was that the way a scoring method controls for systematic error strongly influences the internal consistency reliability. This strengthens the concerns about the use of coefficient alpha as a reliability estimate for an SJT score. Changing only the scoring method could alter the acceptability of the resulting reliability estimate from poor to sufficient, even for an SJT that was specifically constructed to measure one dimension. This large variety in internal consistency reliability is likely explained by the dependence of coefficient alpha on the total score variance (Streiner, 2003). Standardised and dichotomous consensus and percent agreement were associated with a reduction in total score variance, which is demonstrated by the lower standard deviations in Table 2. This reduction in total score variance will most likely lead to a lower coefficient alpha.

This line of reasoning implies that coefficients alpha reported in previous studies on SJTs may be strongly influenced by irrelevant aspects, such as the total score variance generated by the scoring method used. Assuming that most studies on SJTs arbitrarily choose one scoring method rather than another, choice of scoring method contributes to the limited usefulness of coefficient alpha as a reliability estimate for SJTs. Future studies should investigate whether the large variation in coefficient alpha caused by different scoring methods also occurs in other reliability estimates (e.g. alternate forms reliability) to find out whether this large variation is an artefact of coefficient alpha only.

A more accurate reliability estimate might be obtained by a combination of a more thoroughly construct-based SJT development (Christian et al., 2010) and a reliability estimate that takes into account the imposed factor structure of the SJT, for example a structural equation modelling (SEM) reliability estimate (Yang & Green, 2011) or stratified alpha (Catano et al., 2012). Future research is required on the application of construct-based development methods and alternative internal consistency estimates for SJTs.

Adverse impact

Although all scoring methods led to significant ethnic differences in SJT score, the way a scoring method controlled for systematic error influenced the size of these effects. Specifically, the effect size decreased when using raw consensus instead of standardised or dichotomous consensus. This result is not in line with the findings of McDaniel et al. (2011) who found lower ethnic subgroup differences for scoring methods that controlled for systematic error (i.e. standardised and dichotomous consensus), which they explained by the removal of ethnicity related response tendencies in the use of Likert scales. However, the uncorrected effect sizes do show some support for this line of reasoning with the lowest effect sizes reported for the scoring methods using dichotomous consensus. The absence of lower effect sizes for standardised consensus might be caused by the low number of scale points (i.e. four) on the Likert scale that was used. Narrow Likert scales may not be as strongly affected by response tendencies as Likert scales with more scale points (Flaskerud, 1988), resulting in no differences when controlling for the response tendencies. A study on script concordance tests recommended a reduction of the Likert scale from five to three points in order to decrease the influence of construct-irrelevant factors such as examinee response styles (Lineberry, Kreiter, & Bordage, 2013). Dichotomising the Likert scale does seem to

have some effect on adverse impact, but at the cost of low internal consistency reliability, leading to a similar issue as the diversity-validity dilemma (De Soete et al., 2013).

Another noteworthy finding is that adverse impact was similar for both reference groups (SMEs and respondents). Previous studies which compared different reference groups found similar validity coefficients for the scores of both groups (Legree et al., 2005; Motowidlo & Beier, 2010), but did not study the effect of the reference group on adverse impact. Most SJTs use SMEs as a reference group under the assumption that they have considerable experience in a relevant setting and therefore know what kind of behaviours are appropriate in the described situations. Our results suggest that the use of a reference group of inexperienced respondents (i.e. secondary school students) does not affect the adverse impact of an SJT.

A possible explanation for this comparable adverse impact is the better representativeness of the group of respondents with respect to ethnicity. All our SMEs in this study were native Dutch, while only 57 % of the applicants were native Dutch. Little is known about the cultural susceptibility of integrity. However, medical professionalism has been found to depend on cultural context (Chandratilake, McAleer, & Gibson, 2012; Jha, McLean, Gibbs, & Sandars, 2015) and since integrity is an important aspect of medical professionalism, it too might depend on cultural context (Arnold & Stern, 2006). A reference group that is more representative of the demographic characteristics of the applicant group may lead to a more accurate measurement of the targeted construct and may therefore result in equal or less adverse impact. Future research should investigate the effect of the demographic composition of the reference group on the psychometric quality of an SJT.

Another explanation for the equal adverse impact for both type of reference groups might be that there were too few SMEs to be able to achieve proper consensus on the difficult dilemmas described in the scenarios. This was reflected by the non-perfect agreement in the SMEs' evaluation of the response options (ICC = .65). A group of 931 individuals might result in more meaningful consensus. This contention is supported by Legree et al. (2005), who stated that in light of equal validity coefficients, an examinee-based scoring standard gives more reliable values than an expert-based scoring standard, due to the larger number of examinees.

Correlation with personality

Our last finding was that 18 scoring methods showed a correlation with agreeableness and four scoring methods showed a correlation with conscientiousness, which was in line with previous research (Marcus et al., 2007; McDaniel et al., 2007). However, these correlations must be interpreted with caution, since all correlation coefficients represent small effects and it is likely that the large sample size has contributed to the statistical significance of these small effects. The larger number of significant correlations among scoring methods using standardised consensus is in line with the findings of McDaniel et al. (2011) and might be explained by the removal of systematic error from the SJT score. However, the small effect size of these correlations between the integrity-based SJT score and the three Big Five personality traits precludes any conclusive statements about the effect of scoring method on the correlation with personality.

The small number of significant correlations between the SJT score and the Big Five personality traits is in consonance with a previously reported non-association between the Big Five personality traits and the score on a multiple mini interview (MMI), another widely used selection instrument for medical school (Kulasegaram, Reiter, Wiesner, Hackett, & Norman, 2010). This non-association might be explained by the fact that personality tests assess noncognitive traits, whereas MMIs and SJTs assess noncognitive behaviours.

Noncognitive behaviours are more dependent on situational factors than personality traits (Eva, 2005). This is in line with a previous study which demonstrated that a contextualised personality measure had higher criterion validity for academic performance and counterproductive academic behaviour than a generic personality measure (Holtrop et al., 2014). The lack of contextualisation of the NEO-PI-R limits the usefulness of personality tests in medical school selection and may be an explanation for the absence of any meaningful correlations between the SJT score and personality.

Scoring method aspects revisited

Four scoring method aspects were examined. Differences in internal consistency reliability and adverse impact were found for only one aspect: the way of controlling for systematic error, with raw consensus leading to scores with the highest coefficient alpha and the smallest ethnic subgroup differences. As mentioned above, these differences might be explained by the effect of this scoring method aspect on the total score variance and the negligible effect of response tendencies due to the narrow Likert scale used in this study. No differences were found for the other three aspects (i.e. reference group, distance and central tendency statistic).

As stated before, the absence of differences for reference group might be caused by the larger size and better representativeness of the group of respondents itself, which might remove the benefits of using a highly experienced but small group of SMEs. Another potential reason is that integrity-related issues in the beginning stage of medical school do not require specific knowledge but more general knowledge which can be possessed by both reference groups, which is reflected by a correlation of .90 between the group of SMEs and group of respondents itself in their average rating.

The absence of differences for the scoring method aspect of distance (absolute vs. squared) may be explained by the low number of scale points on the Likert scale (i.e. four), which means that the maximum distance between an applicant's rating and the overall rating can never exceed three. This may not be sufficient to get a significant difference in the outcome measure when squaring the distance between both ratings. Future research should examine the scoring method aspect of distance for SJTs using Likert scales with more scale points.

Lastly, the similar results for the three different central tendency statistics may be explained by the distribution of the ratings across the Likert scale. Data with a symmetric distribution are best summarised using the mean. Since the mean is strongly influenced by extreme scores (Field, 2013), asymmetrically distributed data are better summarised using the median or mode. A four-point Likert scale precludes extreme scores leading to similar values for the mean, median and mode and likely causes the comparable results for this scoring method aspect.

Practical implications

The most important practical implication of this study is that it creates awareness about the importance of carefully considering the immense number of possibilities for converting the judgements on an SJT to a score. Instead of arbitrarily choosing one of the many existing methods, researchers and practitioners should accompany the development of an SJT with a thorough examination of the scoring method to be used. In addition, this study demonstrated that the results when using the group of respondents itself are similar to those obtained when using a group of SMEs as reference. Using the group of respondents has practical and economic advantages, since the collection of data from SMEs can be difficult.

Unfortunately, this study does not allow any conclusive statements about which scoring method is best, because the findings are highly dependent on this particular SJT measuring

this particular construct in this particular setting. Firstly, this study was conducted in the Netherlands, where medical school applicants are relatively young (17–18 years). The use of more mature applicants may lead to different results for scoring methods that use the group of respondents itself as a reference. Secondly, the cultural context may influence the way the reference group judges integrity-related dilemmas (Chandratilake et al., 2012; Jha et al., 2015). Finally, SJTs measuring other constructs than integrity might be differentially influenced by changing the scoring method. Future research should replicate this study with other SJTs measuring different constructs in other settings to investigate the generalisability of these findings and to provide clarity on which scoring method is best for which situation.

Strengths and limitations

To our knowledge, this is the first study to compare such a large number of scoring methods, varying not only the way of controlling for systematic error and the type of reference group, but also the type of distance and central tendency statistic. Next to the large number of scoring methods examined, this study also contributes to previous research by the examining the effect of scoring method on internal consistency reliability. Embedding the administration of the SJT into the selection procedure led to a very high response rate, ensuring that our results were not influenced by a volunteer bias. The credibility of our results is further supported by a relatively small restriction of range. Unlike many other selection procedures, the current selection procedure was not preceded by a pre-selection on cognitive competencies.

Although this study compared a large number of scoring methods, we do not claim that this list is exhaustive. Examples of other approaches for scoring SJTs are the squared Mahalanobis distance (Barbot et al., 2012) and the use of paired comparisons (Gold & Holodynski, 2015). It seems that the possibilities are endless and future studies should investigate these other scoring methods. For practical reasons, the number of scoring methods in this study was limited to 28.

Conclusion

In conclusion, although the SJT scoring method is often chosen arbitrarily, this study shows that changing the scoring method strongly influences the internal consistency reliability and adverse impact of an SJT score. The most influential characteristic of a scoring method is the way of controlling for systematic error. Given the increasing use of SJTs for selection into medical school, it is crucial to thoroughly examine which scoring method is best to use.

Chapter 3

Integrity situational judgement test for medical school selection: Judging 'what to do' versus 'what not to do'

This chapter has been published as:
De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2018). Integrity situational judgement test for medical school selection: Judging 'what to do' versus 'what not to do'. *Medical Education, 52*, 427-437.

Abstract

Despite their widespread use in medical school selection, there remains a lack of clarity on exactly what situational judgement tests (SJTs) measure. We aimed to develop an SJT that measures integrity by combining critical incident interviews (inductive approach) with an innovative deductive approach. The deductive approach guided the development of the SJT according to two established theoretical models, of which one was positively related to integrity (honesty-humility; HH) and one was negatively related to integrity (cognitive distortions; CD). The Integrity SJT covered desirable (HH-based) and undesirable (CD-based) response options. We examined the convergent and discriminant validity of the Integrity SJT and compared the validity of the HH-based and CD-based subscores. The Integrity SJT was administered to 402 prospective applicants at a Dutch medical school. The Integrity SJT consisted of 57 scenarios, each followed by four response options, of which two represented HH facets and two represented CD categories. Three SJT scores were computed, including a total, an HH-based and a CD-based score. The validity of these scores was examined according to their relationships with external integrity-related measures (convergent validity) and self-efficacy (discriminant validity). The three SJT scores correlated significantly with all integrity-related measures and not with self-efficacy, indicating convergent and discriminant validity. In addition, the CD-based SJT score correlated significantly more strongly than the HH-based SJT score with two of the four integrity-related measures. An SJT that assesses the ability to correctly recognise CD-based response options as inappropriate (i.e. what one should not do) seems to have stronger convergent validity than an SJT that assesses the ability to correctly recognise HH-based response options as appropriate (i.e. what one should do). This finding might be explained by the larger consensus on what is considered inappropriate than on what is considered appropriate in a challenging situation. It may be promising to focus an SJT on the ability to recognise what one should not do.

Introduction

In addition to the cognitive instruments used in selection for medical school, there is an increasing need for tools that assess noncognitive attributes (e.g. integrity). This growing need has led to the introduction of new medical school selection instruments such as multiple mini-interviews, selection centres, personality and emotional intelligence assessments and situational judgement tests (SJTs) (Patterson et al., 2012; Patterson, Knight, et al., 2016). The SJT presents applicants with challenging situations they may encounter during medical school. These situations are followed by a number of possible responses for which applicants need to judge the appropriateness (Motowidlo et al., 1990). Previous studies on SJTs in medical school selection demonstrated predictive validity and incremental validity over cognitive ability tests (Koczwara et al., 2012; Lievens, 2013; Lievens & Sackett, 2012; Patterson et al., 2009). Furthermore, SJTs result in less adverse impact than traditional cognitive tests with respect to applicants with backgrounds of low socioeconomic status (Lievens et al., 2016).

The application of an SJT in medical school selection necessitates the identification of what is measured by an SJT because this high-stakes process requires clarity on the constructs used for selection. However, few studies elaborate on exactly what SJTs measure (Christian et al., 2010). This limited attention can be explained by the fact that most SJTs use an inductive development approach in which the content of the SJT is matched as closely as possible to the criterion domain (e.g. job performance) (Christian et al., 2010; Weekley et al., 2006). Most inductive development approaches base the content of the SJT on critical incidents (i.e. anecdotal incidents of exceptionally good and exceptionally poor behaviour) (Flanagan, 1954; Weekley et al., 2006). This point-to-point correspondence with the criterion contributes to the perceived job-relatedness of an SJT (Lievens, 2013) and the contextualisation may strengthen its predictive validity (Holtrop et al., 2014). However, the inductive development method gives little insight into which constructs are measured because the criterion domain tends to be highly heterogeneous and to consist of various technical, interpersonal and motivational aspects (Chan & Schmitt, 2002).

By contrast, a deductive development approach bases the content of an SJT on a specific construct by using a literature review, a job analysis or an existing theory (Weekley et al., 2006). The deductive approach has several advantages. Firstly, it facilitates better understanding of why an SJT is related or unrelated to the criterion domain (Christian et al., 2010). Secondly, it supports more meaningful comparisons with other predictors of future performance (Chan & Schmitt, 2002), which are valuable when an admission board intends to apply different weights to the various components of a selection battery (Chan & Schmitt, 2005). Finally, it enables the comparison of different SJT formats (e.g. written versus video-based) designed to measure the same construct (Chan & Schmitt, 1997). A possible disadvantage of the deductive method is reduced realism.

To benefit from the strengths of both methods, we combined the inductive and deductive approaches to develop an SJT measuring applicants' knowledge of appropriate and inappropriate responses to integrity-related situations in medical school (henceforth: Integrity SJT). Integrity is considered a core competency for medical doctors across various medical specialties (Frank et al., 2015; Patterson, Ferguson, & Thomas, 2008) and is therefore considered a relevant construct for selection. Integrity was characterised by honesty, sincerity, fairness and modesty (Ashton & Lee, 2007) and the absence of inaccurate self-serving thoughts and antisocial and counterproductive behaviour (Barriga & Gibbs, 1996). We are aware of three deductively developed SJTs to measure integrity, including two

outside and one within medical education. Firstly, Becker (2005) applied a set of integrity values in developing an SJT measuring employee integrity. This SJT was associated with integrity-related work outcomes. Secondly, De Meijer et al. (2010) developed a video-based SJT for the Dutch police consisting of scenarios depicting police integrity violations. This SJT was related to established integrity-related measures and unrelated to cognitive ability and thus demonstrated both convergent and discriminant validity (De Meijer et al., 2010). Finally, Husbands et al. (2015) developed an integrity SJT for medical school admission based on a literature review on integrity constructs (e.g. honesty). This SJT correlated to honesty-humility, the integrity-related subscale of the HEXACO personality inventory (Husbands et al., 2015). By contrast with the traditional Big Five personality model, the HEXACO personality model consists of six dimensions as a result of the addition of the honesty-humility dimension (Ashton & Lee, 2005).

The present study contributes to the existing research on two points. Firstly, we developed an SJT that covers appropriate and inappropriate responses. In this way, the SJT assesses the ability to identify appropriate responses, as well as the ability to identify inappropriate responses. We distinguished these two abilities because previous researchers suggested that they involve different skills (Stemler et al., 2016). Secondly, we used an innovative deductive development approach to create the desirable and undesirable response options in the SJT whereby two established theoretical models (one positively and one negatively related to integrity) were used to guide the development of the response options.

The deductive approach was combined with an inductive approach (i.e. critical incident interviews) to ensure the realism of the SJT. Next, we addressed the research question: What are the convergent and discriminant validity levels of the Integrity SJT? Convergent validity was examined according to the relationship with external integrity-related measures. Discriminant validity was investigated using the relationship with an unrelated external measure (i.e. self-efficacy). The validity levels of scores based on the appropriate and inappropriate response options of the SJT were compared. With the combination of the inductive and the innovative deductive development approach, we aimed to enhance the convergent and discriminant validity of an SJT measuring integrity. In addition, we aimed to investigate the effect of the distinction between ‘what to do’ and ‘what not to do’ on the construct validity of the Integrity SJT. The outcomes of this study will add to the knowledge about this increasingly popular tool in medical school selection.

Methods

Context

This study was conducted at the Erasmus Medical Centre (MC) Medical School, Rotterdam, the Netherlands. In the Netherlands, all entry to medical school is predominantly at the undergraduate level. Admission to the Erasmus MC Medical School at the time of the study was based on three aspects: pre-university grade point average; extracurricular activities (e.g. work-related activities in health care), and performance on five cognitive study skill tests (e.g. scientific reading) administered during three testing days (Urlings-Strop et al., 2009). The SJT was not part of the admission procedure but was administered solely for research purposes. Approximately 50% of the applicants were admitted to the Erasmus MC Medical School.

Six months before the testing days, the Erasmus MC Medical School organised a selection orientation day to inform medical school applicants about the selection process. Participation in the selection orientation day was voluntary and free of charge.

Participants and procedure
The Integrity SJT was administered to the 402 participants at the 2015 selection orientation day. Participation in the SJT was voluntary. Participants were informed about the purpose of the administration and that their answers would not influence the admission decision. Informed consent was obtained from all participants. The data in this study were confidentially processed. The pencil-and-paper administration took place in a lecture hall at the Erasmus MC Medical School campus. The Ethics Committee of the Institute of Psychology, Erasmus University Rotterdam, deemed this study to have no need for further ethical approval by the Medical Ethics Committee.

Measures

Demographic questionnaire
A demographic questionnaire was administered to determine the participants' ethnic and socioeconomic backgrounds. An individual was classified as belonging to an ethnic minority if at least one of his or her parents had been born outside the Netherlands (i.e. the definition used by Statistics Netherlands). Otherwise, an individual was classified as Dutch. Socioeconomic background was determined according to the level of education of the participants' parents. First-generation university students are individuals whose parents did not attend higher education (Stegers-Jager, Steyerberg, Cohen-Schotanus, & Themmen, 2012).

Development of the Integrity SJT
The deductive development approach was guided by two integrity-related models: the honesty-humility (HH) subscale of the HEXACO personality inventory, and the How I Think questionnaire measuring cognitive distortions (CDs). The HH dimension has been demonstrated to be positively related to integrity (Lee, Ashton, & De Vries, 2005) and was used to create desirable responses. The CDs describe inaccurate thinking styles which may lead to antisocial behaviours (Barriga & Gibbs, 1996) that are negatively associated with integrity (Ones, Viswesvaran, Schmidt, & Reiss, 1994). Therefore, these were used to create undesirable responses. Specifically, sets of response options were written to represent each of seven response option categories assembled according to three HH facets (i.e. sincerity, fairness and modesty) and four CD categories (i.e. self-centredness, blaming others, minimising and assuming the worst). These response option categories are described in Table 1.

The inductive development approach consisted of critical incident interviews with nine subject matter experts (SMEs), who were individuals directly involved in the assessment of professional behaviour of medical students (e.g. clinical skills teachers). These SMEs described incidents in which a medical student behaved unprofessionally (e.g. by cheating). Further questions were asked to provide elaboration on these critical incidents following the technique described by Flanagan (1954). These incidents formed the basis of the SJT scenarios. The scenarios were presented to a group of medical students and staff (n = 41) to gather input for realistic response options. To stimulate the development of response options, scenarios were presented with a number of prompts (e.g. What would be the best/worst/most likely response to this situation?) (Lievens & Schollaert, 2008).

Table 1
Short description of each response option category including the number of items per category for both SJT versions.

Response option category	Short description	Version A	Version B
HH facet			
Sincerity	Being honest and genuine	20	18
Fairness	Being fraud and corruption avoidant	20	18
Greed avoidance	Being unmaterialistic	-	-
Modesty	Not claiming special treatment	18	20
CD category			
Self-centeredness	Putting one's own needs and desires above those of others (egocentrism)	15	15
Blaming others	Misattributing antisocial behaviour to outside sources	14	13
Minimizing/ Mislabelling	Regarding antisocial behaviour as harmless/using dehumanizing labels on others	15	14
Assuming the worst	Interpreting antisocial behaviour as a reaction to hostile intentions attributed to others	14	14

Note. HH = honesty-humility CD = cognitive distortion Greed avoidance and Mislabelling were not used for the SJT in this study.

The resulting Integrity SJT consisted of 57 scenarios. This pilot version of the Integrity SJT was randomly split into two versions (i.e. Version A and Version B) because of the large number of scenarios. Each scenario was followed by four response options, of which two represented HH facets and two represented CD categories. Table 1 presents the distribution of items across the seven response option categories. All scenarios described situations at the beginning of medical school. No medical knowledge was required to understand the scenarios because the target population of this study were applicants for undergraduate entry who, in general, have limited experience in health care. On average, scenarios were described in 56.4 words and response options in 12.9 words. An example SJT item is given in Box 1. Five additional example items are presented in Appendix 3A.

Each SJT item was scored by calculating the squared distance between a participant's judgement and the average judgement across all other participants. To ensure that the SJT score was not influenced by responder tendencies to use the rating scale in a certain manner (e.g. extreme response style), this calculation was preceded by a within-person *z* standardisation so each participant had a mean score of 0 and a standard deviation (*SD*) of 1 (McDaniel et al., 2011). Unlike most SJTs, SMEs did not contribute to the scoring key as previous research has demonstrated the similarity of judgements of novices and experts (De Leng et al., 2017; Motowidlo & Beier, 2010). However, to guarantee the comparability of novices and experts in this study, we compared item scores based on the average judgement of the group of participants with item scores based on the average judgement of a group of general practice (GP) residents (n = 63). These residents were chosen as a reference group because this group includes a relatively large number of residents who are trained as generalists. For the GP residents, the SJT was split into three versions of 19 scenarios (n_I = 23, n_{II} = 18, n_{III} = 22) in order to reduce the time investment. The mean (SD) age of the GP

residents was 28.6 (2.7) years and 52 (82.5%) of them were female. Fifty-one (81.0%) GP residents were Dutch and 21 (33.0%) were first-generation university students. Appendix 3B presents the intraclass correlation coefficients for the GP residents for the total SJT score, the subscore based on the HH SJT items and the subscore based on the CD SJT items.

Box 1

Example scenario [including corresponding response option categories]							
John finds out that Mary has a copy of the exam paper that will be given next week. She tells him that she has already sold the exam to some fellow students and asks him if he also wants to look at the exam paper. *Judge for each of the following response options how appropriate they would be for John.*							
		Very inappropriate					Very appropriate
1.	Look at the exam paper because everyone would do that. [Minimizing]	1	2	3	4	5	6
2.	Don't look at the exam since he is not entitled to do so. [Modesty]	1	2	3	4	5	6
3.	Look at the exam and tell no one you did. [Self-centeredness]	1	2	3	4	5	6
4.	Don't look at the exam and inform the teacher. [Fairness]	1	2	3	4	5	6

Convergent and discriminant validity

Convergent validity was examined by the relationship between the Integrity SJT and the two integrity-related measures used for assembling the response option categories: the HH subscale of the HEXACO Simplified Personality Inventory (HEXACO-SPI) (De Vries & Born, 2013) and the How I Think (HIT) questionnaire measuring CDs (Barriga & Gibbs, 1996; Barriga, Gibbs, Potter, & Liau, 2001). To thoroughly analyse the convergent validity, we examined the relationship with two additional integrity-related measures: the student-related items of the Inventory of Counterproductive Behaviour (ICB) (Hakstian, Farrell, & Tweed, 2002; Marcus et al., 2007) and the workplace deviance measure (Bennett & Robinson, 2000). The student-related items of the ICB assess counterproductive academic behaviour (i.e. intentional behaviours in conflict with the objectives of an educational institution) (Gruys & Sackett, 2003). Workplace deviance refers to the deliberate violation of the norms of an organisation (Robinson & Bennett, 1995). The items of the workplace deviance measure were rewritten to fit the context and two items were deleted because they were considered irrelevant to an academic context.

Discriminant validity was examined according to the relationship with the self-efficacy subscale of the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich, 1991). Self-efficacy is a person's belief in his or her ability to reach desired goals (Bandura, 1994). Self-efficacy is an important predictor of medical school performance (Mavis, 2001; Stegers-Jager, Cohen-Schotanus, & Themmen, 2012), but is expected to be unrelated to integrity. The items were slightly adapted to fit the context of the study. The characteristics of these measures are described in Appendix 3C.

Statistical analyses
Three SJT scores were computed by adding up scores across: i) all items (i.e. total SJT score), ii) all HH-based items, and iii) all CD-based items. Scores were reversed so that higher scores indicated better performance on the Integrity SJT. Pearson's correlation coefficients were calculated between the three SJT scores and the integrity-related measures and self-efficacy subscale. The correlation coefficients were merged across the two versions of the Integrity SJT using a random-effects meta-analytic approach.

The difference between the HH-based and CD-based SJT scores in their correlations with the integrity-related measures was analysed with the Williams' test (Steiger, 1980). Given the large number of correlations, a stricter alpha level was used ($\alpha = 0.01$). Correlation analyses were conducted using IBM SPSS Statistics for Windows Version 21.0 (IBM Corp., Armonk, NY, USA). R Version 3.1.0 (www.R-project.org) was used to meta-analytically merge the correlation coefficients ('metacor' package) and to conduct the Williams' test ('psych' package).

Results

Demographics
The numbers of participants completing Versions A and B of the SJT were 186 (response rate: 92.5%) and 181 (response rate: 90.0%), respectively. There were no significant differences in age, gender, ethnicity or socioeconomic background between participants completing Versions A and B (Table 2). The mean age of the undergraduate entry applicants was 17.8 years, 271 participants were female (73.8%), 132 came from ethnic minorities (36.0%) and 108 were first-generation university students (29.4%). Scores on the integrity-related measures and self-efficacy subscale were comparable for the participants of the two versions, except for the HIT questionnaire ($t(354) = -2.77$, $p = .006$, $d = 0.29$). However, the effect size of this difference was small and both groups scored well below the average score of a normative sample of 412 youths (mean score: 2.39) (Barriga et al., 2001). Of the participants in the selection orientation day, 352 applied to medical school (87.6%), indicating that the participants were suitably representative of medical school applicants. For both Versions A and B, examination of the skewness and kurtosis of the SJT score distributions showed negative skewness (Table 2) (i.e. most participants obtained a high score on the SJT).

Preliminary analyses
For each SJT item, two scores were generated: one of these used the GP residents as a reference and the other used the group of participants itself as a reference. Correlations between these two scores were calculated. For Version A, the average correlation across the 116 items was 0.93 (range: 0.27-1.00). All but three items had a correlation above 0.50 (i.e. large effect size) (Cohen, 1988). For Version B, the average correlation across the 112 items was 0.93 (range: 0.11-1.00). Only two items had a correlation below 0.50. The negligible number of correlations below 0.50 was deemed sufficient to confirm the use of a scoring key with the group of participants itself as a reference.

Table 2
Respondent demographics and descriptive data for the study's measures.

	Version A (N = 186)	Version B (N = 181)	Range (min. – max.)
Gender (% female)	75.1	72.9	
Age (Mean (SD))	17.8 (2.2)	17.7 (1.8)	
Ethnicity (% XXX)	63.4	63.9	
First-generation university	28.1%	31.1%	
Integrity-related measures			
HEXACO-SPI HH	43.36 (6.01)	44.36 (6.06)	16-80
HIT questionnaire	**1.63 (0.42)**	**1.75 (0.41)**	1-6
ICB student-related items	2.89 (0.84)	2.82 (0.91)	1-6
WD measure	2.38 (0.81)	2.23 (0.86)	1-7
MSLQ Self-efficacy	46.27 (5.76)	45.33 (6.64)	8-56
Skewness			
Total	-1.97	-1.83	
HH-based	-2.03	-1.62	
CD-based	-1.59	-1.94	
Kurtosis			
Total	4.09	3.67	
HH-based	4.65	2.48	
CD-based	2.31	4.37	

Note. SD = Standard deviation HEXACO-SPI = HEXACO Simplified Personality Inventory HIT = How I Think ICB = Inventory Counterproductive Behaviour WD = Workplace Deviance MSLQ = Motivated Strategies of Learning Questionnaire HH = honesty-humility CD = cognitive distortions Bold numbers indicate a significant difference ($p < .01$, two-tailed)

Main analyses

All SJT scores (i.e. total, HH-based and CD-based) correlated significantly with the four external integrity-related measures (Table 3). The correlations were in the expected direction and indicated a moderate effect size ($0.22 \leq r \leq 0.40$). Appendix 3D presents the correlations between the individual response option categories, HH facets and CD categories.

All correlation coefficients with the integrity-related measures were – in absolute terms – larger for the CD-based SJT score than for the HH-based SJT score (Table 3). The Williams' test indicated that the CD-based SJT score correlated significantly more strongly than the HH-based SJT score with the HIT questionnaire ($t(168) = 3.07$, $p = .003$, $d = 0.47$) and with the ICB ($t(171) = 2.69$, $p = .008$, $d = 0.41$). The CD-based SJT score correlated more strongly than the HH-based SJT score with the honesty–humility subscale, but this difference was only marginally significant ($t(173) = -2.54$, $p = .011$, $d = 0.39$). No significant difference was found between the HH-based and CD-based SJT scores in their correlation with the workplace deviance measure ($t(169) = 1.50$, $p = .130$).

As expected, none of the SJT scores were significantly correlated to the self-efficacy subscale (Table 3).

Table 3

Descriptive data for the total score, the HH-based and CD-based SJT scores and correlations between the total score, the HH-based and CD-based SJT scores and the integrity-related measures and self-efficacy subscale.

	Version A		Version B		Integrity-related measures				
Score	M/max.	SD	M/max.	SD	HH	HIT*	ICB*	WD	SE
Total	82.44	19.13	77.77	22.85	**.37** [.27; .45]	**-.35** [-.50; -.17]	**-.34** [-.55; -.09]	**-.27** [-.40; -.13]	.01 [-.10; .11]
HH	75.45	10.15	80.79	12.43	**.29** [.19; .38]	**-.26** [-.41; -.13]	**-.26** [-.44; -,06]	**-.22** [-.32; -.11]	-.01 [-.11; .10]
CD	81.21	10.08	81.44	11.68	**.40** [.31; .49]	**-.40** [-.58; -.18]	**-.38** [-.60; -.10]	**-.29** [-.44; -.13]	.02 [-.08; .13]

Note. M/Max. = mean as a percentage of the maximum score (because version A and B have a different number of items) SD = standard deviation HH = honesty-humility CD = cognitive distortions HIT = How I Think ICB = Inventory Counterproductive Behaviour (academic) WD = Workplace deviance SE = Self-efficacy Integrity-related measures with an asterisk (*) had a significantly different correlation with the CD-based SJT score than the HH-based SJT score ($p < .01$) 95% confidence interval between square brackets Descriptive data are presented for each version separately, correlations are meta-analytically merged across both versions Bold coefficients depict a significant correlation ($p < .01$, two-tailed)

Discussion

The results of this study indicate that the Integrity SJT had convergent and discriminant validity. This is evidenced by a significant correlation with integrity-related measures and no correlation with a self-efficacy subscale. Additionally, the findings indicate that an SJT score representing CD categories has stronger convergent validity than an SJT score representing HH facets. This is demonstrated by significantly higher correlations with two of the four integrity-related measures for the CD-based SJT score than for the HH-based SJT score.

The first finding implies that the use of a deductive development approach based on established theoretical models together with a traditional inductive approach generates an SJT that has convergent validity. The correlation with the HH subscale found in this study appears to be somewhat stronger than the correlation coefficient reported in the study by De Meijer et al. (2010) and is similar to the uncorrected correlation coefficient reported in the study by Husbands et al. (2015). The strength of the correlation with the HIT questionnaire found in this study is similar to that of the correlation reported in the study by De Meijer et al. (2010). However, a prior study demonstrated a negative association between the score on the HIT questionnaire and a person's level of education (Nas, Brugman, & Koops, 2008). Thus, the correlation with the HIT questionnaire in this study might be attenuated by the high pre-university education level of the participants. Different SJTs and contexts in these studies make it difficult to perform a direct comparison of the correlation coefficients.

Nonetheless, the established integrity-related models proved to be a useful guide to deductively develop the Integrity SJT. Moreover, the convergent validity of the Integrity SJT was at least as strong as the correlations reported in prior studies (De Meijer et al., 2010; Husbands et al., 2015). The use of theoretical models for the development of an SJT is supported by previous studies on SJTs outside the medical domain measuring constructs other than integrity. For example, an SJT developed on the basis of eight dimensions of an existing leadership model was significantly correlated to an external leadership questionnaire (Peus, Braun, & Frey, 2013). Additionally, an SJT developed on the basis of a conflict management model was significantly related to supervisor ratings of on-the-job conflict management (Olson-Buchanan et al., 1998). Overall, these findings suggest that a deductive development approach based on established theoretical models enhances the construct and predictive validity of an SJT. Future research is required to identify which characteristics of the deductive development approach positively influence the SJT's validity and should attempt to make a more direct comparison of the two development approaches. The positive findings with respect to the use of theoretical models in SJT development should not diminish the importance of the inductive development approach. The inductive approach uses empirical data to contextualise the SJT's content. The contextualisation could lead to stronger predictive validity (Robie, Risavy, Holtrop, & Born, 2017), higher perceived job-relatedness (Lievens et al., 2008) and lower susceptibility to socially desirable responding than, for example, non-contextualised personality tests (Hooper et al., 2006). The strengths of an SJT are enhanced by a combination of both development methods.

The second finding of this study indicates that an SJT score based on the ability to identify what one should not do has stronger convergent validity than an SJT score based on the ability to identify what one should do. This finding is in line with that in a prior study on sales and management SJTs, which demonstrated stronger predictive validity for the ability to identify the worst response option than for the ability to identify the best response option (Stemler et al., 2016). A similar finding was reported in another SJT study on teachers' tacit knowledge in which a subscale assessing the ability to detect bad responses was better able to

discriminate experts from novices than a subscale assessing the ability to detect good responses (Elliott et al., 2011). This finding might be explained by a larger consensus on what is considered inappropriate than on what is considered appropriate in a challenging situation. There exist a variety of reactions that may be considered appropriate but the eventual response depends on the type of job, organisation and culture (e.g. appropriately solving a problem with one's supervisor differs between vertical and horizontal organisational structures). However, inappropriate reactions are most likely to always lead to negative outcomes regardless of the type of job, organisation or culture (Stemler et al., 2016). Indeed, the GP residents in this study showed greater agreement in their judgements of the CD-based response options than in their judgements of the HH-based response options. Unlike prior studies that empirically determined the best and worst responses (e.g. using SMEs) (Becker, 2005; Elliott et al., 2011), the present study deductively established desirable and undesirable responses. The deductive development approach does not require the input of SMEs, which may be beneficial because it can be difficult to determine who is best placed to serve as an expert and practically inconvenient to collect data from this group. However, we have not yet examined the relationship of the Integrity SJT with future performance and therefore further research is necessary to determine if the stronger predictive validity for the ability to identify what one should not do is also observed for the SJT in this study.

Strengths, limitations and recommendations for future research

An important strength of this study lies in its combination of two development approaches, which allows us to benefit from the advantages of both methods and results in an SJT with realistic contextualised scenarios measuring an explicit construct. A second strength is the large number of integrity-related measures used in this study, which supports the credibility of our statements regarding convergent validity. A third strength refers to the fact that, unlike most previous studies, the current work not only examined convergent validity, but also investigated discriminant validity, thereby indicating that the Integrity SJT is associated with theoretically related constructs and not associated with theoretically unrelated constructs.

Despite its strengths, this study has some limitations. Firstly, the response options of the Integrity SJT were written to represent response option categories by aligning the wording and reasoning of response options belonging to the same category. Future research might improve the accuracy of this categorisation by performing an additional classification by an independent group. Secondly, the assumption that the HH facets reflect good responses and that CDs reflect bad responses may be too simplistic. For example, an HH-based response might entail the betrayal of one's friend and a CD-based response might seem to be made inevitable by group pressure. The influence of these subtleties on the functioning of an SJT should be further investigated. Thirdly, the investigation of systematic ethnic differences in the score on the Integrity SJT was beyond the scope of this paper, but future research is necessary to examine the 'what to do' versus 'what not to do' distinction with regard to adverse impact. Fourthly, critical incident interviews were conducted with only nine SMEs. Although the critical incident interviews produced a wealth of data, interviews with more SMEs may have led to a wider coverage of the professional issues encountered by medical students. Finally, the results of this study are derived solely from its administration within an admission context with undergraduate entry. As a result, the patient-centeredness of the SJT scenarios was limited, which may reduce the generalisability of the present results to SJTs used for graduate entry into medical school. Although the Integrity SJT involved some patient-related scenarios, future research should investigate the generalisability of this study's findings to other settings.

These findings elicit the following recommendations for future research. Firstly, the Integrity SJT showed stronger convergent validity for the CD-based score than for the HH-based score. However, it is possible that for other constructs (e.g. empathy), a score based on the correct identification of desirable responses will have stronger convergent validity than a score based on the correct identification of undesirable responses, perhaps because desirable responses are more obvious for certain constructs. Future research is necessary on the generalisability of the CD-based score's stronger convergent validity to SJTs measuring other constructs. Finally, future research on the predictive validity is a necessary requirement before an SJT can be considered for inclusion in medical school selection.

Practical implications

A first practical implication for medical schools using or planning to use a construct-based SJT in their selection procedures is the use of established theoretical models to guide the deductive development of an SJT. The theoretical models may be related to integrity, but may also involve other constructs (e.g. social competence).

A second practical implication is that an SJT might be used to assess the ability to correctly identify what one should not do in a challenging situation. This implication could support the proposal to use an SJT for screening out medical school applicants (Patterson et al., 2016) as SJTs appear to be more informative at the lower end of the distribution (Cousans et al., 2017; Tiffin & Carter, 2015). Only a small group of medical students behaves unprofessionally and is unresponsive to remediation activities as a result of poor insight and poor adaptability (Mak-Van der Vossen et al., 2016). An SJT that assesses the ability to identify inappropriate response options may improve the ability to accurately identify unsuitable applicants. The application of an SJT as a screen-out test must take into account the high base rate of suitable applicants (Niessen & Meijer, 2016) and the low prevalence of unprofessional behaviour (Norman, 2015). Future research to indicate the precise use of the SJT in medical selection procedures is necessary.

Conclusions

The combination of a traditional inductive and an innovative deductive development approach resulted in an Integrity SJT which had convergent and discriminant validity. Categorising the response options of the SJT according to two established theoretical models – one positively and one negatively related to integrity – resulted in a wide range of appropriate (HH-based) and inappropriate (CD-based) response options. The CD-based SJT score had stronger convergent validity than the HH-based SJT score. It may be promising to focus SJTs on the ability to correctly identify inappropriate response options (i.e. what one should not do).

Chapter 4

Faking on a situational judgement test in a medical selection setting: Effect of different scoring methods?

This chapter has been published as:
De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2019). Faking on a situational judgment test in a medical school selection setting: Effect of different scoring methods? *International Journal of Selection and Assessment, 27*, 235-248.

Abstract

We examined the occurrence of faking on a rating situational judgement test (SJT) by comparing SJT scores and response styles of the same individuals across two naturally occurring situations. An SJT for medical school selection was administered twice to the same group of applicants (N = 317) under low-stakes (T1) and high-stakes (T2) circumstances. The SJT was scored using three different methods that were differentially affected by response tendencies. Applicants used significantly more extreme responding on T2 than T1. Faking (higher SJT score on T2) was only observed for scoring methods that controlled for response tendencies. Scoring methods that do not control for response tendencies introduce systematic error into the SJT score, which may lead to inaccurate conclusions about the existence of faking.

Introduction

The predictive validity evidence on situational judgement tests (SJTs) in personnel selection stimulated the introduction of SJTs in educational selection settings. SJTs instruct individuals to judge the appropriateness of potential response options to challenging situations (Weekley & Ployhart, 2006). These dilemma-like situations take place in the context of the organisation or the educational programme for which an individual applies. Generally, SJTs are used to measure noncognitive attributes. SJTs demonstrate sufficient criterion-related validity in personnel selection (McDaniel et al., 2007) and educational admissions (Lievens et al., 2005a). Additionally, SJTs have incremental validity over traditional cognitive predictors such as high-school grade point average (GPA) (Schmitt et al., 2009). Finally, SJT scores show smaller socioeconomic group differences than traditional predictors (Lievens et al., 2016).

Parallel to other noncognitive measures, concerns have been raised about faking on SJTs (Weekley & Ployhart, 2006). Faking is defined as conscious response distortion in order to make a favourable impression and to increase the chance of getting hired (Goffin & Boyd, 2009). Concerns about faking on noncognitive measures are a consequence of the use of self-report formats that are prone to faking.

Faking on personality measures

Faking in high-stakes selection settings has been extensively investigated on personality measures. Considerable research has been devoted to answering the research questions "can people fake?" and "do people fake?" (Cook, 2016). Regarding the first question, studies that instructed respondents to deliberately "fake good" demonstrated that most people can increase their personality scores (McFarland & Ryan, 2000; Viswesvaran & Ones, 1999). Regarding the second question, studies comparing the personality test scores of incumbents and applicants found more desirable scores for applicants, indicating that people do fake in high-stakes settings (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Rosse, Stecher, Miller, & Levin, 1998). Both questions have been addressed in between-subjects and within-subjects study designs. Between-subjects designs compare the personality scores of two groups that receive different instructions (e.g. fake or respond honestly) or are from distinct settings (e.g. applicant or incumbent), whereas within-subjects designs compare different instructions or settings within the same individual. The main advantage of within-subjects over between-subjects designs is the possibility to control for existing group differences which may confound the score differences (Donovan et al., 2014). A disadvantage of within-subjects designs in real-life settings is the difficulty to control for order effects, because counterbalancing is often not feasible, and for other retest effects (e.g. caused by practice effects or less test anxiety). However, retest effects on noncognitive instruments are generally viewed as a result of faking (Landers, Sackett, & Tuzinski, 2011; Van Iddekinge & Arnold, 2017). Overall, selection settings drive individuals to convey desirable impressions of themselves, but individuals may differ in their tendency to fake. These individual differences have been described in various models of applicant faking (Goffin & Boyd, 2009; McFarland & Ryan, 2000; Mueller-Hanson, Heggestad, & Thornton, 2006; Roulin et al., 2016).

Consequences of faking

Although researchers reached considerable consensus with respect to peoples' ability and willingness to fake, differing perspectives exist on the influence of faking on the construct

and predictive validity of personality measures. One perspective considers the influence of faking on the predictive validity of personality measures to be negligible, calling the concerns on social desirability in the use of personality tests a "red herring" (Ones et al., 1996). Other studies have indicated that faking does not affect the construct validity (Ones & Viswesvaran, 1998) or the factor structure of personality measures (Hogan, Barrett, & Hogan, 2007). Additionally, Ingold et al. (2015) demonstrated a positive relation between faking and job performance and thus proposed that faking should be viewed as socially adequate behaviour. In contrast, the other perspective regards faking as detrimental to the use of personality measures for selection purposes, because faking affects the rank order of the applicants and reduces the quality of hiring decisions (Donovan et al., 2014; Griffith, Chmielowski, & Yoshita, 2007). In addition, concerns have been raised about the adverse effect of faking on the construct validity (Rosse et al., 1998) and criterion-related validity (Morgeson et al., 2007; Mueller-Hanson, Heggestad, & Thornton, 2003) of personality test scores. So far, no consensus has been reached regarding the consequences of faking on personality measures.

Measures against faking

Several studies investigated approaches to deal with faking on personality measures. First, warning respondents about the potential identification and consequences of faking resulted in lower personality scores than not warning respondents (Dwight & Donovan, 2003). However, warnings may also reduce the convergent validity of a personality measure (Robson, Jones, & Abraham, 2007). Second, faking has been tackled by correcting personality test scores for the score on a faking measure (e.g. a social desirability scale) (Goffin & Christiansen, 2003; Schmitt & Chan, 2006). The success of this approach is often limited due to the poor construct validity of faking measures, as the variance in faking measures is often not only explained by faking, but also by personality test scores and the criterion (Cook, 2016; Griffith & Peterson, 2008; Schmitt & Chan, 2006). Finally, another approach to reduce the influence of faking is the use of forced-choice response formats, forcing respondents to choose between equally desirable responses (Jackson, Wroblewski, & Ashton, 2000; O'Neill et al., 2017). A disadvantage of forced-choice response formats is their ipsative nature which impedes the comparison of applicants, because the total score is equal for each applicant (Heggestad, Morrison, Reeve, & McCloy, 2006). However, one can perform interindividual comparisons through partially ipsative measurement using scoring formats that allow total score variability (Heggestad et al., 2006). To summarise, research on the effectiveness of various approaches to deal with faking on personality measures has mixed results.

Faking on SJTs

Unlike the extended research on faking on personality tests, the number of published studies on faking on SJTs is limited (Table 1). As with personality tests, lab studies showed that individuals are able to obtain higher SJT scores if they are instructed to fake (Lievens & Peeters, 2008; Nguyen et al., 2005; Oostrom et al., 2017; Peeters & Lievens, 2005). The size of the faking effects seems to depend on the order in which the fake and honest conditions are presented to the respondent. On a would-do SJT (i.e. which asks respondents what they would actually do), Nguyen et al. (2005) found a larger effect size when respondents received the instructions to respond honestly first ($d = 0.34$) than when respondents received the faking instructions first ($d = 0.15$). In contrast, Oostrom et al. (2017) found a larger faking effect size on a would-do SJT when faking instructions preceded honest instructions ($d = 1.09$) than vice versa ($d = 0.82$). A faking effect on a should-do SJT (i.e. which asks respondents what they should do) was only found in the fake-first condition ($d = 0.45$), whereas the reverse

Table 1
Overview of published papers on faking on Situational Judgement Tests (SJTs).

Study	Study type	Study design	Study setting	Response format	Response instruction	Results
Lievens et al. (2008)	Lab	BS	Edu	Pick one	WD	Smaller faking effect when students elaborated on their judgement, but only for items familiar to the students (elaboration condition: d =0.61 vs. non-elaboration condition: d = 1.04).
Nguyen et al. (2005)	Lab	WS	Edu	Pick two	WD & SD	Larger faking effect for behavioural tendency instruction (i.e. WD; honest condition first: d = 0.34; fake condition first: d = 0.15) than for knowledge instruction (i.e. SD; no significant difference between honest and fake conditions).
Oostrom et al. (2017)	Lab	WS	Pers	Pick two	WD, SD & OWD	Larger faking effect for WD instruction (d = 0.92) than for the SD (d = 0.71) and false consensus (i.e. OWD; d = 0.65) instructions.
Peeters et al. (2005)	Lab	BS	Edu	Pick two	WD	Higher SJT scores for students in the fake condition than in the honest condition (d = 0.89).
Ployhart et al. (2003)	Field	BS	Pers	Pick two	WD	Applicants scored significantly higher than incumbents. Larger faking effect when the applicants made the SJT as a paper-and-pencil test (d = 0.88) as opposed to a web-based test (d = 0.59).
Vasilopoulos et al. (2000)	Field	BS	Pers	Pick one	WD	Low versus high IM applicants: faking effect larger for applicants scoring higher on job familiarity (low job familiarity: d = 0.18; moderate job familiarity: d = 0.22; high job familiarity: d = 0.73).
Weekley et al. (2004)	Field	BS	Pers	Pick two	WD	Applicants scored significantly lower than incumbents (d = 0.60).

Note. Conference papers are not included Lab = Faking-induced study Field = Real-life study BS = Between-subjects WS = Within-subjects WD = Would do SD = Should do OWD = Others would do Edu = Educational Pers = Personnel IM = Impression management d = Cohen's d (effect size)

(i.e. higher SJT scores under honest instructions than under faking instructions) was found in the honest-first condition (d = -0.34) (Nguyen et al., 2005). Oostrom et al. (2017) found a faking effect in both conditions, but the effect size was much smaller in the honest-first condition (d = 0.11) than in the fake-first condition (d = 1.31). Field studies comparing existing groups of applicants and nonapplicants showed mixed results, with one study reporting better SJT performance for applicants (Ployhart et al., 2003) and another study reporting better SJT performance for nonapplicants (Weekley et al., 2004).

Several faking studies on SJTs attempted to reduce faking (Lievens & Peeters, 2008; Oostrom et al., 2017). The most common approach is asking individuals what they should do (i.e. knowledge instructions) as opposed to asking individuals what they would actually do (i.e. behavioural tendency instructions) (Nguyen et al., 2005). Knowledge instructions may reduce the influence of faking because these instructions convert the SJT to a cognitively loaded knowledge test and knowledge is difficult to fake (McDaniel et al., 2007). Although promising, knowledge instructions might not fully solve the faking issue since SJTs are not traditional knowledge tests with clear-cut right and wrong answers. In fact, the dilemma-like nature of SJT items causes even experts to disagree on the effectiveness of a response option. In addition, the meta-analysis of McDaniel et al. (2007) indicated that SJTs with knowledge instructions still have noncognitive correlates, although to a lesser extent than SJTs with behavioural tendency instructions. Moreover, the differences between both types of response instructions are not replicated in high-stakes settings, like a medical school selection setting (Lievens et al., 2009). Finally, due to the higher susceptibility to faking, behavioural tendency instructions are of limited practical value in high-stakes medical school selection and examining faking effects on SJTs using these instructions would, therefore, have little ecological validity.

Present study

This study examined the fakability of an SJT in a medical school selection setting. Prior studies in the medical education domain indicated that applicants showed more response distortion on personality tests than nonapplicants (Anglim, Bozic, Little, & Lievens, 2018; Griffin & Wilson, 2012). The current study investigates whether applicants also distort their responses to an SJT. Prior faking research on SJTs is extended in three different ways.

First, unlike the SJT studies mentioned in Table 1, this study used a within-subjects design without different instructional sets (i.e. a field study). Although previous studies have used within-subjects designs in the educational field to examine faking on personality measures (Griffin & Wilson, 2012; Niessen et al., 2017b), this is one of the first field studies using a within-subjects design to examine faking on an SJT. As mentioned above, the disadvantage of between-subjects designs is the complexity to determine if group differences are caused by faking or by existing individual differences (e.g. in job experience), especially in field studies where random assignment to applicant and nonapplicant groups is not possible. Within-subjects designs control for these individual differences. Additionally, lab studies examine whether applicants can fake, but not whether applicants actually do fake in real-life high-stakes selection settings. The present field study investigated the actual occurrence of faking by comparing the SJT scores of the same individuals across two naturally occurring situations (i.e. low-stakes and high-stakes). Although the combination of a within-subjects design and a field study will extend previous faking research on SJTs, the real-life setting of the present study does not allow counterbalancing the order of the low and high-stakes settings. Earlier exposure to an identical or comparable test may cause retest effects (Lievens, Buyse, & Sackett, 2005b). Retest effects may reflect faking, but may, for example, also

encompass practice effects, due to familiarisation with the test format (Hooper et al., 2006). The present study examined retest effects using a between-subjects analysis comparing the SJT score of first-time test takers to second-time test takers (Lievens et al., 2005b).

Second, this study investigated differences in faking between desirable and undesirable response options because prior research proposed that there might be differences in the extent to which positive traits are exaggerated and unflattering traits are de-emphasised (Goffin & Boyd, 2009). A comparison of desirable and undesirable response options was also performed because previous research has indicated that SJT scores based on desirable items have lower construct and predictive validity than SJT scores based on undesirable items (De Leng, Stegers-Jager, Born, & Themmen, 2018; Elliott et al., 2011; Stemler et al., 2016). Stronger validity for undesirable than desirable response options is possibly a result of larger consensus on what not to do than on what to do in challenging situations (Stemler et al., 2016). A survey regarding faking behaviours during job applications revealed that the proportion of respondents indicating to de-emphasise negative traits was larger than the proportion of respondents indicating to exaggerate positive characteristics (Donovan et al., 2003). Accordingly, we hypothesised the following:

Hypothesis 1 The influence of faking on SJT scores will be more pronounced for undesirable than for desirable response options.

Third, the present study examined faking on an SJT that uses a rating format as opposed to a pick-one or pick-two format (e.g. most and least likely to perform). To our knowledge, no prior faking studies have been published on a rating SJT (Table 1). A rating SJT enables the investigation of faking not only by examining differences in mean scores but also in extreme responding on the rating scale. Since prior research demonstrated a positive relationship between faking and extreme responding (Van Hooft & Born, 2012), we formulated the following hypothesis.

Hypothesis 2a Applicant use more extreme responding in a high-stakes than in a low-stakes setting.

Whether differences in extreme responding relate to differences in the SJT score is likely to depend on the method used for scoring the SJT. Of the many SJT scoring methods that exist (De Leng et al., 2017), most use consensus judgement to determine the scoring key (McDaniel et al., 2011) and calculate the distance on the rating scale between an individual's judgement and the consensus judgement. Prior research has demonstrated that these scoring methods may be affected by response tendencies (e.g. extreme response style), introducing a source of systematic error, which may decrease the criterion-related validity of an SJT (McDaniel et al., 2011; Weng et al., 2018). In the present study, we examined how faking (i.e. higher SJT score in a high-stakes setting than in a low-stakes setting) is influenced by three different scoring methods that are differentially affected by response tendencies in the use of a rating scale. Based on previous findings, we formulated the following hypothesis.

Hypothesis 2b More extreme responding is related to a larger score difference between low-stakes and high-stakes settings for a scoring method that is more strongly affected by response tendencies (henceforth: a scoring method that does not control for response tendencies).

Finally, as an additional exploratory test, we examined whether a scoring method controlling for response tendencies had stronger construct validity than a scoring method not controlling for response tendencies (Weng et al., 2018). We expect that the systematic error introduced by response tendencies will lower the construct validity of scoring methods not controlling for response tendencies.

Methods

Context and procedure

This study was conducted at a Dutch medical school, where the selection was based on pre-university GPA, extracurricular activities and three cognitive tests on mathematics, logical reasoning, and a video lecture. Three months before the selection testing day, applicants had the opportunity to participate in a selection orientation day, where they received information about the selection procedure. Participation in the selection orientation day was voluntary and free of charge. The same SJT scenarios were administered twice – on the 2017 selection orientation day (T1) and on the 2017 selection testing day (T2) (interval: three months). On both occasions, the SJT was administered for research purposes only and participation was voluntary. However, the stakes were higher on T2 as the SJT was administered among the admission tests for which test performance did determine the selection outcome. Because the selection context was more obviously present on T2, it was expected that applicants would be more motivated to fake on T2. Applicants were informed that their answers would not influence the selection decision, because ethical regulations precluded misleading the applicants about the true purpose of the SJT administration. Applicants were asked to sign an informed consent form before participation. The data in this study were processed confidentially.

Participants

The T1 sample consisted of 362 applicants (73.5% females) and was on average 18.55 years old (SD = 2.38). The T2 sample consisted of 591 applicants (69.5% females) and was on average 18.96 years old (SD = 2.25). In total, 317 applicants were present in both samples (74.4% females). On T2, the average age of this overlapping group was 18.75 years (SD = 2.46). On T1, the sample that only provided data on T1 (N = 45) was comparable to the sample that provided data on T1 and T2 with respect to age ($t(360) = 0.56$, $p > .05$) and gender ($X^2(1) = 1.22$, $p > 0.05$). On T2, the sample that provided data on T1 and T2 was significantly younger ($t(589) = 2.43$, $p = .015$, $d = 0.20$) and consisted of significantly more females ($X^2(1) = 7.77$, $p = .005$, $\varphi = 0.11$) than the sample that only provided data on T2 (N = 274). The results of this study were based on the overlapping group (N = 317).

Situational judgement test

The SJT was designed to measure integrity and was developed using a combination of critical incident interviews and two established theoretical models. The first model comprised the honesty-humility dimension of the HEXACO personality inventory. Unlike the well-known Big Five personality dimensions, the HEXACO assumes six dimensions of personality: honesty-humility, emotionality, extraversion, agreeableness, conscientiousness, and openness to experience (Lee & Ashton, 2004). The sixth factor, honesty-humility, is defined as "sincere, fair and unassuming versus sly, greedy and pretentious" (Ashton & Lee, 2005, p. 1324) and is positively related to integrity (Lee et al., 2005). The desirable response options of the SJT were written based on three of the four facets of the honesty-humility dimension

(i.e. sincerity, fairness, and modesty). The fourth facet, greed avoidance, was not used because this facet was considered less relevant for medical school applicants. The second model comprised the cognitive distortions measured by the How I Think questionnaire (Barriga & Gibbs, 1996). Self-serving cognitive distortions are inaccurate thinking styles that may lead to the violation of social norms (Nas et al., 2008) and are, therefore, negatively related to integrity. The undesirable response options of the SJT were written based on the four categories of cognitive distortions (i.e. self-centeredness, blaming others, minimising, and assuming the worst). See De Leng et al. (2018) for an extensive description of the development of the integrity SJT. The construct validity of the SJT was demonstrated by the significant correlations with four external integrity-related measures on honesty-humility, cognitive distortions, counterproductive academic behaviour, and workplace deviance (De Leng et al., 2018). On T1, the SJT consisted of 10 scenarios, each followed by four response options (two desirable and two undesirable) that had to be judged on a six-point rating scale (1: *Very inappropriate* - 6: *Very appropriate*). On T2, the same 10 SJT scenarios were administered plus 21 additional scenarios. Two example items of the integrity SJT can be found in Appendix 4A. Appendix 4B shows the intercorrelations between SJT scores and the other variables collected during the selection procedure.

Scoring methods

The SJT was scored using three methods that were differently affected by response tendencies in the use of a rating scale. First, the raw consensus scoring method calculated the absolute distance on the rating scale between an applicant's judgement and the average judgement of a group of subject matter experts (SMEs). The SMEs were residents in training to become general practitioners. The size of the SME group ranged between 18 and 23. The characteristics of the SME sample were described in De Leng et al. (2018). Distances were summed across the response options to obtain a scale score. Scale scores based on raw consensus were reverse coded (i.e. subtracted from the maximum possible score), for higher scores to indicate better SJT performance. For raw consensus scoring methods, extreme responding generally relates to lower SJT scores because more extreme ratings result in larger deviations from the scoring key (Weng et al., 2018). Second, the standardised consensus scoring method calculated the absolute distance between the applicant's judgement and the average SMEs' judgement, but first performed a within-person *z* standardisation such that each respondent has a mean of zero and a standard deviation of one across the items (McDaniel et al., 2011). Like the raw consensus scoring method, distances were summed across the response options and the scale scores were reverse coded. Standardised consensus scoring methods control for response tendencies and should, therefore, not be affected by extreme responding. Third, the dichotomous consensus scoring method divided the rating scale in half. Applicants received one point if their judgement was located on the same side of the rating scale as the average judgement across the SMEs (McDaniel et al., 2011). Otherwise, applicants received no points. The dichotomous consensus scoring method is not affected by response tendencies because it does not matter whether an applicant's judgement is located at the extremes or near the midpoint of the rating scale. The distribution of the SMEs' judgements across the rating scales of the SJT items is presented in Appendix 4C. Internal consistency reliability estimates for all scoring methods are reported in Table 2. The SJT scores based on all response options showed sufficient to good reliability and the SJT subscores based on only the desirable or undesirable response options showed poor to sufficient reliability. Caution in the interpretation of these estimates is warranted as research indicated that internal consistency may be an unsuitable reliability estimate due to the multidimensional nature of SJTs (Whetzel & McDaniel, 2009). The multidimensional nature

was verified in a principal component analysis of the SJT, which revealed an uninformative component structure.

Personality
The HEXACO simplified personality inventory (HEXACO-SPI) (De Vries & Born, 2013) was administered online after the selection testing day (T2), but before applicants received the admission decision. The honesty-humility dimension of the HEXACO-SPI was used to examine the construct validity of the integrity SJT. The honesty-humility subscale consisted of 16 items (e.g. "I find it hard to lie") which need to be judged on a five-point rating scale (1: *Strongly disagree* – 5: *Strongly agree*). The internal consistency reliability of the honesty-humility subscale was sufficient ($\alpha = 0.74$). Participation in the online administration of the HEXACO-SPI was voluntary and did not affect the admission decision. Respondents were informed that their answers would not affect the admission decision and signed informed consent before participation. Among the applicants who provided SJT data on T1 and T2 (N = 317), 171 responded to the personality questionnaire. The responders were comparable to the nonresponders with regard to gender ($X^2(1) = 0.36$, $p = .551$) and age ($t(315) = -0.51$, $p = .610$). Responders had a significantly higher pu-GPA than nonresponders ($t(239) = -2.08$, $p = .039$, $d = 0.27$). Additionally, responders obtained a significantly higher standardised consensus score on T2 than nonresponders for all response options ($t(315) = -3.21$, $p = .002$, $d = 0.37$), for desirable response options ($t(315) = -2.79$, $p = .006$, $d = 0.31$) and for undesirable response options ($t(315) = -2.87$, $p = .004$, $d = 0.32$). Responders also obtained a significantly higher dichotomous consensus score on T2 than nonresponders for all response options ($t(315) = -2.49$, $p = .016$, $d = 0.28$) and undesirable response options only ($t(315) = -2.64$, $p = .010$, $d = 0.29$), indicating that the SJT score had a positive but weak association with the voluntary participation in the online administration of a personality inventory. Responders and nonresponders did not significantly differ in the raw SJT scores, the standardised and dichotomous SJT scores on T1 and the dichotomous SJT score based on the desirable response options on T2.

Results

Mean differences
The mean raw consensus SJT score was significantly lower (worse) on T2 than T1 (Table 2), $t(316) = 3.82$, $p < .001$, $d_{RM} = 0.23$ (small effect). The effect size for repeated measures (d_{RM}) was calculated using the method described by Morris and DeShon (2002). A comparable raw consensus SJT score was found for desirable response options ($t(316) = 1.41$, $p = .161$, $d_{RM} = 0.11$). For undesirable response options, respondents obtained a significantly lower score on T2 than T1, $t(316) = 4.88$, $p < .001$, $d_{RM} = 0.28$ (small effect). On the contrary, the mean standardised consensus SJT score was significantly higher (better) on T2 than T1 for all response options ($t(316) = -4.45$, $p < .001$, $d_{RM} = -0.25$, small effect), for desirable response options ($t(316) = -4.72$, $p < .001$, $d_{RM} = -0.27$, small effect), and for undesirable response options ($t(316) = -2.59$, $p = .010$, $d_{RM} = -0.15$). The T1-T2 difference in the dichotomous consensus SJT score was also significant, $t(316) = -5.01$, $p < .001$, $d_{RM} = -0.28$ (small effect) with a higher (better) SJT score on T2 than on T1. In addition, a significantly higher score on T2 than T1 was found for desirable response options ($t(316) = -8.25$, $p < .001$, $d_{RM} = -0.46$, medium effect), but not for undesirable response options ($t(316) = -0.73$, $p = .469$, $d_{RM} = -0.04$). Thus, a faking effect (i.e. higher score in a high-stakes than in a low-stakes setting) was detected for the standardised and dichotomous consensus scoring method, but not for the

raw consensus scoring method. In contrast to Hypothesis 1, the faking effect on the standardised and dichotomous SJT scores was larger for desirable response options than undesirable response options.

Table 2

Average raw, standardised and dichotomous consensus SJT scores and average percentage of items judged with the extreme rating scale points on T1 and T2 based on all response options, the desirable response options and the undesirable response options.

	α_{T1}	T1	α_{T2}	T2
Raw consensus				
All response options	.74	**121.94 (7.14)**	.81	**120.22 (7.32)**
Desirable response options	.52	61.74 (3.88)	.70	61.39 (3.73)
Undesirable response options	.65	**60.20 (4.19)**	.73	**58.63 (4.75)**
Standardised consensus				
All response options	.81	**84.31 (4.42)**	.83	**85.33 (4.11)**
Desirable response options	.66	**41.71 (2.55)**	.67	**42.38 (2.13)**
Undesirable response options	.75	**42.60 (2.46)**	.79	**42.95 (2.53)**
Dichotomous consensus				
All response options	.72	**36.38 (2.98)**	.83	**37.21 (2.93)**
Desirable response options	.42	**17.62 (1.60)**	.66	**18.37 (1.35)**
Undesirable response options	.72	18.76 (1.88)	.81	18.84 (2.07)
% Extreme responding				
All response options		**52.50 (21.17)**		**61.41 (25.26)**
Desirable response options		**55.32 (21.53)**		**65.78 (25.12)**
Undesirable response options		**49.78 (25.41)**		**57.12 (29.52)**

Note. T1 = selection orientation day (low motivation-to-fake context) T2 = selection testing day (high motivation-to-fake context) α_{T1} = alpha coefficients on T1 α_{T2} = alpha coefficients on T2 Standard deviations between brackets Bold numbers indicate significant T1-T2 difference

Extreme responding

Extreme responding was measured by the percentage of extreme rating scale points (i.e. 1 or 6) from the total number of rating scale points. In line with Hypothesis 2a, the use of extreme rating scale points was significantly higher on T2 than T1, $t(316) = -7.36$, $p < .001$, $d_{RM} = -0.46$ (medium effect; Table 2). A significantly higher percentage of extreme ratings was found for both desirable ($t(316) = -8.10$, $p < .001$, $d_{RM} = -0.50$, medium effect) and undesirable response options ($t(316) = -5.24$, $p < .001$, $d_{RM} = -0.32$, small effect).

For each item, we calculated the distance between an individual's position on the rating scale and the outer rating scale point to compare the opportunity to fake for desirable and undesirable response options (Pelt, Van der Linden, & Born, 2017). For desirable items, the distance was calculated from the rating scale point representing "very appropriate" (6). For undesirable items, the distance was calculated from the rating scale point representing "very inappropriate" (1). The opportunity to fake for desirable ($M = 16.72$, $SD = 9.32$) and undesirable response options ($M = 16.85$, $SD = 10.72$) was comparable ($t(316) = -0.25$, $p = .804$, $d_{RM} = 0.02$). So, the difference between desirable and undesirable response options in extreme responding was not explained by a difference in the opportunity to fake.

Association between mean differences and extreme responding

The association between the mean score differences and extreme responding was examined by correlating the T1-T2 difference in the percentage of extreme rating scale points (ERS difference) to the T1-T2 difference in SJT scores (Table 3). The raw consensus SJT score difference was significantly and negatively correlated to the ERS difference, indicating that an increase in extreme responding was associated with a decrease in the SJT score. Significant negative correlations between the ERS difference and the raw consensus score difference were also found for the desirable and undesirable response options. Conversely, the standardised and dichotomous consensus SJT score differences were significantly and positively correlated to the ERS difference, indicating that an increase in extreme responding was associated with an increase in the SJT score. The absolute correlation between the ERS difference and the score difference based on undesirable response options was significantly larger for the raw than for the dichotomous consensus scoring method, $t(316) = -2.50$, $p = .013$. Williams' test was used to test the difference between two dependent correlation coefficients (Steiger, 1980). No significant difference between raw and dichotomous consensus in the absolute correlation between the ERS difference and the score difference was found for the score based on all response options ($t(316) = 1.72$, $p = .087$) or on desirable response options ($t(316) = 0.89$, $p = .370$). Additionally, the correlation between the ERS difference and the standardised consensus score difference was significantly stronger than the correlation between the ERS difference and the dichotomous consensus score difference for all response options ($t(316) = 2.42$, $p = .016$) and undesirable response options ($t(316) = 2.03$, $p = .043$), but not for desirable response options ($t(316) = 1.42$, $p = .160$). In addition, the correlation between the ERS difference and the raw consensus score difference was not significantly different from the correlation between the ERS difference and the standardised consensus score difference for all response options ($t(316) = 0.37$, $p = .710$), desirable response options ($t(316) = -0.15$, $p = .880$), and undesirable response options ($t(316) = 1.10$,

Table 3

Correlation between the T1-T2 difference in extreme responding and the T1-T2 difference in SJT scores for three scoring methods and for all, desirable and undesirable response options.

	Difference % extreme responding T1-T2
Raw consensus	
All response options	**-.47**
Desirable response options	**-.37**
Undesirable response options	**-.43**
Standardised consensus	
All response options	**.45**
Desirable response options	**.37**
Undesirable response options	**.37**
Dichotomous consensus	
All response options	**.38**
Desirable response options	**.32**
Undesirable response options	**.30**

Note. T1 = selection orientation day (low motivation-to-fake context) T2 = selection testing day (high motivation-to-fake context) Bold coefficients indicate a significant correlation ($p < .001$, two-tailed)

p = .270). Thus, more extreme responding on T2 than T1 related to lower SJT scores when using the raw consensus scoring method and related to higher SJT scores when using the standardised and dichotomous consensus scoring method. In other words, a faking effect was only detected when using a scoring method that controls for response tendencies. A stronger influence of extreme responding on the T1-T2 score difference for the raw consensus scoring method was solely found in comparison with the dichotomous consensus scoring method for the undesirable response options, only partially confirming Hypothesis 2b.

Construct validity

As expected, the correlation of the raw consensus SJT scores with honesty-humility was smaller than the correlation of the standardised and dichotomous SJT scores with honesty-humility (Table 4). The standardised consensus score based on all response options had a significant and positive correlation to honesty-humility on T1 (r = .17, p = .029) and T2 (r = .24, p = .001). For the dichotomous consensus scoring method, the overall SJT score was also significantly and positively correlated to honesty-humility, but only on T2 (r = .25, p = .001). The SJT score based on undesirable response options correlated significantly and positively to honesty-humility on both T1 and T2 for the standardised consensus scoring method (r_{T1} = .22, p_{T1} = .004 and r_{T2} = .30, p_{T2} < .001) and for the dichotomous consensus scoring method (r_{T1} = .17, p_{T1} = .023 and r_{T2} = .33, p_{T2} < .001). For the standardised and dichotomous consensus score based on desirable response options, no significant correlations to honesty-humility were found on T1 or T2. Stronger construct validity for undesirable than desirable response options was in line with our expectations based on the previous research (De Leng et al., 2018; Elliott et al., 2011; Stemler et al., 2016).

Table 4

Correlation to honesty-humility for three scoring methods and for all, desirable and undesirable response options on T1 and T2.

		Scoring method		
	Response options	Raw consensus	Standardised consensus	Dichotomous consensus
T1	All	-.01	**.17**	.13
	Desirable	.01	.07	.04
	Undesirable	-.03	**.22**	**.17**
T2	All	-.12	**.24**	**.25**
	Desirable	-.14	.08	.01
	Undesirable	-.08	**.30**	**.33**

Note. T1 = selection orientation day (low motivation-to-fake context) T2 = selection testing day (high motivation-to-fake context) Bold coefficients indicate a significant correlation, two-tailed, p < .05

Retest effects

Finally, retest effects were investigated as an alternative or complementary explanation for the T1-T2 differences because the real-life setting of the present field study prevented counterbalancing the order of the low and high-stakes settings. Possible retest effects were examined by comparing the T2 SJT score for repeat test takers (i.e. applicants who also

participated on T1) and novel test takers (i.e. applicants who did not participate on T1) (cf. Lievens et al., 2005b). For the raw consensus scoring method, repeat test takers did not significantly differ from novel test takers in the SJT score on T2 (Table 5). For the standardised consensus scoring method, a significant difference was found for the overall SJT score ($t(589) = -3.28$, $p = .001$, $d = 0.27$, small effect), desirable response options ($t(589) = -2.98$, $p = .003$, $d = 0.24$, small effect), and undesirable response options ($t(589) = -2.93$, $p = .004$, $d = 0.24$, small effect), all in favour of repeat test takers. In addition, a significant difference favouring repeat test takers was found for the dichotomous consensus score based on all response options ($t(589) = -3.23$, $p = 0.001$, $d = 0.27$, small effect), desirable response options ($t(589) = -2.87$, $p = .004$, $d = 0.24$, small effect), and undesirable response options ($t(589) = -2.73$, $p = .007$, $d = 0.23$, small effect). Finally, repeat test takers used significantly more extreme rating scale points on T2 than novel test takers, $t(589) = -2.44$, $p = .015$, $d = -0.20$ (small effect). Prior exposure to an SJT resulted in no retest effects when using the raw consensus score and in small retest effects when using the standardised and dichotomous consensus score. Thus, retest effects – faking or practice – were only detected for scoring methods that controlled for response tendencies.

Table 5
Average SJT scores and extreme responding on T2 for repeat test takers (participation on T1) and novel test takers (no participation on T1).

	Repeat test takers ($N = 317$)	Novel test takers ($N = 274$)
Raw consensus score on T2		
All response options	120.22 (7.32)	119.83 (8.54)
Desirable response options	61.39 (3.73)	61.06 (4.84)
Undesirable response options	58.83 (4.63)	58.77 (4.90)
Stan. consensus on T2		
All response options	**85.33 (4.11)**	**84.03 (5.50)**
Desirable response options	**42.38 (2.13)**	**41.76 (2.90)**
Undesirable response options	**42.95 (2.53)**	**42.27 (3.12)**
Dich. consensus score on T2		
All response options	**37.21 (2.93)**	**36.29 (3.85)**
Desirable response options	**18.37 (1.35)**	**17.98 (1.88)**
Undesirable response options	**18.84 (2.07)**	**18.31 (2.57)**
Extreme responding (%)	**61.5 (25.3)**	**56.3 (25.7)**

Note. T1 = selection orientation day (low motivation-to-fake context) T2 = selection testing day (high motivation-to-fake context) Stan. = Standardised Dich. = Dichotomous Standard deviations between brackets Bold numbers indicate a significant difference

Discussion

The present study describes a within-subjects investigation of faking on an SJT in a real-life setting. Additionally, in contrast to previous research, this study examined faking on an SJT that uses a rating response format, enabling the examination of faking through extreme

responding. Applicants used more extreme rating scale points on the high-stakes selection testing day than on the low-stakes selection orientation day, indicating that applicants responded differently to the SJT during the second administration. More extreme responding relates to a T1-T2 increase in the SJT score (i.e. a faking effect) for the scoring methods that controlled for response tendencies (i.e. standardised and dichotomous consensus). Conversely, for the raw consensus scoring method, more extreme responding relates to a lower SJT score. These results suggest that statements about the existence of a faking effect on a rating SJT depend on the method used for scoring the SJT. The nonsignificant correlation with honesty-humility for the raw consensus scoring method may indicate that systematic error caused by response tendencies interferes with the construct validity of a traditionally scored SJT. In addition, our findings indicate that a raw consensus scoring method may obscure the presence of a faking effect. Finally, the faking effect seemed stronger for desirable response options than for undesirable response options.

Faking

Because the standardised and dichotomous SJT scores were not affected by systematic error due to response tendencies, we will focus on these SJT scores in the discussion below. The higher SJT scores on T2 than T1 seems to demonstrate a small faking effect for the standardised (d = -0.25) and dichotomous (d = -0.28) scoring methods, indicating that on the same SJT, the same applicants obtained a higher score in a high-stakes setting than in a low-stakes setting. The effect size is smaller than most effect sizes reported in Table 1. Unfortunately, a direct comparison with these published effect sizes is problematic because none of the previous SJT faking studies used a within-subjects design in the field (i.e. not using different instructional sets). Consequently, dissimilar effect sizes are likely caused by differences in study design and study type. Between-subjects designs may produce larger faking effects than within-subjects designs if the compared groups also differ on other variables, for example, job experience (Ployhart et al., 2003; Vasilopoulos, Reilly, & Leaman, 2000). Additionally, lab studies may generate larger effect sizes than field studies, because different instructional sets involve a stronger intervention (Birkeland et al., 2006). Another possible explanation for the smaller faking effect found in this study is that the integrity SJT used knowledge response instructions, whereas most previous SJT faking studies used behavioural tendency instructions. Two SJT faking studies that compared both response instructions (Nguyen et al., 2005; Oostrom et al., 2017) demonstrated that the faking effect is smaller for knowledge than for behavioural tendency instructions. McDaniel et al. (2007) describe SJTs with knowledge instructions as maximal performance tests and SJTs with behavioural tendency instructions as typical performance tests and argue that self-reports of typical behaviour are more susceptible to faking than self-report predictors of knowledge. Our findings seem to support the lower susceptibility to faking of SJTs with knowledge instructions, but also indicate that knowledge instructions do not completely cancel out the faking effect, presumably because SJTs are not pure knowledge tests due to their noncognitive content.

Desirable and undesirable response options

The T1-T2 increase in the SJT score based on desirable response options was significant for the standardised (d = -0.27) and dichotomous (d = -0.46) scoring method. For undesirable response options, only the T1-T2 increase in the standardised SJT score was significant (d = -0.15), albeit considerably smaller than the T1-T2 increase for desirable response options. Additionally, the T1-T2 increase in extreme responding was larger for desirable (d = -0.50) than undesirable items (d = -0.32). A possible explanation for these findings is that it might

be harder to fake on items that require the identification of what not to do than the identification of what to do, possibly because the undesirable items have greater cognitive loading than the desirable items. Prior research indicated that there appears to be more consensus on what not to do than on what to do in a challenging situation (Stemler et al., 2016). SJTs consisting of undesirable items that are unambiguously ineffective could be viewed as measures of maximum performance, whereas SJTs consisting of desirable items – for which the appropriateness is more dependent on personal style and preference – could be viewed as measures of typical behaviour. Measures of typical behaviour are assumed to be more prone to faking than measures of maximum performance (McDaniel et al., 2007). Future research is necessary to replicate our findings and to investigate the reasons of why faking might be more difficult on undesirable items.

A stronger faking effect for desirable response options was not in line with our expectations based on the survey of Donovan et al. (2003). A possible explanation for this inconsistent finding is that what respondents say they do (e.g. moderately exaggerating positive traits) is not what they actually do when they are in a high-stakes situation. In other words, it is probable that respondents fake – consciously or unconsciously – on a survey regarding faking behaviours. Social desirable responding consists of intentional faking and unconscious self-deception (Paulhus & John, 1998). Desirable items are potentially more affected by self-deception than undesirable items. An interesting avenue for future research is to unravel the influence of faking and self-deception on de-emphasising negative traits and exaggerating positive traits. Additionally, an explanation for the stronger faking effects for desirable than undesirable response options might be found in the self-discrepancy theory (Higgins, Roney, Crowe, & Hymes, 1994). The self-discrepancy theory describes that discrepancies between one's perceived actual self and one's desired self result in negative feelings (Higgins, Shah, & Friedman, 1997). The desired self may be characterised by aspirations and wishes (i.e. the ideal self) or by obligations and responsibilities (i.e. the ought self). Individuals who are predominated by the ideal self are more oriented toward approaching a desired end state, whereas individuals who are predominated by the ought self are more oriented toward avoiding an undesired end state (Higgins et al., 1994). Applicants' responses to the SJT might have been more strongly affected by the ideal self than the ought self, possibly caused by characteristics of the selection context leading to self-enhancement. To our knowledge, no previous faking studies have referred to the self-discrepancy theory, so more research is necessary to elucidate the influence of ideal and ought selves on faking positive and negative traits.

Scoring methods

A rating SJT allowed us to examine faking through extreme responding. Extreme responding is unaffected by the scoring method of the SJT, which is useful because our findings indicate that conclusions about faking heavily depend on how an SJT is scored. Extreme responding occurred more often in a high-stakes than in a low-stakes setting, which is in line with previous faking research on personality measures (Levashina, Weekley, Roulin, & Hauck, 2014; Van Hooft & Born, 2012). For a traditional raw consensus scoring method, extreme responding is related to lower scores, because it creates more distance from the consensus judgement, which is often located near the midpoint of the rating scale (Weng et al., 2018). Consequently, one coaching strategy to improve the score on a rating SJT instructs respondents to avoid the extreme responses on the rating scale (Cullen, Sackett, & Lievens, 2006). Additionally, our results indicate that a raw consensus SJT score may have weaker construct validity than a standardised or dichotomous SJT score, which is in line with previous research demonstrating lower criterion-related validity for scoring methods that do

not control for response tendencies (McDaniel et al., 2011; Weng et al., 2018). Response tendencies introduce systematic error in a rating SJT score, which may result in lower construct and criterion-related validity coefficients. In addition, findings indicate that systematic error caused by response tendencies may lead to inaccurate conclusions about faking on an SJT.

Hypothesis 2b that scoring methods that do not control for response tendencies are more strongly affected by a change in extreme responding than scoring methods that do control for response tendencies is only confirmed for the dichotomous SJT score based on undesirable response options. Apparently, controlling for response tendencies within one test administration does not reduce the influence of a change in response tendencies across test administrations. Additionally, an explanation for the significant influence of extreme responding on the dichotomous SJT score might be that, for 11 out of 40 response options, the consensus judgement was located near the midpoint of the rating scale (i.e. between 2.5 and 4.5 on a 6-point rating scale). For these ambiguous midrange items, an applicant might be close to but on the "incorrect" side of the rating scale, yielding no points. More extreme responding in the high-stakes setting would shift the applicant's judgement to the "correct" half of the rating scale, producing a higher SJT score. Weng et al. (2018) showed that the dichotomous consensus scoring method is more appropriate for non-midrange items, supporting this potential explanation.

A last notable finding was that – for the standardised and dichotomous SJT score – the construct validity was stronger on T2 than T1, possibly because applicants are familiarised with the SJT format on T2 which reduces construct-irrelevant variance due to unfamiliarity with the test format (Lievens et al., 2005b). SJTs are relatively new in admission procedures to higher education and use a test format that is quite different from test formats used by traditional admission tests. Medical school admission committees should consider acquainting applicants with the SJT format before administering it for admission purposes. Another possible explanation for the stronger correlation with honesty-humility on T2 than T1 is that applicants have faked on the personality measure, which was administered after the selection testing day, but before applicants received the admission decision. Applicants might have been motivated to fake on the personality measure, because admission was not yet certain. The stronger construct validity on T2 might, therefore, also be a result of overlapping variance caused by faking in both scores (i.e. SJT score on T2 and honesty-humility score). Finally, the larger correlation with honesty-humility on T2 might be caused by the stronger common frame of reference produced by the high-stakes selection context (Ones & Viswesvaran, 1998). Even though the stakes were lower on T1 than T2, some applicants might still have felt a tendency to fake. In contrast, a high- stakes setting may present a stronger frame of reference that is shared by all applicants. Ones and Viswesvaran (1998) emphasise the importance of standardising the test administration to generate a common frame of reference and to enhance the reliability. This explanation is supported by higher estimates of internal consistency reliability for the SJT score on T2 than T1. More research is necessary to examine which of these processes give rise to the stronger construct validity on T2 than T1.

Overall, each scoring method has pros (e.g. raw consensus scores have more variance and dichotomous consensus scores have stronger construct validity) and cons (e.g. raw consensus scores rely on suboptimal difference scores and dichotomous consensus scores may neglect relevant variance), that must be taken into account when using SJTs in selection settings (see De Leng et al. (2017) for an overview).

Faking versus retest effect

Due to the real-life setting of this study, the order of the selection orientation day and selection testing day could not be counterbalanced. We examined the possibility of a retest effect as an alternative explanation by comparing the SJT scores of first-time and second-time test takers (cf. Lievens et al., 2005b). The significantly higher score for second-time test takers (d = 0.27) provides evidence for a small retest effect when using the standardised or dichotomous consensus scoring method, which corresponds to previous research on retest effects on SJTs (Dunlop, Morrison, & Cordery, 2011; Lievens et al., 2005b). Retest effects could represent faking, but could also represent a practice effect or actual improvement in the relevant construct (Hooper et al., 2006). The stronger construct validity on T2 than on T1 provides some support for a practice effect caused by familiarisation with the SJT format. However, studies on retest effects involve multiple similar test administrations, whereas in the present study, the SJT is deliberately administered across two dissimilar test conditions. It is probable that the T1-T2 increase in the standardised and dichotomous SJT score is partially caused by both a faking and a practice effect. Future research is necessary to unravel the influence of faking and practice on score changes across low- and high-stakes conditions, for example, by ensuring that applicants are already familiar with the SJT format. Another possible method for disentangling the sources of the T1–T2 score change involves administering the SJT twice under the same conditions to establish a baseline for the retest effect (Ellingson, Sackett, & Connelly, 2007). Despite the problems with disentangling the causes of the T1-T2 score difference, our findings do indicate that retest effects – faking or practice – are obscured when scoring a rating SJT with a method that does not control for response tendencies.

Implications for future research and practice

First, we recommend future investigations of faking or retest effects on rating SJTs to use scoring methods that control for response tendencies. Examples of other scoring methods that control for response tendencies are mode consensus or proportion consensus (Weng et al., 2018). Second, research on the consequences of faking for the construct validity of personality measures should take into account the influence of response tendencies (i.e. extreme responding) and scoring methods. Our findings indicate that response tendencies and scoring methods might be contributing factors to the mixed evidence concerning the influence of faking on the construct validity. Third, we advise researchers to make a distinction between desirable and undesirable response options as this may affect the conclusions on SJT faking. The distinction between desirable and undesirable items can be based on empirical data (Stemler et al., 2016) or on a theoretical framework (De Leng et al., 2018). Practitioners of SJTs are also recommended to use scoring methods that control for response tendencies and undesirable response options because these modifications may increase the construct and criterion-related validity of the SJT.

Limitations

This study is not without limitations. First, the main limitation of this study is the inability to rule out other possible sources of a retest effect. A between-subjects analysis comparing the SJT scores of novel and repeat test takers indicated a retest effect of similar size as the faking effect. The investigation of retest effects on noncognitive instruments has primarily interpreted these effects as a result of applicant faking (Van Iddekinge & Arnold, 2017). However, retest effects may have many different causes, such as practice effects, genuine improvement in the construct, reduction in test anxiety, or test familiarisation (Lievens et al., 2005b; Van Iddekinge & Arnold, 2017). Even though the retest effect found in the current

study is likely produced by faking as T1 and T2 were deliberately chosen to have substantial contextual differences, it is not feasible to exclude other potential sources of a retest effect. Future studies should use research designs that allow the separation of these different sources. Second, the scoring methods used rely on the difference between a respondent's rating and the average rating across a group of SMEs. Difference scores, however, have several limitations, such as low reliability, reduced effect sizes, and loss of information from the separate component scores (Edwards, 2001). The limitations of the raw consensus scoring method were confirmed in the present study as shown by obscured faking or retest effect and the weak construct validity. The standardised and dichotomous consensus scoring methods solved some of the problems of the raw consensus scoring method. Nonetheless, future research is advised to examine polynomial regression methods as an alternative method for scoring SJTs, because these methods provide a more direct solution to the problems of difference scores (Edwards, 2001; Kulas, 2013).

Third, due to the real-life setting of this study, we investigated faking using only one order of conditions: low-stakes on T1 and high-stakes on T2. Other within-subjects studies on faking mainly use the reversed order, i.e. high-stakes among applicants on the first occasion and low-stakes among incumbents on the second occasion. We believe that the low-stakes-first order of the current study has some important benefits. First, because most T1 respondents were also present on T2, it is unlikely that our findings are affected by a restriction of range. Second, because our T2 respondents were not medical school incumbents, it is unlikely that T1-T2 score differences are caused by experience at medical school. The within-subjects field study by Ellingson et al. (2007) examined response distortion on a personality inventory using both orders and found a larger score change for the low-stakes- first condition than for the high-stakes-first condition. In contrast, within-subjects studies using directed-faking instructions demonstrated a faking effect on a should-do SJT for the fake-first condition, but not for the respond-honestly-first condition (Nguyen et al., 2005; Oostrom et al., 2017). A faking effect observed only in the fake-first condition was explained by respondents' tendency to respond deliberately different during the second condition after they responded to the best of their ability during the first condition. The tendency to respond differently might be less strong in the current study because no directed faking instructions were used and due to the longer time period between both conditions than in the previous studies. Nonetheless, these contrasting findings require more research on the effect of the order of the low- and high-stakes settings.

Fourth, during both test administrations, the SJT was administered for research purposes only, which might reduce the generalisability of our findings to real selection settings. However, Niessen et al. (2017b) found large score differences on several noncognitive measures between a research and an admission context, even though applicants were informed that the noncognitive measures were not used for selection. Additionally, the difference in extreme responding indicated that the applicants responded differently on T2. The selection testing day, therefore, appears to be a sufficient proxy of a high-stakes situation.

Finally, it might be too simplistic to assume that faking is limited to extreme responding (König, Merz, & Trauffer, 2012). Moreover, prior research has demonstrated that response styles differ across individuals (Ziegler, 2015) and cultures (He, Bartram, Inceoglu, & Van de Vijver, 2014). Further research is required to examine how other response styles apart from extreme responding relate to faking.

Chapter 5

Influence of response instructions and response format on applicant perceptions of a situational judgement test for medical school selection

This chapter has been published as:
De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2018). Influence of response instructions and response format on applicant perceptions of a situational judgement test for medical school selection. *BMC Medical Education, 18*, 282.

Abstract

This study examined the influence of two Situational Judgement Test (SJT) design features (response instructions and response format) on applicant perceptions. Additionally, we investigated demographic subgroup differences in applicant perceptions of an SJT. Medical school applicants (N = 372) responded to an online survey on applicant perceptions, including a description and two example items of an SJT. Respondents randomly received one of four SJT versions (should do-rating, should do-pick-one, would do-rating, would do-pick-one). They rated overall favourability and items on four procedural justice factors (face validity, applicant differentiation, study relatedness and chance to perform) and ease-of-cheating. Additionally, applicant perceptions were compared for subgroups based on gender, ethnic background and first-generation university status. Applicants rated would-do instructions as easier to cheat than should-do instructions. Rating formats received more favourable judgements than pick-one formats on applicant differentiation, study-relatedness, chance to perform and ease of cheating. No significant main effect for demographic subgroup on applicant perceptions was found, but significant interaction effects showed that certain subgroups might have more pronounced preferences for certain SJT design features. Specifically, ethnic minority applicants – but not ethnic majority applicants – showed greater preference for should-do than would-do instructions. Additionally, first-generation university students – but not non-first-generation university students – were more favourable of rating formats than of pick-one formats. Findings indicate that changing SJT design features may positively affect applicant perceptions by promoting procedural justice factors and reducing perceived ease of cheating and that response instructions and response format can increase the attractiveness of SJTs for minority applicants.

Introduction

An increasing number of medical schools implement a Situational Judgement Test (SJT) in their admission procedures (Dore et al., 2009; Fröhlich et al., 2017; Patterson et al., 2017; Schripsema et al., 2017). The growing popularity of the SJT is a result of the test's psychometric qualities, in terms of its predictive validity, incremental validity and low adverse impact, from the perspective of medical school admission committees (Patterson et al., 2016). Yet, the quality of an SJT should also be investigated from the perspective of medical school applicants, since applicant perceptions may influence test-taking motivation, test performance and applicant withdrawal (Chan & Schmitt, 1997; Schmit & Ryan, 1997). Furthermore, minority applicants may hold more negative applicant perceptions (Ryan, Sacco, McFarland, & Kriska, 2000), which could lead to adverse impact through diminished test-taking motivation and test performance. The current study examines the influence of two SJT design features, namely response instructions and response format, on applicant perceptions. Additionally, the perceptions of the SJT are compared for applicants belonging to different demographic subgroups.

SJTs require respondents to judge the appropriateness of responses to challenging situations (Weekley & Ployhart, 2006). The situations are contextualised to the setting for which an individual applies, such as medical school. In general, SJTs are added to admission procedures for the measurement of noncognitive attributes, for instance integrity and interpersonal skills (Patterson et al., 2012). Prior research has demonstrated that SJTs have predictive validity for future medical school performance (Lievens, 2013), that they have incremental validity over traditional cognitive admission instruments (Koczwara et al., 2012) and smaller ethnic and socioeconomic subgroup differences than cognitive admission tests (Lievens et al., 2016; Oswald et al., 2004).

In addition to these psychometric findings, several studies have demonstrated that medical school applicants hold favourable perceptions of SJTs (Husbands et al., 2015; Lievens, 2013; Lievens & Sackett, 2006; Luschin-Ebengreuth et al., 2015). Moreover, some studies indicated that SJTs are perceived more positively than cognitive admission tests (Lievens, 2013; Luschin-Ebengreuth et al., 2015). Favourable perceptions of SJTs are likely caused by the test content, which is closely related to the criterion domain for which an individual applies (Lievens et al., 2008). Furthermore, previous research demonstrated that certain SJT features might affect applicants' perceptions (Bauer & Truxillo, 2006). For example, Chan and Schmitt (1997) and Kanning, Grewe, Hollenberg, and Hadouch (2006) found that applicants perceived the same SJT more positively when it was administered in a video-based format than in a text-based format. Additionally, Neal et al. (2018) showed that medical students felt that an SJT with a short-answer-question or an interview response format would better reflect their future behaviour as a junior doctor than a ranked-order or a single-best-answer response format. Response formats using short-answer or interview questions received the most favourable ratings, probably because applicants believe these formats provide a good opportunity to demonstrate their knowledge, skills and abilities (Schleicher, Venkataramani, Morgeson, & Campion, 2006). No prior research has examined the influence of SJT response instructions on applicant perceptions.

The importance of applicant perceptions is evidenced by the influence of these perceptions on test-taking motivation and test performance (Chan & Schmitt, 1997) and possible applicant withdrawal (Schmit & Ryan, 1997). Additionally, prior research indicated that applicant perceptions might differ across demographic subgroups. For example, ethnic minorities tend to perceive selection procedures at large more negatively than ethnic

majorities (Chan & Schmitt, 1997; Ryan et al., 2000), possibly due to differences in cultural values and beliefs on testing (Chan, 1997) or by perceptions of stereotype threat (Ployhart, Ziegert, & McFarland, 2003), which refers to impaired test performance caused by the salience of a negative stereotype (Steele & Aronson, 1995). More negative perceptions of the admission procedure might reduce test performance through decreased test-taking motivation (Chan et al., 1997). Unfavourable perceptions of the admission process among ethnic minorities might also result in disproportionally more withdrawal from the admission procedure among ethnic minority applicants (Schmit & Ryan, 1997). Thus, if minority applicants – based on either gender, ethnic or socioeconomic background – perceive an admission procedure more negatively than majority applicants, they might also be less motivated to perform well or more inclined to withdraw from the admission procedure. Consequently, more negative applicant perceptions among minority applicants may lead to adverse impact. It is therefore crucial to examine which design features of an SJT reduce subgroup differences in applicant perceptions. We are not aware of previous studies that have focused on how response instructions and response format may influence subgroup differences in applicant perceptions of an SJT.

The dominant theoretical framework on applicant perceptions is the organisational justice theory (Gilliland, 1993). This theory has been applied to studies on applicant perceptions of selection practices for postgraduate medical training (Patterson et al., 2011) and of admission methods in higher education (Niessen, Meijer, & Tendeiro, 2017a). The organisational justice theory encompasses distributive justice, that is the fairness of the distribution of desired outcomes (e.g. admission spots in medical school) and procedural justice, referring to the fairness of procedures used to allocate desired outcomes (Gilliland, 1993). In the model of applicant reactions proposed by Gilliland, procedural justice perceptions are influenced by the formal characteristics of the selection system, like job relatedness and opportunity to perform. According to the organisational justice model, formal characteristics are influenced by test type. Therefore, the formal characteristics component was used to study the influence of SJT design features on applicant perceptions.

The aim of the present study is two-fold. Firstly, we examined the effect of the response instructions (i.e. should do or would do) and the response format (i.e. pick-one or rating) of an SJT on applicant perceptions. The influence of response instructions was examined because previous research showed that SJTs with should-do instructions are less susceptible to faking than SJTs with would-do instructions (McDaniel et al., 2007). Additionally, admission procedures that are perceived as more difficult to fake receive more favourable applicant perceptions (Schreurs, Derous, Proost, Notelaers, & Witte, 2008; Uggerslev, Fassina, & Kraichy, 2012). Therefore, we hypothesised that applicants have more positive perceptions of an SJT using should-do instructions than an SJT using would-do instructions. The influence of response format on applicant perceptions of an SJT was previously investigated by Neal et al. (2018). However, these researchers did not include a rating format in their investigation, even though this response format is commonly used by SJTs (Chan & Schmitt, 1997; Husbands et al., 2015). We expected the pick-one format to receive more favourable applicant perceptions than the rating format because applicants – at least in Western cultures – are more familiar with the use of pick-one (i.e. multiple-choice) formats in college admission, such as in cognitive ability tests (Patterson et al., 2011). Additionally, we assumed that applicants associate rating formats with self-report measures, which are prone to faking and therefore perceived less favourably.

Secondly, to determine if SJTs are perceived differently by minority applicants than majority applicants, we examined the influence of demographic variables (i.e. gender, ethnic background and first-generation university status) on applicant perceptions. Based on

previous research, we hypothesised to find no gender differences in applicant perceptions (Hausknecht et al., 2004; Niessen et al., 2017a). The meta-analysis of Hausknecht et al. (2004) indicated that the correlation between ethnic background and applicant perceptions was near zero. However, Chan (1997) found that among a US sample the predictive validity perceptions of a cognitive ability test – but not of a personality test – were significantly more favourable for White than for Black examinees. Since SJTs – like personality tests – focus on noncognitive attributes, we expected no ethnic differences in applicant perceptions. Prior research on subgroup differences in applicant perceptions has mainly focused on gender and ethnicity, but not on socioeconomic characteristics such as the educational level of the applicant's parents. Therefore, we pose the following research question: do applicant perceptions of an SJT differ across subgroups from different socioeconomic backgrounds?

Methods

Setting and procedure

This study was conducted at a Dutch medical school, whose admission procedure consisted of three equally-weighted parts: i) pre-university grade point average (pu-GPA), ii) extracurricular activities and iii) performance on three cognitive tests during an on-site testing day. Applicants with a pu-GPA $\geq$ 7.5 (on a scale from 1 (low performance) to 10 (high performance)) were directly admitted. The applicants of the 2018 admission procedure comprised the sample of this study. After the on-site testing day but before the applicants received the selection decision, applicants were invited to participate in an online survey on applicant perceptions. Participation in the survey was voluntary. Applicants were informed about the aim of the survey and that their answers would not influence the selection decision. Applicants gave informed consent before they were navigated to the survey. The data in this study were processed anonymously.

Survey

The online survey started with a questionnaire on the applicants' demographic characteristics. The demographic questions were administered online for the applicants with pu-GPA $\geq$ 7.5 and on-site for other responders. Applicants were categorised as first-generation university student, if both their parents had not attended university. The ethnic background of the applicants was categorised as Dutch if both parents were born in the Netherlands, as non-Western if at least one parent was born in Africa, Asia or South-America, or as Western if at least one parent was born in Europe (but not in the Netherlands), North-America or Oceania (Statistics Netherlands). The applicants' gender was retrieved from the student administration system.

The second part of the survey covered applicant perceptions. Applicant perceptions were measured using seven items. Firstly, overall process favourability was assessed using two items: perceived predictive validity and perceived fairness (Steiner & Gilliland, 1996). Steiner and Gilliland (1996) reported a coefficient alpha of .73 for the two process favourability items. Secondly, four items were administered measuring formal characteristics of the procedural justice dimension: i) face validity, ii) applicant differentiation (Steiner & Gilliland, 1996), iii) study relatedness and iv) chance to perform (Bauer et al., 2001). These items were selected because previous research demonstrated the influence of these formal characteristics on process favourability (Hausknecht et al., 2004; Niessen et al., 2017a; Schleicher et al., 2006). Finally, one item measuring ease of cheating (Niessen et al., 2017a) was added because a prior meta-analysis showed that ease of cheating/difficulty to fake has

a negative influence on applicant perceptions (Uggerslev et al., 2012). Each item was judged on a seven-point anchored rating scale. The items and rating scales are depicted in Appendix 5A.

The survey asked respondents to answer the seven applicant perception items separately for eleven admission instruments (CV, motivation letter, pre-university GPA, cognitive capacity test, skills test, curriculum sample test, personality questionnaire, interview, weighted lottery, unweighted lottery and SJT). The order in which the admission instruments were presented to the respondents was randomised.

Survey respondents received a short description of the SJT followed by two examples of SJT items. These example items were identical, with the exception of two design features that were manipulated. Firstly, the response instructions: the example items asked either which response should be given in the described situation (i.e. should do) or which response the respondents were most likely to perform (i.e. would do). Secondly, the response format: the example items had to be judged either by rating each separate response option (i.e. rating format) or by picking out the best response option (i.e. pick-one). In total, there were four versions of the SJT example items (i.e. should do-rating, should do-pick-one, would do-rating, would do-pick-one). Each respondent randomly received two SJT example items representing one of the four versions.

Statistical analyses

Two-way ANOVAs were used to examine the influence of SJT response instructions (should do versus would do) and SJT response format (rating versus pick one) on process favourability, the four procedural justice items (i.e. face validity, applicant differentiation, study relatedness, chance to perform) and ease of cheating. Main and interaction effects were examined. Pu-GPA $\geq$ 7.5 status (i.e. directly admitted) was included in the analyses as a control variable. Partial eta-squared was used to examine the effect sizes, where $\eta_p^2 = .01$, $\eta_p^2 = .06$ and $\eta_p^2 = .14$ indicated a small, medium and large effect, respectively (Cohen, 1988).

ANOVAs were used to examine subgroup differences (based on gender, ethnic background and first-generation university status) on the applicant perception items. Pu-GPA $\geq$ 7.5 status was again included as a control variable. In addition, the demographic variables were investigated in relation to the SJT design features by examining if the subgroup variables had an interaction effect with either the response instructions or the response format. Partial eta-squared was used to examine the effect size.

Results

Participants

In total, 872 applicants were invited to participate in the survey. Three-hundred seventy-two applicants responded to the survey (response rate = 42.7%). The average age of this group was 18.35 years (SD = 1.19) and 75.3% were women. Among the 372 respondents, 26.6% were first-generation university students, 70.2% had a Dutch ethnic background, 21.5% had a non-Western ethnic background, 8.3% had a Western ethnic background and 38.7% were directly admitted to medical school (i.e. pu-GPA $\geq$ 7.5). The group of respondents was significantly younger (18.35 vs. 18.64 years, $t(870) = 3.39$, $p = .001$, $d = 0.24$) and consisted of significantly more women (75.3% vs. 65.7%, $X^2(1) = 8.91$, $p = .003$, $\varphi = 0.10$) than the group of non-respondents, but the effect sizes were small. Respondents and non-respondents were comparable with respect to first-generation university status ($X^2(1) = 1.30$, $p = .254$) and ethnic background ($X^2(2) = 2.47$, $p = .291$).

Applicant perception items

The alpha coefficients of the two process favourability items (i.e. perceived predictive validity and perceived fairness) indicated sufficient to good internal consistency (should do-rating: $\alpha = .66$, should do-pick-one: $\alpha = .75$, would do-rating: $\alpha = .72$, would do-pick-one: $\alpha = .90$). The intercorrelations between the process favourability score (i.e. average of the two process favourability items) and the other applicant perception items are depicted in Table 1. Intercorrelations were controlled for pu-GPA $\geq$ 7.5 status (i.e. directly admitted). All intercorrelations were statistically significant. The correlations between process favourability and the procedural justice items were all above .6 (large effect size). As expected, the ease-of-cheating item correlated significantly and negatively with process favourability, but the effect size was smaller ($r = -.20$).

Table 1

Intercorrelations between overall process favourability and the other applicant perception items.

	1.	2.	3.	4.	5.
1. Process favourability					
2. Face validity	.76				
3. Applicant differentiation	.67	.69			
4. Study relatedness	.62	.64	.62		
5. Chance to perform	.63	.59	.67	.63	
6. Ease of cheating	-.20	-.24	-.20	-.26	-.24

Note. All correlations are significant, $p < .01$ (two-tailed) Correlations are controlled for pu-GPA $\geq$ 7.5 status (i.e. directly admitted)

Preliminary analysis: Comparison to other admission methods

Prior to the main analyses, we compared the overall process favourability of the SJT to the other admission methods included in the online survey, in order to determine if the SJT was perceived more or less positively than the other admission methods (Table 2). Repeated-measures ANOVAs were used to examine the differences in process favourability between the SJT and each of the other admission methods. We controlled for pu-GPA $\geq$ 7.5 status by including it as a between-subjects factor. The average process favourability rating (on a seven-point scale) ranged between 3.21 (unweighted lottery) and 5.29 (interview). The SJT was judged significantly more favourable than pu-GPA ($F(1, 364) = 7.04$, $p = .008$, $\eta_p^2 = .02$), a personality questionnaire ($F(1, 365) = 17.89$, $p < .001$, $\eta_p^2 = .05$), weighted lottery ($F(1, 365) = 67.07$, $p < .001$, $\eta_p^2 = .16$) and unweighted lottery ($F(1, 366) = 114.31$, $p < .001$, $\eta_p^2 = .24$) and significantly less favourable than a motivation letter ($F(1, 365) = 22.11$, $p < .001$, $\eta_p^2 = .06$), cognitive capacity test (F(1, 363) = 17.68, p < .001, η 2 = .05), skills test (F(1, 364) = 87.78, $p < .001$, $\eta_p^2 = .19$), curriculum sample test ($F(1, 367) = 105.17$, $p < .001$, $\eta_p^2 = .22$) and an interview ($F(1, 364) = 119.50$, $p < .001$, $\eta_p^2 = .25$). CV was judged as equally favourable as the SJT. Thus, among the other admission methods included in the online survey, the SJT takes a middle position with respect to overall process favourability.

Table 2
Comparison of the Situational Judgement Test with the other admission methods on process favourability.

	Process favourability
Situational Judgement Test	4.39 (1.28)
Curriculum vitae	4.44 (1.42)
Motivation letter	**4.76 (1.22)**
Pre-university GPA	**3.93 (1.46)**
Cognitive capacity test	**4.69 (1.15)**
Skills test	**5.11 (1.09)**
Curriculum sample test	**5.20 (1.05)**
Personality questionnaire	**4.02 (1.26)**
Interview	**5.29 (1.20)**
Weighted lottery	**3.40 (1.55)**
Unweighted lottery	**3.21 (1.82)**

Note. GPA = grade point average Bold averages indicate a significant difference from the average judgement of process favourability for the Situational Judgement Test (repeated-measures ANOVA with GPA ≥ 7.5 as between-subjects factor, $p < .01$)

Response instructions and format

Applicant perceptions of the four SJT versions are depicted in Figure 1. The mean and standard deviations corresponding to Figure 1 can be found in Appendix 5B. A significant main effect of response format was found on the applicant differentiation item ($F(1, 362) = 4.08$, $p = .044$, $\eta^2 = .01$) with a more positive judgement for the rating format ($M = 4.30$, $SD = 1.53$) than for the pick-one format ($M = 3.94$, $SD = 1.59$). Response format also had a significant influence on the study-relatedness item ($F(1, 362) = 4.23$, $p = .040$, $\eta^2 = .01$), again indicating more favourable perceptions for the rating format ($M = 3.73$, $SD = 1.33$) than for the pick-one format ($M = 3.44$, $SD = 1.41$). The rating format ($M = 3.81$, $SD = 1.52$) was also judged significantly more favourable than the pick-one format ($M = 3.42$, $SD = 1.61$) on the chance-to-perform item ($F(1, 361) = 5.16$, $p = .024$, $\eta^2 = .01$). Finally, the pick-one format ($M = 5.31$, $SD = 1.81$) was judged as significantly easier to cheat than the rating format ($M = 4.94$, $SD = 1.84$; $F(1, 362) = 5.29$, $p = .022$, $\eta^2 = .01$). Overall, an SJT with a rating response format was rated more favourably than an SJT with a pick-one format on applicant differentiation, study-relatedness, chance to perform and ease of cheating. Thus, the rating format was – in contrast to our hypothesis – judged more favourable than the pick-one format. Finally, response instructions had a significant main effect on the ease-of-cheating item ($F(1, 362) = 4.53$, $p = .034$, $\eta^2 = .01$) with the would-do instructions ($M = 5.33$, $SD = 1.79$) judged as easier to cheat than the should-do instructions ($M = 4.92$, $SD = 1.86$). With regard to our hypothesis, no differences between should-do and would-do instructions were found for the overall process favourability, but should-do instructions were judged as more difficult to cheat than would-do instructions. Two-way ANOVAs revealed no significant interaction effects between response instructions and response format.

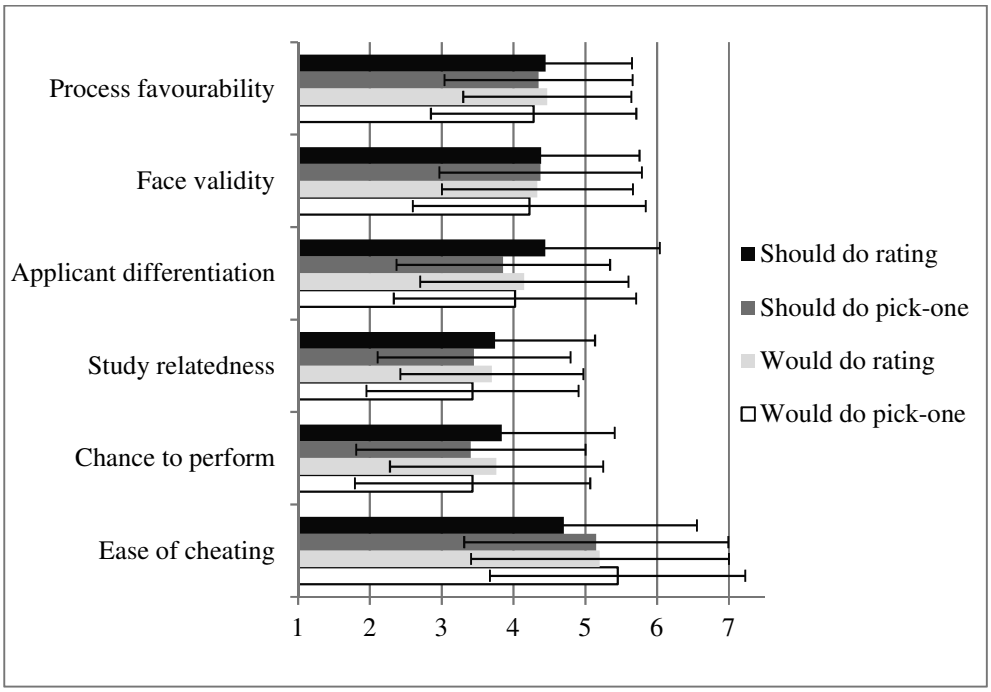

Figure 1. Process favourability and judgements on the other applicant perception items for the four SJT versions. Error bars reflect standard deviations.

Subgroup differences

The demographic subgroup differences in applicant perceptions are shown in Table 3. No significant main effects were found for gender, ethnic background or first-generation university status on the judgements of process favourability, the procedural justice factors and ease of cheating. However, significant interaction effects between subgroup and either response instructions or response format were found. Demographic subgroup differences for the four separate SJT versions are depicted in Appendix 5B.

Gender and response format had a significant interaction effect on the applicant differentiation item ($F(1, 362) = 4.80$, $p = .029$, $\eta^2 = .01$) and the study-relatedness item ($F(1, 362) = 7.64$, $p = .006$, $\eta^2 = .02$). The more positive judgement of the rating format than the pick-one format was stronger for men than for women on the applicant differentiation item ($d = 0.46$ vs. $d = 0.15$) and on the study- relatedness item ($d = 0.61$ vs. $d = 0.08$). Ethnic background and response instructions had a significant interaction effect on process favourability ($F(2, 336) = 4.42$, $p = .013$, $\eta_p^2 = .03$), the face validity item ($F(2, 333) = 3.61$, $p = .028$, $\eta_p^2 = .02$) and the study-relatedness item ($F(2, 335) = 3.10$, $p = .046$, $\eta_p^2 = .02$). For applicants from a Dutch background, should-do and would-do instructions were rated similarly on process favourability ($d = 0.03$), the face validity item ($d = 0.04$) and the study-relatedness item ($d = 0.02$). In contrast, applicants from a non-Western background were more positive on should-do than would-do instructions (process favourability: $d = 0.36$; face validity: $d = 0.41$; study relatedness: $d = 0.27$), whereas applicants from a Western background were more positive on would-do than should-do instructions (process favourability: $d = -0.42$; face validity: $d = -0.12$; study relatedness: $d = -0.36$). First-generation university status and response format had a significant interaction effect on

Table 3
Average judgement on process favourability and the other applicant perception items for the different subgroups.

		Gender		First-generation university		Ethnic background		
	Overall	M	W	Yes	No	Dutch	NW	W
Process favourability	4.39 (1.28)	4.38 (1.34)	4.39 (1.26)	4.47 (1.29)	4.36 (1.23)	4.41 (1.26)	4.35 (1.25)	1.36 (1.44)
Face validity	4.33 (1.43)	4.28 (1.53)	4.34 (1.40)	4.56 (1.31)	4.21 (1.47)	4.27 (1.43)	4.49 (1.50)	4.10 (1.32)
Applicant differentiation	4.12 (1.57)	4.04 (1.74)	4.15 (1.51)	4.10 (1.61)	4.12 (1.56)	4.04 (1.54)	4.35 (1.66)	4.17 (1.56)
Study relatedness	3.58 (1.38)	3.55 (1.52)	3.59 (1.33)	3.70 (1.33)	3.55 (1.37)	3.57 (1.32)	3.75 (1.47)	3.45 (1.45)
Chance to perform	3.61 (1.58)	3.64 (1.66)	3.60 (1.56)	3.74 (1.61)	3.54 (1.55)	3.56 (1.57)	3.83 (1.57)	3.38 (1.50)
Ease of cheating	5.13 (1.83)	5.22 (1.89)	5.10 (1.81)	5.05 (1.73)	5.20 (1.84)	5.22 (1.83)	5.11 (1.76)	4.86 (1.85)

Note. M = Men W = Women NW = non-Western W = Western Standard deviations between brackets

process favourability ($F(1, 341) = 5.23$, $p = .023$, $\eta^2 = .02$), the face validity item ($F(1, 338) = 9.80$, $p = .002$, $\eta^2 = .03$) and the applicant differentiation item ($F(1, 340) = 4.25$, $p = .040$, $\eta^2 = .01$). First-generation university students judged an SJT with a rating format more favourably than an SJT with a pick-one format on process favourability ($d = 0.45$), the face validity item ($d = 0.51$) and the applicant differentiation item ($d = 0.42$). In contrast, for non-first-generation university students, both response formats were judged similarly on process favourability ($d = 0.05$), the face validity item ($d = -0.15$) and the applicant differentiation item ($d = 0.13$). Thus, as stated by our hypotheses, subgroups based on gender and ethnic background did not significantly differ in their applicant perceptions of an SJT. Regarding our research question, we found no significant difference in applicant perceptions between the subgroups based on first-generation university status. Nonetheless, the findings do indicate that subgroups might differ in their preference for certain SJT design features.

Discussion

The present study indicates that response format and – to a lesser extent – response instructions influence applicants' perceptions of an SJT. The results show that asking applicants to rate each separate response option leads to more favourable perceptions of an SJT than asking applicants to pick one of the responses as the best option. Additionally, when instructed to respond according to what they would actually do in the described situation, applicants perceive an SJT as easier to cheat than when instructed to respond according to what should be done in the described situation. The applicant subgroups based on gender, ethnic background or first-generation university status were comparable regarding their perceptions of the SJT. However, our results do show that applicants from a non-Western ethnic background hold more positive perceptions of an SJT with should-do instructions than of an SJT with would-do instructions. On the contrary, applicants from a Western ethnic background appear to be more positive about an SJT with would-do instructions than an SJT with should-do instructions. Additionally, men and first-generation university students perceive an SJT with a rating response format more favourably than an SJT with a pick-one response format.

Response instructions had a significant influence on the perceived ease of cheating, indicating that should-do instructions are not only statistically less susceptible to faking (Nguyen et al., 2005; Oostrom et al., 2017), but are also perceived as more difficult to fake than would-do instructions. Previous research has shown that applicants' perceptions of a test do not always correspond to the actual psychometric qualities of that test (Smither et al., 1993). For example, Chan (1997) found that personality tests were perceived as more predictive than cognitive ability tests, whereas empirical studies show that cognitive ability tests are more predictive than personality tests. Apparently, ease of cheating is more obvious to applicants than the predictive value of a test and might therefore provide a more effective means to enhance applicant perceptions. Response instructions had no significant effect on the overall process favourability of the SJT. Nevertheless, the negative association between process favourability and ease of cheating indicates that applicant perceptions may be enhanced by reducing the SJT's susceptibility to faking.

In contrast to our hypothesis, a rating response format was perceived more positively than a pick-on response format on three of the procedural justice factors and ease of cheating. We had expected applicants to be more positive on pick-one formats because applicants are more familiar with this response format in traditional multiple-choice admission tests (Patterson et al., 2011; Ryan & Huth, 2008) and because rating formats are commonly used by easier-to-

fake self-report measures. However, the results of this study indicate that applicants perceive rating formats as a better measure to differentiate between applicants, as more strongly related to medical school, as a better means to show skills and abilities and as more difficult to cheat than pick-one formats. Possible explanations for this finding are that rating formats provide applicants the possibility to give more nuanced responses and allow applicants to give a rating of all response options. The challenging situations described in SJTs may be solved using multiple approaches, causing pick-one formats to be perceived as unrealistic (Ryan & Greguras, 1998). Response formats that allow for more nuanced answers might better fit the dilemma-like nature of SJTs. Likewise, medical students preferred an SJT with a short-answer-question format over an SJT with a single-best-answer format (Neal et al., 2018). Unlike our expectations, the rating format was not judged as easier to cheat than the pick-one format. Apparently, when used in SJTs, rating formats are not associated with the negative characteristics of self-report measures in a selection context. More favourable perceptions of the rating format are desirable as previous research has demonstrated that rating formats are superior to other response formats on a variety of psychometric outcomes (Arthur et al., 2014).

Applicant perceptions did not differ across subgroups based on gender, ethnic background and first-generation university status. The absence of subgroup differences is in line with findings of previous studies (Hausknecht et al., 2004; Niessen et al., 2017a; Smither et al., 1993). Nevertheless, the significant interaction effects do indicate that certain subgroups might have more pronounced preferences for certain SJT design features. Specifically, men seem to perceive rating formats more positively than pick-one formats regarding applicant differentiation and study relatedness. Prior research on cognitive ability tests showed that open-ended response formats resulted in less gender differences in test performance than multiple-choice response formats (Stumpf & Stanley, 1996). Arthur et al. (2014) found that the gender difference in an SJT score was larger for a ranking format than for a rating format and most/least-effective format. This interaction effect between gender and response format on test performance might translate into a gender-response format interaction on applicant perceptions. More research is required to unravel this interaction effect.

Non-Western ethnic minority applicants appear to be more positive on should-do than would-do instructions. Although a previous study demonstrated that the administration method (paper-and-pencil vs. video-based) affected the Black-White difference in applicants perceptions of an SJT (Chan & Schmitt, 1997), this is the first study showing that response instructions might also affect ethnic differences in applicant perceptions of an SJT. McDaniel et al. (2007) demonstrated that SJTs with knowledge instructions (i.e. should do) had higher correlations with cognitive ability test, whereas SJTs with behavioural tendency instructions (i.e. would do) had higher correlations with personality. Applicants from a non-Western background might feel that knowledge-based tests are more face valid and stronger related to medical school than personality-based tests and therefore perceive should-do instructions more favourably. Another possible explanation for more positive perceptions of should-do instructions among non-Western ethnic minority applicants might be found in differences between individualistic and collectivistic cultures (Hofstede, 2001). We presume that non-Western minority applicants may have a stronger collectivistic cultural orientation than majority applicants and might therefore be more comfortable to judge the SJT response options according to the group norms instead of according to their own individual norms (Jetten, Postmes, & McAuliffe, 2002). Additionally, results seem to indicate that Western ethnic minority applicants are more favourable of would-do than should-do instructions. However, the sample size of the Western minority applicant group was very small, making it

difficult to draw strong conclusions from this finding. Future research is necessary to replicate these findings and to examine potential explanations.

First-generation university students perceive rating formats more positively than pick-one formats. It appears that applicants from a low socioeconomic background have a stronger preference for response formats that permit more nuanced responses than applicants from a high socioeconomic background. A possible explanation might be that applicants whose parents did not attend university have more negative test-taking attitudes on traditional formats of testing. SJTs with pick-one formats might be more strongly associated with traditional tests and therefore receive more negative perceptions. Nevertheless, prior research on demographic differences in applicant perceptions has mainly focused on gender and ethnic background. Thus, future research should take into account socioeconomic background when examining subgroup differences in applicant perceptions and should examine why first-generation university students are more favourable of rating formats.

Practical implications

Our findings present two practical implications for medical school admission committees which use an SJT and are concerned with the applicant perceptions of that SJT. Firstly, using should-do instructions as opposed to would-do instructions increases the SJT's favourability among ethnic minority applicants. Secondly, men and first-generation university students perceived an SJT with a rating format more positively than an SJT with a pick-one format. Moreover, applicant perceptions did not differ between the two response instructions and the two response formats for the majority applicants. Therefore, using these SJT design features to positively influence applicant perceptions among minority applicants does not lead to unfavourable perceptions among majority applicants.

Limitations and directions for future research

Although applicant perceptions in this study are solely based on a short description and two example items of an SJT, minor changes in the example items led to significant differences in applicant perceptions. Nonetheless, future research should assess the applicants' perspective after completing a full version of an SJT, preferably one that is used for the actual selection into medical school, to obtain a more thorough picture of the influence of changing the SJT design features on applicant perceptions.

Prior research has indicated that applicant perceptions may influence applicant behaviour (e.g. applicant withdrawal, recommendations to others) (Ababneh, Hackett, & Schat, 2014; Schmit & Ryan, 1997). However, the present study is limited to examining the influence of SJT design features on applicant perceptions. The behavioural consequences of positive or negative applicant perceptions of an SJT need to be addressed in future research.

In general, the average judgements on the applicant perception items were situated close to the midpoints of the rating scales. Additionally, the SJT was judged significantly less favourable than five of the ten other admission methods included in the online survey (i.e. motivation letter, cognitive capacity test, skills test, curriculum sample test and interview). Even though this study demonstrated that changing the design features of an SJT may enhance applicant perceptions, future research is advised to examine the influence of other SJT characteristics that may positively affect perceptions of SJTs.

Finally, perceptions of procedural justice are not only determined by the formal characteristics of the admission procedure, but also by the treatment of applicants and the explanations of admission procedures and decisions (i.e. interactional justice) (Gilliland, 1993). Enhancing applicants' perceptions of an SJT must be accompanied by devoting attention to these other aspects of the medical school admission procedure.

Conclusions

The applicant's perspective on the use of SJTs in medical school admission procedures should not be underestimated, because applicant perceptions might influence test-taking motivation, test performance and applicant withdrawal. The current study demonstrated that changing the response format of an SJT may positively affect applicant perceptions through advancing the procedural justice factors of applicant differentiation, study relatedness and chance to perform and by reducing the perceived ease of cheating. Additionally, applicant perceptions may be altered by using response instructions that are less susceptible to faking. Finally, this study indicated that certain design features may lead to more favourable perceptions of an SJT among minority applicants, presenting another potential measure for promoting widening access to medical school.

Chapter 6

The base rate problem in medical school admissions: A machine learning approach

This chapter has been submitted for publication as:
De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (submitted). The base rate problem is medical school admissions: A machine learning approach.

Abstract

Medical school admission committees increasingly include admission predictors to measure attributes such as professionalism. A methodological difficulty in predicting professionalism is the low occurrence of unprofessionalism. Retrospective studies approached this base rate, or class imbalance, problem by matching cases to controls, but the predictive value of the indicators identified by these studies is often low. The present study approaches this problem using machine learning, because it offers novel methods to address class imbalance. Different machine learning algorithms and methods to address class imbalance are compared for the classification of unprofessionalism in first-year medical students based on admission variables. Participants were 410 first-year medical students. All received a rating whether their professional behaviour 'deserves attention' or not. Predictor variables included pre-university GPA, scores on extracurricular activities, cognitive tests, an SJT and a personality measure, and participation in a coaching day about the admission process. Six machine learning algorithms and three class imbalance methods were applied. True positive rate and true negative rate were calculated for each algorithm-class-imbalance-method combination. Class imbalance was reflected by low classification accuracy for the minority (unprofessional) group and high classification accuracy for the majority (professional) group. Most effective class imbalance method was the synthetic minority oversampling technique (SMOTE). Algorithms which resulted in the highest classification accuracy for the minority group were *k*-nearest neighbourhood (KNN) and neural networks (NN). Controlled oversampling of minority data using SMOTE appears a more effective solution to the base rate problem than the extensive removal of majority data in retrospective case-control studies. KNN and NN algorithms show the most accurate classification of unprofessionalism, but the accuracy is still suboptimal (max. 74.5%). More effort should be put in the development and collection of more reliable and valid input and output data to optimise the use of machine learning in medical school admissions.

Introduction

In recent times, most medical school admission procedures have been extended to include noncognitive admission methods such as personal statements (Ferguson, McManus, James, O'Hehir, & Sanders, 2003), multiple mini-interviews (Eva, Rosenfeld, Reiter, & Norman, 2004) and situational judgement tests (SJTs) (Lievens, 2013). The inclusion of these novel admission tools reflects the perceived relevance of noncognitive attributes to the medical profession (Walsh, Arnold, Pickwell-Smith, & Summers, 2016) and in medical education (Irby, Cooke, & O'Brien, 2010). One noncognitive attribute that has received considerable attention is professionalism (Shrank, Reed, & Jernstedt, 2004). This attribute is characterised by its multidimensional nature, visible in various conceptualisations. For example, the American Board of Internal Medicine defines medical professionalism as a set of ten commitments, for instance the commitment to honesty with patients and the commitment to patient confidentiality (ABIM Foundation, 2002). Another examples is the conceptualisation by Swick (2000), who describes medical professionalism as being constituted of nine behaviours, including the demonstration of humanistic values such as honesty and integrity. In addition, Kanter, Nguyen, Klau, Spiegel, and Ambrosini (2013) consider professionalism as consisting of six principles: excellence, accountability, altruism, humanitarianism, respect for others, and honour and integrity. Despite these varying conceptualisations, there is general agreement on the relevance of professionalism to the medical domain, as is evidenced by the incorporation of this attribute in many competency frameworks (e.g. ACGME Competencies, CanMEDS, Good Medical Practice) (Iobst et al., 2010). As a result, measures of professionalism and professionalism-related components (e.g. integrity, ethics) are increasingly introduced in medical school admissions (Finn, Mwandigha, Paton, & Tiffin, 2018; Jerant et al., 2012; Knights & Kennedy, 2006; Patterson et al., 2017).

In general, evidence regarding the predictive value of admission tools measuring professionalism is mixed, possibly caused by the difficulty to reliably measure noncognitive attributes (Kulatunga-Moruzi & Norman, 2002). Another likely explanation for the low predictive validity of professionalism-based admission instruments is the skewed distribution of this competence, with most medical students displaying appropriate professional behaviour and only a small group of students (estimated between 5 and 15%) exhibiting unprofessional behaviour (Fargen, Drolet, & Philibert, 2016; Mak-Van der Vossen et al., 2016). The skewed distribution of a criterion may diminish the accuracy and utility of prediction models, a complication which is called the base rate problem (Niessen & Meijer, 2016; Rosenfeld, Sands, & Gorp, 2000).

In the present study, we will examine the predictive value of several admission variables in the prediction of future professionalism in medical school. In our pursuit to tackle the base rate problem, the data are analysed using machine learning methods, which we compare to a traditional statistical approach (logistic regression analysis). Machine learning involves the application of automatic self-learning algorithms to extract the underlying patterns in the observed data (Alpaydin, 2009). As machine learning is often applied to criteria with highly skewed distributions, several approaches to address the base rate problem have been introduced to the machine learning literature. The current study will demonstrate the application of several machine learning algorithms and approaches to address the base rate problem in examining the prediction of future professionalism in medical school. We are not aware of prior research that has applied machine learning to the context of medical school admissions.

Predicting professionalism

The inclusion of professionalism in medical school admission procedures invokes the following question: how predictive are admission tools that aim to measure professionalism? In general, admission committees apply noncognitive tools with the intention to predict similar noncognitive criteria. Several retrospective studies have linked indicators of professionalism to such criteria during the medical professional career path. Most notable is the research of Papadakis and colleagues, who indicated that medical doctors disciplined by a medical-licensing board were two to three times as likely as non-disciplined doctors to have received a negative evaluation of their professional behaviour during medical school (Papadakis, Hodgson, Teherani, & Kohatsu, 2004; Papadakis et al., 2005). Another study showed that students who experienced academic or personal difficulties during medical school had more negative comments in their academic references than students without difficulties (Yates & James, 2006). Additionally, Chang, Boscardin, Chou, Loeser, and Hauer (2009) found that medical students who eventually failed an examination of a patient-physician interaction had more often received a low rating of communication/professionalism skills during a previous clerkship than medical students who passed the examination.

Although these studies point to some potential noncognitive predictors of professionalism, several other studies examining the predictive value of noncognitive measures showed less convincing results. For example, Stern, Frohna, and Gruppen (2005) found no significant relationship between on the one hand cognitive and noncognitive variables during admissions and on the other hand professional behaviour in the third year of medical school. In addition, Ainsworth and Szauter (2018) mention the difficulty to prospectively identify medical students who show repetitive problematic behaviour. In another study, the occurrence of conduct-related issues in the registration forms of first-year medical doctors was significantly and positively associated with the score on a self-esteem scale administered during medical school admissions (Paton, Tiffin, Smith, Dowell, & Mwandigha, 2018). However, the positive predictive value of the self-esteem scale turned out to be a low 4.4%, indicating that, of all doctors who were predicted to have conduct-related issues in their registration forms, only 4.4% actually had such issues and thus 95.6% were unjustly predicted to have such issues. Because of these results, several researchers have started questioning the usefulness of medical school admission instruments measuring professionalism and other noncognitive constructs (Benbassat & Baumal, 2007; Colliver, Markwell, Verhulst, & Robbs, 2007; Niessen & Meijer, 2016; Norman, 2015; Prasad, 2011).

Probable explanations for these contradictory findings are the use of different research designs in combination with the prevalence of unprofessionalism as an outcome measure. In the search for predictors of future professionalism, a considerable number of studies have used retrospective case-control designs. In such research designs, the target group (e.g. disciplined medical doctors) is matched to a group of ‘professional controls’. Various case:control ratios have been used in these studies, for example 1:2 (Chang et al., 2009; Papadakis et al., 2005) or 1:4 (Yates & James, 2006). Yet, the problem of these ratios is that they are not always realistic. Based on a literature review, Fargen et al. (2016) estimated the prevalence of unprofessional behaviour among medical students and residents to be much lower, namely between 5% and 15% (i.e. a case:control ratio of approximately 1:20 to 1:7). Another study among 2460 medical students found reports of unprofessional behaviour for 7.9% of the students (Mak-Van der Vossen et al., 2016). The prevalence of medical doctors who are disciplined appears to be even lower; Khaliq, Dimassi, Huang, Narine, and Smego (2005) report a prevalence of 2.8% across 14314 medical doctors. Further, Papadakis et al.

(2005) mention a prevalence of disciplinary action among medical doctors of 0.3%, corresponding to a ratio of approximately 1:333.

Base rate problem

The classification of relatively rare events poses a problematic situation (see Box 1 for an illustration). A large classification accuracy could be achieved by simply classifying all individuals as the majority class (in the present study: professional medical students). For instance, if the prevalence of unprofessionalism among medical students is 5% and we would predict all applicants to become professional students, 95% of our predictions would be accurate. However, this large prediction accuracy would not apply to individuals belonging to the minority class (in this study: unprofessional medical students), who would all be wrongly classified as the majority class. Thus, the accurate identification of who will behave professionally as a medical doctor or medical student is hampered by the low prevalence of unprofessionalism, which is problematic because unprofessionalism is of great concern, since it may result in misconduct and subsequent allegations are highly undesirable for anyone involved (Binder, Friedli, & Fuentes-Afflick, 2015). Colliver et al. (2007) re-evaluated the case-control studies of Papadakis et al. (2004; 2005), taking into account the low prevalence of disciplinary action against medical doctors, and concluded that the relationship between professional behaviour in medical school and future disciplinary action is too weak to be of any practical value.

The base rate problem is often tackled by researchers by matching cases (i.e. rare events) to controls (i.e. non-rare events) in order to obtain a more balanced class distribution. Although such retrospective case-control studies may be able to identify indicators of a rare event, they do no guarantee that these indicators will prospectively predict that rare event under realistic low base rates (Mercaldo, Lau, & Zhou, 2007). The present study will focus on the base rate problem in a medical school admission context by examining the prediction of unprofessionalism in first-year medical students based on admission data. We will attempt to address the base rate problem by using techniques from the field of machine learning.

A test that aims to predict a rare event may have a good sensitivity of .90, indicating that the test correctly identifies 90% of all rare events (i.e. true positives) and a good specificity of .90, indicating that the test correctly identifies 90% of all non-rare events (i.e. true negatives). In a hypothetical dataset containing 50 rare events and 950 non-rare events (i.e. prevalence is 5%), the test would result in 45 (= .90 * 50) true positives and 855 (= .90 * 950) true negatives. However, this test will also lead to the incorrect classification of 95 non-rare events (i.e. false positives), meaning that of all predicted rare events (45 + 95) only 32.1% are true rare events. In other words, the positive predictive value (PPV) of the test only equals .32. A lower prevalence (e.g. 1%) would lead to an even lower PPV (in our example leading to PPV = .08). This example shows that even tests with good sensitivity and specificity will have a low positive predictive value when the prevalence is low (Norman, 2015; Prasad, 2011).

Box 1. Illustration of the effect of the base rate problem on the positive predictive value of a test.

Machine learning

Machine learning is a subdomain, of the field of artificial intelligence, that is increasingly used to acquire knowledge from data (Witten, Frank, & Hall, 2011), through the extraction of an algorithm which captures the underlying patterns in the data (Alpaydin, 2009). Instead of programming pre-determined pathways, in machine learning a computer automatically

"learns" the underlying patterns in the presented data, indicating the self-programming nature of machine learning (Domingos, 2012; Rowe, 2019). Machine learning aims to uncover structural patterns in data which can be generalised to new data (Domingos, 2012) and which may help to understand the data and make future predictions.(Witten et al., 2011) In machine learning terminology, variables are named features or attributes, and subjects are named instances or examples (Kotsiantis, 2007; Mitchell, 1997). The prediction of professionalism in medical school is a classification task in which we use a set of features to assign each instance to one of two classes, labelled 'professional' and 'unprofessional'. The algorithm in a classification task – the classifier – is built on a set of training data consisting of instances for which the class labels are known (Alpaydin, 2009). The case in which the class labels of the instances are known is called supervised machine learning, whereas the label *un*supervised machine learning refers to the situation in which class labels are unknown (Kotsiantis, Zaharakis, & Pintelas, 2006).

When building a classifier on the training data, it is often possible to obtain high classification accuracy, but this does not guarantee that the classifier will generalise to a set of test data consisting of new instances (Domingos, 2012). The difference in classification accuracy between training and test data is generally caused by overfitting, referring to the situation in which an overcomplex algorithm makes a classifier too reliant on the quirks in the training data (Alpaydin, 2009). Hence, the performance of classifiers is evaluated using cross-validation by dividing the dataset into separate folds, holding out one fold as the test data and building the classifier on the remaining folds which comprise the training data. This procedure is repeated until every fold has once been the test data and the results across the different folds are averaged (Kotsiantis, 2007). The various algorithms and the automatic self-learning nature of machine learning may enable the development of classifiers that capture the structural patterns linking the admission features to future unprofessionalism. For example, in the educational domain, machine learning methods have been used to build classifiers aimed at predicting student drop-out (Baker & Yacef, 2009; Dekker, Pechenizkiy, & Vleeshouwers, 2009; Delen, 2010).

Machine learning is often applied to imbalanced datasets, that is, datasets with a disproportionate distribution of instances across classes (Ramyachitra & Manikandan, 2014), for example in diagnosing rare diseases or detecting fraudulent transactions (Chawla, Lazarevic, Hall, & Bowyer, 2003). Regarding the base rate problem, machine learning provides several ways to address the classification of rare events. Two common approaches to handle imbalanced datasets are resampling and cost-sensitive learning (Weiss & Hirsh, 2000). The resampling approach may encompass either undersampling the majority class or oversampling the minority class (Weiss & Hirsh, 2000). The disadvantage of undersampling the majority class is that the removal of majority instances may lead to the loss of potentially useful data for building the classifier (Batista, Prati, & Monard, 2004). Vice versa, the downside of oversampling the minority class is that adding exact copies of minority instances increases the risk of overfitting (Guo, Yin, Dong, Yang, & Zhou, 2008). However, the risk of overfitting can be reduced by using oversampling methods that create so-called synthetic minority instances, which are not exact copies, but are based on the actual minority instances in the training data.(He & Garcia, 2008) Although evidence on the effectiveness of both resampling approaches is mixed, a number of studies has indicated better performance for oversampling the minority class (Batista et al., 2004; García, Sánchez, & Mollineda, 2012; Japkowicz & Stephen, 2002; Marqués, García, & Sánchez, 2013).

Another approach to handle imbalanced datasets is cost-sensitive learning, which incorporates different costs for the two possible classification errors: false positives and false negatives (Witten et al., 2011). Research has suggested that changing the costs of these

classification errors may enhance the utility of a classifier (Ciraco, Rogalewski, & Weiss, 2005). Which classification error is more costly depends on the stakeholder. For instance, medical school admission committees may consider the incorrect classification of unprofessional students more costly than the incorrect classification of professional students, whereas the reverse is likely true for medical school applicants. In imbalanced datasets, the costs of misclassifying the minority class are often considered larger than the costs of misclassifying the majority class (He & Garcia, 2008). The different costs of the different errors are depicted in a so-called cost matrix. Cost-sensitive learning builds a classifier after reweighting the instances according to the cost matrix (Witten et al., 2011).

The aim of the present study is to examine the use of different machine learning algorithms in the classification of unprofessionalism in first-year medical students based on several cognitive and noncognitive admission features, and to compare this with a traditional statistical method (logistic regression analysis). Additionally, resampling and cost-sensitive learning are applied to address the base rate problem caused by the low occurrence of unprofessional behaviour.

Methods

Context and procedure

This study was conducted at the Erasmus MC Medical School, Rotterdam, the Netherlands. The admission procedure of this undergraduate-entry-level medical school consisted of three parts: i) pre-university GPA (pu-GPA), ii) extracurricular activities and iii) five cognitive tests administered on site. During the on-site testing days, we also administered an SJT measuring integrity. Additionally, a personality measure (HEXACO-SPI; see below) was administered online between the testing days and the admission decision. The SJT and personality measure were administered for research purposes only; participation was voluntary and did not affect the admission decision. Applicants were informed about the study's aim and that their answers to the SJT and personality measure would not influence the admission decision. Respondents signed informed consent before participation. The data in this study were anonymised and processed confidentially. The Ethical Committee of the Institute of Psychology, Erasmus University Rotterdam, judged that this study could be excluded from further ethical approval of the Medical Ethical Committee.

The medical school curriculum consists of a three-year pre-clinical bachelor programme and a three-year predominantly clinical master programme. During the first bachelor year, teachers of small-scale educational groups (e.g. on practical clinical skills) evaluate the professional behaviour of their students. Teachers rated professional behaviour on three dimensions: i) commitment and communication, ii) participation and iii) reflection and feedback. Three possible ratings indicate whether a student's professional behaviour i) is going very well, ii) is going well or iii) deserves attention. The study sample was split into two classes, a class of students who received at least one 'deserves attention' rating on one of the three professional behaviour dimensions and another class of students who received only 'going well/very well' ratings. The class of a student was linked to the available admission features (i.e. pu-GPA and score on extracurricular activities, cognitive tests, SJT and personality measure) using the students' identification numbers.

Participants

First-year medical students of cohort 2016 comprised this study's sample (N = 410). This sample was on average 18.5 years old (SD = 1.1) and consisted of 72.2% women. In total, 36

students (8.8%) received a 'deserves attention' rating for at least one of the three professional behaviour dimensions. Thirty students received one 'deserves attention' rating, four students received two 'deserves attention' ratings and two students received three 'deserves attention' ratings.

Admission features

Pre-university grade point average (pu-GPA)
The average grades (on a scale from 1 to 10) obtained during the fifth year of secondary education (total duration: six years) on the subjects Dutch, English, chemistry, physics, biology and mathematics.

Extracurricular activities (ECA)
A score based on the quality and quantity of extracurricular activities carried out by applicants in addition to their educational activities.

Cognitive tests
The scores on five cognitive tests regarding logical reasoning, scientific reading, anatomy, mathematics and a curriculum sample test comprising of a trial lecture.

Situational Judgement Test (SJT)
The score on an SJT designed to measure a key component of professionalism, namely integrity. SJTs consist of scenarios that describe dilemma-like situations and ask respondents to judge the appropriateness of several responses to those situations (Lievens, 2013). The integrity SJT consisted of 31 scenarios, four response options per scenario and a six-point rating scale (1: *Very inappropriate* - 6: *Very appropriate*). Applicants had to judge the response options in terms of what should be done given the situation (i.e. knowledge instructions). The development of the SJT is detailed in De Leng et al. (2018) The SJT was scored using three different methods because previous research indicated the relevance of the scoring method for the psychometric quality of an SJT (De Leng et al., 2017; McDaniel et al., 2011; Weng et al., 2018). Firstly, a raw consensus scoring method calculated the absolute distance between the rating of an applicant and the average rating across the other applicants. Secondly, a standardised consensus scoring method first performed a within-person *z* standardisation before calculating the absolute distance. The within-person *z* standardisation minimises the effect of response tendencies in the use of a rating scale by ensuring that each applicant has a mean of 0 and a standard deviation of 1. Thirdly, a dichotomous consensus scoring method divided the rating scale in half and applicants received a score of 1 when they were on the same side as the average rating across the other applicants, and a score of 0 when they were on the other half.

Additionally, the SJT consisted of two types of response options: desirable response options describing generally appropriate responses and undesirable response options describing generally inappropriate responses. SJT scores were calculated based on all response options, only the desirable response options and only the undesirable response options. These three different scores were calculated because prior studies found that the construct and predictive validity differed between SJT scores based on knowing what to do and SJT scores based on knowing what *not* to do (De Leng et al., 2018; Elliott et al., 2011; Stemler et al., 2016). The desirable and undesirable response options of the integrity SJT were based on theoretical models that were either positively or negatively related to integrity

(see De Leng et al. (2018) for more information). In total, we calculated nine different SJT scores (three scoring methods × three subscores).

Personality measure
The score on the six personality dimensions of the HEXACO Simplified Personality Inventory (HEXACO-SPI) was used: Honesty-humility ('I find it hard to lie'), Emotionality ('I am often worried that something will go wrong'), eXtraversion ('I like to talk to others'), Agreeableness ('I quickly agree with others'), Conscientiousness ('I work very accurately') and Openness to experience ('I like poems') (De Vries & Born, 2013). The HEXACO-SPI consists of 16 statements per dimension and a five-point rating scale (1: *strongly disagree* – 5: *strongly agree*). The internal consistency reliability (alpha coefficient) was .70 for Honesty-humility, .81 for Emotionality, .87 for eXtraversion, .71 for Agreeableness, .84 for Conscientiousness and .82 for Openness to experience.

Participation in coaching day
A score indicating whether an applicant voluntarily participated in a free-of-charge coaching day on which applicants received information on the admission procedure.

Analyses
This subsection will first provide a brief description of some of the machine learning algorithms that were applied to the dataset, followed by a description of the procedures that were used to address the class imbalance in the dataset. We will then describe the metrics used to evaluate the performance of the classifiers. The analyses were conducted using WEKA Version 3.8.3 (Eibe, Hall, & Witten, 2016).

Machine learning algorithms
Six machine learning algorithms, discussed in the review of Kotsiantis (2007), were applied to our dataset: *k*-nearest neighbourhood (KNN), neural networks (NN), decision trees (DT), rule induction (RI), naive Bayes (NB) and support vector machines (SVM). The algorithms are briefly explained below. Technical terminology is reduced to a minimum. For a more extensive description and technical details, we refer to relevant textbooks on machine learning (Alpaydin, 2009; Witten et al., 2011).

K-nearest neighbourhood
The *k*-nearest neighbourhood (KNN) algorithm assigns a new instance to the most common class among a group of *k* similar instances in the training data (Witten et al., 2011). The similarity between instances is determined by the distance between them in an instance space (Figure 1), because the KNN algorithm assumes that instances in close proximity are more similar to each other than instances which are further apart (Mitchell, 1997). A larger value for *k* (i.e. more neighbours) can make a classifier less sensitive to noisy instances (Kotsiantis, 2007). The KNN algorithm was performed in WEKA using the IBk classifier with a Euclidean distance metric and $k = 1$, $k = 3$ and $k = 5$. In general, the value for *k* is small and odd to avoid ties (Phyu, 2009).

Neural networks
Neural network (NN) algorithms are composed of interconnected nodes which are structured in an input layer, one or more hidden layers and an output layer (Figure 2). The number of nodes in the input layer equals the number of features and the number of nodes in the output layer equals the number of classes (Witten et al., 2011). The nodes in an NN are generally

represented by so-called perceptrons, elements that produce an output value of either +1 or -1 based on the weighted sum of the input values and the weight connections (Mitchell, 1997). The input values may flow directly from the input layer or from the output values of other perceptrons (Alpaydin, 2009). The weight of the connections are often determined using a back propagation algorithm, which calculates the error based on the output of the neural network and the actual output and then propagates back an error signal that gradually modifies the weights of the connections in the network (Lykourentzou, Giannoukos, Nikolopoulos, Mpardis, & Loumos, 2009). The NN algorithm was applied to our dataset using the MultilayerPerceptron classifier in WEKA.

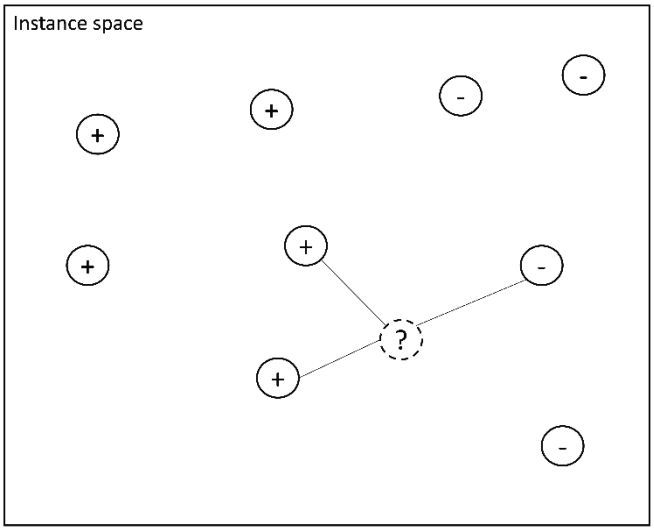

Figure 1. Schematic presentation of the *k*-nearest neighbourhood (KNN) algorithm.

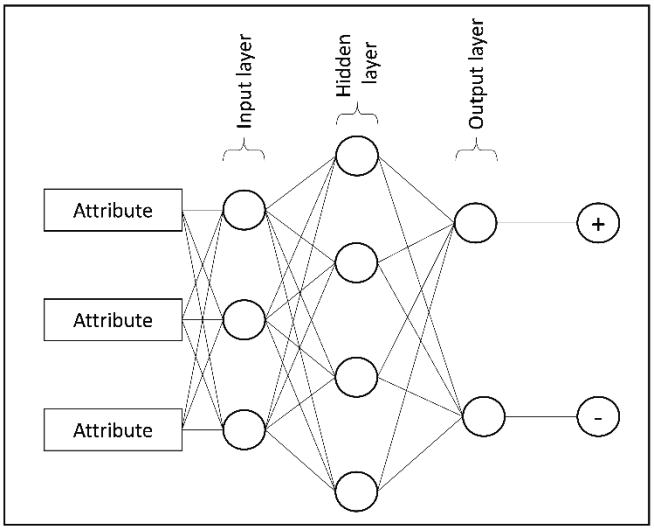

Figure 2. Schematic presentation of the neural networks (NN) algorithm.

Decision trees

Decision trees (DTs) consist of nodes, branches and leaves (Figure 3). The classification of a new instance starts at the root node of the DT which corresponds to a particular feature. The branches that flow from the root node correspond to the feature values. An instance moves through the DT from top to bottom following its feature values until it ends at a leaf which corresponds to a class (Mitchell, 1997). The algorithm creates a DT by searching for the feature that provides the best split between the classes by using a function called information gain. The feature with the largest information gain will be the root node of the DT. The training data are divided according to the feature values at the root node and then the process is repeated for these smaller subsets of training data until the data can no longer be divided (Witten, Frank, & Hall, 2011). Most DTs are subsequently pruned by removing nodes to avoid overfitting and make the decision tree better generalisable to new instances (Kotsiantis, 2007). The DT algorithm was performed in WEKA using the J48 classifier with pruning.

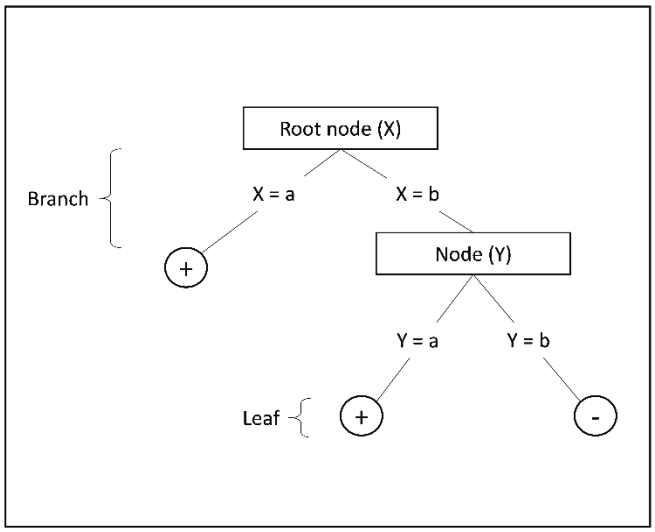

Figure 3. Schematic presentation of the decision tree (DT) algorithm.

Rule induction

Rule induction (RI) algorithms create a list of consecutive if-then rules that assign instances to classes (Alpaydin, 2009). For a group of instances belonging to the same class, the algorithm identifies a rule that maximizes a rule quality measure (e.g. accuracy) (Kotsiantis, 2007). In the following step, all instances of this class that are covered by the rule are removed from the dataset and RI continues creating a rule based on the remaining instances (Witten et al., 2011). This procedure is called sequential covering and is repeated until all instances are covered by a rule (Alpaydin, 2009). As with DTs, the risk of overfitting is reduced by pruning the rule set, generalising rules that are too restrictive (Kotsiantis, 2007). Afterwards, rules are generated for the other class or classes. A new instance is classified by running it through a decision list of consecutive rules until one of the rules applies and classifications stops (Witten et al., 2011). In WEKA, the PART algorithm with pruning was used to learn rules.

Naive Bayes

The Naive Bayes (NB) algorithm uses Bayesian statistics to calculate the probability of each possible class given the feature values, i.e. the posterior probability. The algorithm assigns a new instance to the class with the highest posterior probability (Alpaydin, 2009). In Bayesian statistics, the calculation of the posterior probability of class C given feature values F is based on Bayes' rule:

$$P(C|F) = \frac{P(F|C)P(C)}{P(F)}$$

where P(C) is the prior probability of class C or the frequency of C in the training data (Witten et al., 2011). P(F|C) is the likelihood that an instance belonging to class C has a particular feature value F and P(F) is the marginal probability of feature value F (Alpaydin, 2009). The NB algorithm is called naive because the classical Bayes' rule assumes conditional independence, meaning that all features are independent given a class value, an assumption that is violated in most real-world applications. Despite this violation, the NB algorithm often achieves high classification accuracies, because even if the posterior probabilities are inaccurate, classification will be accurate as long as the correct class has the highest posterior probability (Zhang, 2004). The NB algorithm was applied to our training data using the NaiveBayes classifier in WEKA.

Support vector machine

Support vector machine (SVM) is an algorithm that finds a linear line – a hyperplane – in the instance space that separates two classes (Alpaydin, 2009). This optimally separating hyperplane is found by maximising the margin around the hyperplane until it touches the closest instances (Figure 4). The instances that lie on the margin are called the support vectors and other instances are ignored (Kotsiantis, 2007). Since the support vectors are the only relevant instances, the chance of overfitting is reduced which results in relatively stable

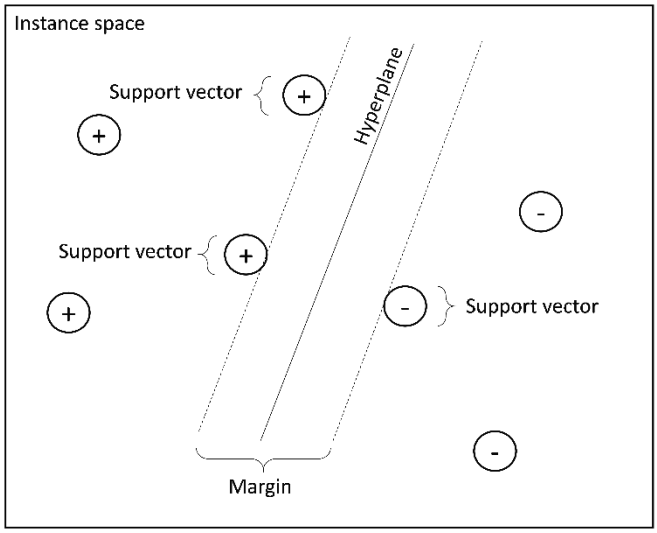

Figure 4. Schematic presentation of the support vector machine (SVM) algorithm.

hyperplanes (Witten et al., 2011). Maximising the margin around the hyperplane improves the generalisation of the classifier to new instances (Alpaydin, 2009). Nonetheless, linear lines are often too simple to separate classes in many real-world applications and more complex, nonlinear boundaries are required. SVM handles this problem by applying a nonlinear transformation to the instance space. A linear hyperplane in this higher-dimensional transformed feature space will correspond to a nonlinear boundary in the original instance space (Witten et al., 2011). The function used to map the training data from a low-dimensional space to a higher-dimensional space is called a kernel function (Noble, 2006). The SVM algorithm was applied to the training dataset using the SMO classifier with a polynomial kernel in WEKA.

Approaches used to address class imbalance

Three approaches were used to address the class imbalance in our dataset: random undersampling, synthetic minority oversampling technique (SMOTE) and cost-sensitive learning. These approaches are briefly described below. For more information on these approaches, we refer to the articles cited below.

Random undersampling

In line with the retrospective case-control studies described above (Chang et al., 2009; Papadakis et al., 2004; Papadakis et al., 2005; Yates & James, 2006), we undersampled majority instances to obtain case:control ratios of 1:4, 1:6 and 1:8. These ratios were picked based on the assumption that the true ratio of unprofessional to professional students might be larger than the ratio found in the present study (i.e. approximately 1:10), due to teachers' reluctance to fail students based on professional criteria, in other words, the 'failure to fail' issue (Cleland, Knight, Rees, Tracey, & Bond, 2008; Mak-Van der Vossen, Peerdeman, Van Mook, Croiset, & Kusurkar, 2014). Additionally, the ratio of unprofessional to professional students might differ for different types of unprofessional behaviour (Rennie & Crosby, 2001). For example, posting of unprofessional online content (e.g. evidence of intoxication) was reported by 34.7% of the medical students in one study (Barlow et al., 2015), dishonest clinical behaviours (e.g. reporting an omitted physical examination as normal) were reported by 43.3% of the medical students in a different study (Dyrbye et al., 2010), and yet another study showed that 14% of the respondents had engaged or would consider engaging in plagiarism (Rennie & Crosby, 2001). Based on these different percentages, we selected varying but still realistic ratios of unprofessional to professional students for the resampling methods.

Synthetic Minority Oversampling Technique (SMOTE)

As mentioned earlier, random oversampling by adding exact copies of minority instances to the dataset may increase the chance of overfitting (Guo et al., 2008). SMOTE tackles this problem by finding the k nearest minority class neighbours of each minority instance and creating new synthetic minority instances along the lines between each minority instance and its k nearest minority neighbours (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Thus, the new synthetic minority instances are based on similarities between existing minority instances (He & Garcia, 2008). Using SMOTE with default settings, we oversampled the minority instances to obtain case:control ratios of 1:4, 1:6 and 1:8.

Cost-sensitive learning

Cost-sensitive learning starts with a cost matrix that defines the costs of both misclassification errors, i.e. false positives (C_{FP}) and false negatives (C_{FN}) (McCarthy, Zabar,

& Weiss, 2005). For most imbalanced datasets, incorrectly classifying a minority instance is considered more costly than vice versa (Sun, Wong, & Kamel, 2009). However, from the perspective of a medical school applicant, misclassifying a majority instance (that is, identifying a professional student as unprofessional) is more costly than misclassifying a minority instance (that is, identifying an unprofessional student as professional). Therefore, we applied four different cost matrices to our dataset, two considering $C_{FP} > C_{FN}$ (serving interest of the medical school and society) and two considering $C_{FN} > C_{FP}$ (serving in the interest of medical school applicants) with misclassification error ratios 5:1, 3:1, 1:3 and 1:5. We used CostSensitiveClassifier in WEKA to convert the classifiers into cost-sensitive classifiers.

Evaluation metrics

Classification accuracy is an inappropriate measure for evaluating classifier performance in imbalanced datasets, because it leads to misleading conclusions (Weiss & Hirsh, 2000). Therefore, classifier performance was evaluated by measures that are not impacted by class distribution: sensitivity, specificity and area under the receiver operating characteristic (ROC) curve. The sensitivity, or true positive rate (TPR), depicts the percentage of correctly identified unprofessional instances among all unprofessional instances, whereas the specificity, or true negative rate (TNR), depicts the percentage of correctly identified professional instances among all professional instances. The ROC curve combines the information on sensitivity and specificity in a two-dimensional graph with the TPR on the vertical axis and the false positive rate (FPR = 1 – TNR) on the horizontal axis (Guo et al., 2008). ROC curves are used to assess trade-offs between TPRs and FPRs. The area under the curve (AUC) in an ROC graph is used to evaluate the overall performance of a classifier, with larger values indicating better performance (Weiss & Provost, 2001). We calculated the evaluation metrics using ten-fold cross-validation (default).

Feature selection

Feature selection is used in machine learning to eliminate irrelevant and redundant features. Advantages of feature selection are increased classification performance and a better understanding of the underlying patterns in the data (Chandrashekar & Sahin, 2014). Feature selection was performed using two approaches, namely the evaluation of individual features and the evaluation of subsets of features (Yu & Liu, 2004). The evaluation of individual features involves the ranking of features based on a ranking criterion, whereas subset evaluation involves the selection a subset of features that collectively have good classification performance (Guyon & Elisseeff, 2003). Individual feature evaluation helps to reduce irrelevant features, but does not take into account the redundancy between features (Chandrashekar & Sahin, 2014). In contrast, subset evaluation searches for a minimum subset of relevant and non-redundant features that achieves good classification performance (Yu & Liu, 2004). In WEKA, InfoGainAttributeEval was used for individual feature evaluation and CfsSubsetEval was used for subset evaluation. InfoGainAttributeEval ranks the features based on their information gain, which is the increase in information value in the output caused by the inclusion of a feature (Chandrashekar & Sahin, 2014). CfsSubsetEval tries to find the subset of features which are strongly associated with the class, but not with each other (Witten et al., 2011). Individual subset evaluation was performed separately for the nine different SJT versions (three scoring methods × three subscores), whereas subset evaluation was performed once for all different SJT versions since subset evaluation takes into account the redundancy between features.

Results

This section will start with a description of the results from traditional logistic regression analyses and retrospective case-control comparisons, followed by a description of the results from the machine learning approach.

Traditional statistical analysis

To compare the results of the machine learning techniques to commonly used statistical analysis techniques, we performed logistic regression analysis with professional behaviour as a dichotomous outcome variable and the admission variables as predictors. This analysis revealed no significant relationship between the admission variables and the rating on professional behaviour, when including the SJT score based on all response options scored using the raw consensus method ($X^2(15) = 9.46$, $p = .852$). A significant relationship with the outcome variable was neither found when including the score on any of the eight other SJT versions. Additionally, we used a retrospective case-control approach to analyse these data. Specifically, we used gender and pu-GPA to match students who received at least one 'deserves attention' regarding their professional behaviour with students who did not receive a 'deserves attention' rating, using a ratio of 1:3. We compared the admission variables of the cases and controls (Table 1) and found no significant differences.

Table 1

Mean (and standard deviation) on admission variables for students who received at least one 'deserves attention' rating regarding their professional behaviour and for students matched on gender and pu-GPA who did not receive a 'deserves attention' rating (ratio 1:3).

	Professionalism rating			
	≥ 1 'deserves attention'	n	no 'deserves attention'	n
Pre-university GPA	7.01 (0.63)	29	7.04 (0.53)	86
Extracurricular activities	48.51 (17.62)	36	46.15 (12.69)	99
Logical reasoning	57.67 (5.99)	30	57.68 (6.34)	81
Scientific reading	129.07 (12.01)	30	128.15 (11.19)	81
Anatomy	34.13 (5.99)	30	34.07 (4.82)	81
Mathematics	35.93 (5.98)	30	36.49 (5.42)	81
Curriculum sample	52.63 (8.95)	30	53.17 (7.69)	81
SJT raw consensus				
all response options	92.63 (17.55)	26	90.85 (21.84)	71
desirable response options	45.94 (9.57)	26	45.42 (12.03)	71
undesirable response options	46.68 (11.14)	26	45.43 (12.12)	71
SJT standardised consensus				
all response options	43.32 (14.63)	26	42.79 (14.95)	71
desirable response options	21.47 (7.58)	26	21.46 (8.10)	71
undesirable response options	21.84 (7.75)	26	21.32 (7.61)	71

Table 1 continued

	Professionalism rating			
	≥ 1 'deserves attention'	n	no 'deserves attention'	n
SJT *dichotomous consensus*				
all response options	113.38 (8.71)	26	113.22 (9.97)	71
desirable response options	57.77 (4.30)	26	57.77 (5.29)	71
undesirable response options	55.61 (5.52)	26	55.44 (5.67)	71
Honesty-humility	55.33 (6.12)	30	54.76 (8.91)	70
Emotionality	44.70 (7.75)	30	44.97 (9.60)	70
Extraversion	65.83 (7.40)	30	65.20 (9.95)	70
Agreeableness	49.43 (6.88)	30	47.86 (9.42)	71
Conscientiousness	60.23 (8.39)	30	59.41 (11.87)	71
Openness to experience	57.77 (8.10)	30	55.75 (11.20)	71
Coaching day	58.3%	36	66.4%	107

Note. Higher scores indicate lower performance for the raw and standardised consensus scoring methods.

Machine learning algorithms

The imbalance in the dataset was reflected by low true positive rates (TPR = percentage of correctly classified unprofessional students) and high true negative rates (TNR = percentage of correctly classified professional students) for all machine learning algorithms (Table 2). The highest TPR was found for the KNN algorithm with $k = 1$ ($8.3\% \leq TPR \leq 57.4\%$ for the different approaches to address class imbalance). For the imbalanced dataset, a TNR of 100% was found for the DT, RI and SVM algorithms, but this maximum TNR was associated with a TPR of 0%, indicating that these algorithms classify all instances as 'professional'. The higher TPR of the KNN algorithm was associated with a lower TNR ($62.8\% \leq TNR \leq 90.4\%$ for the different approaches to address class imbalance). The NN and NB algorithms resulted in a higher TPR in the resampled datasets than the DT, RI and SVM algorithms, but not as high as the KNN algorithm. For the KNN algorithm, larger values for k were associated with a lower TPR and a higher TNR. The AUC-values for the imbalanced dataset ranged between .39 and .50, indicating low classification performance.

In sum, these results show that imbalanced datasets lead to poor classification accuracy, no matter which algorithm is used. From the six algorithms applied to the data, KNN using $k = 1$ led to the least inaccurate classification of unprofessional students. These results indicated that classifier performance may benefit from approaches aimed at reducing the imbalance in the dataset.

Approaches to address class imbalance

The most successful approach to address the class imbalance in our dataset was synthetically oversampling instances from the minority class to achieve a ratio of unprofessional to professional students of 1:4 as evidenced by the highest TPR ($0\% \leq TPR \leq 57.4\%$) and the highest AUC-values ($.50 \leq AUC \leq .70$) for all algorithms (Table 2). Random undersampling resulted in a higher TPR for the KNN, NN and NB algorithms, but the increase in TPR was modest and AUC-values were lower than the ones found in the imbalanced dataset. For both

Table 2

True positive rate, true negative rate and area under the curve for different machine learning algorithms and different approaches to address class imbalance.

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		1:10.39	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
KNN (k = 1)	TPR (%)	8.3	13.9	16.7	16.7	57.4	50.0	36.2	8.3	8.3	8.3	8.3
	TNR (%)	90.4	75.0	75.0	74.7	62.8	69.8	75.1	90.4	90.4	90.4	90.4
	AUC	.49	.44	.46	.47	.61	.60	.56	.53	.53	.46	.46
KNN (k = 3)	TPR (%)	0	11.1	5.6	5.6	53.2	24.2	14.9	0	0	30.6	30.6
	TNR (%)	99.2	87.5	89.8	96.2	63.6	85.6	87.2	100	100	62.8	62.8
	AUC	.47	.42	.34	.42	.60	.58	.53	.48	.48	.47	.47
KNN (k = 5)	TPR (%)	2.8	11.1	2.8	0	48.9	12.9	10.6	0	0	8.3	44.4
	TNR (%)	100	86.1	95.4	99.3	63.6	91.7	94.1	100	100	91.2	43.9
	AUC	.45	.39	.34	.41	.61	.61	.51	.46	.46	.43	.46
DT	TPR (%)	0	5.6	0	0	30.9	3.2	0	0	0	8.3	11.1
	TNR (%)	100	96.5	100	100	90.1	98.4	100	100	100	94.4	92.0
	AUC	.46	.39	.46	.46	.70	.50	.47	.46	.46	.45	.38

Table 2 continued

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		1:10.39	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
RI	TPR (%)	0	5.6	0	2.8	27.7	4.8	4.3	0	0	16.7	8.3
	TNR (%)	100	90.3	95.4	99.3	90.4	97.3	97.6	100	100	94.4	92.2
	AUC	.39	.45	.36	.44	.64	.52	.57	.46	.46	.46	.47
NN	TPR (%)	2.8	13.9	5.6	5.6	45.7	25.8	17.0	0	0	5.6	13.9
	TNR (%)	93.9	79.9	90.3	91.3	84.8	89.6	93.3	100	99.7	92.0	88.8
	AUC	.50	.42	.46	.50	.68	.61	.60	.47	.49	.51	.49
NB	TPR (%)	0	11.1	2.8	2.8	45.7	30.6	6.4	0	0	11.1	22.2
	TNR (%)	98.9	77.1	95.8	97.6	82.6	89.8	95.5	99.7	99.7	88.2	77.0
	AUC	.45	.42	.39	.42	.69	.60	.53	.45	.45	.45	.45
SVM	TPR (%)	0	0	0	0	0	0	0	0	0	0	0
	TNR (%)	100	100	100	100	100	100	100	100	100	100	100
	AUC	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Note. KNN = *k*-nearest neighbourhood DT = Decision tree RI = Rule induction NN = Neural network NB = Naive Bayes SVM = Support Vector Machine TPR = True positive rate TNR = True negative rate AUC = Area under the curve SMOTE = Synthetic Minority Oversampling Technique C_{FP} = cost of a false positive C_{FN} = cost of a false negative SJT score calculated using the raw

oversampling and undersampling, ratios that included more unprofessional students resulted in a higher TPR and a lower TNR, but higher AUC-values were only found for the oversampling method. Within the oversampled datasets using a 1:4 ratio, the highest AUC-value (.70) was found for the DT algorithm, showing a TPR of 30.9% and a TNR of 90.1%. In the oversampled datasets using ratios 1:6 and 1:8, the largest AUC-values were found for the NN algorithm (AUC = .61 for 1:6 and AUC = .60 for 1:8). For all three ratios, the highest TPR was found for the KNN algorithm using k = *1*.

Cost matrices that assigned higher costs to false negatives (FN, i.e. incorrectly classified professional students) than to false positives (FP, i.e. incorrectly classified unprofessional students) resulted for most algorithms in a TPR of 0% and a TNR of 100%. In other words, all applicants were classified as professional, indicating that these ratios of misclassification costs served the interest of the medical school applicants. In contrast, cost matrices that served the interest of the medical school (i.e. CFP > CFN), led to a slightly higher TPR, but the increase was small in comparison to the resampling methods. An exception is when an FP was considered five times more costly than an FN and when using the KNN algorithm with $k = 5$, which resulted in a TPR of 44.4%, but at the expense of a TNR of 43.9%. In general, cost-sensitive learning had more negative consequences for the TNR than resampling. The different cost matrices had a negligible effect on the AUC-values of the algorithms. Finally, the SVM algorithm classified all instances as 'professional' and was not affected by any of the approaches addressing class imbalance.

To summarise, oversampling was a more effective approach to address class imbalance than undersampling or cost-sensitive learning, evidence by a higher TPR and no large detrimental effect on the TNR of the different algorithms.

Feature selection

For all resampling approaches, the ranking of the individual features indicated the participation in the coaching day as the most relevant feature for the classification of unprofessionalism (Table 3), followed by the personality trait Emotionality and the score on the cognitive admission test on scientific reading. Exceptions were the top rankings for the oversampled datasets with ratio 1:4 (1. coaching day, 2. pu-GPA and 3. SJT) and 1:6 (1. pu-GPA, 2. coaching day and 3. scientific reading). Participation in the coaching day was also the most selected feature in the subset evaluation for the classification of unprofessionalism (Table 4). Again, exceptions in the feature subset were found for the oversampled datasets with ratio 1:4 (subset consisting of pu-GPA, raw SJT score based on all response options, raw SJT score based on desirable response options, dichotomous SJT scores based on desirable response options, Emotionality, Openness to experience and coaching day) and 1:6 (subset consisting of raw SJT score based on desirable response options and coaching day). Scientific reading was excluded when applying subset evaluation. Overall, participation in the coaching day was consistently selected as the most relevant feature in the classification of unprofessional behaviour.

Different SJT versions

In general, no large disparities were found between the three different SJT scoring methods (see Appendix 6A). A slightly higher TPR was obtained when the SJT score was based on desirable response options than when the SJT score was based on all or on the undesirable response options. The SJT was only included in one of the top-three rankings of individual features when the SJT was scored using the raw consensus method based on all response options (see Table 3). See Appendix 6B for the top-three rankings for the other SJT versions. For the subset evaluation, SJT versions were included in the feature subset for the

oversampled dataset with ratio 1:4 (raw SJT score based on all response options, raw SJT score based on desirable response options and dichotomous SJT score based on desirable response options) and ratio 1:6 (raw SJT score based on desirable response options). Appendix 6C and 6D provide a graphical presentation of the TPR for each combination of machine learning algorithm and approach to address class imbalance.

Table 3
First, second and third position in the individual feature ranking for the different resampling methods.

		Undersampling			SMOTE oversampling		
Feature	1:10.39	1:4	1:6	1:8	1:4	1:6	1:8
Pre-university GPA					2	1	
Extracurricular activities							
Logical reasoning							
Scientific reading	3	3	3	3		3	3
Anatomy							
Mathematics							
Curriculum sample							
SJT					3		
Honesty-humility							
Emotionality	2	2	2	2			2
Extraversion							
Agreeableness							
Conscientiousness							
Openness to experience							
Coaching day	1	1	1	1	1	2	1

Note. GPA = Grade Point Average SJT = Situational Judgement Test SJT score calculated using the raw consensus method, based on all response options SMOTE = Synthetic Minority Oversampling Technique

Table 4
Features selected using subset evaluation for the different resampling methods.

		Undersampling			SMOTE		
Feature	1:10.39	1:4	1:6	1:8	1:4	1:6	1:8
Pre-university GPA					×		
Extracurricular activities							
Logical reasoning							
Scientific reading							
Anatomy							
Mathematics							
Curriculum sample							

Table 4 continued

		Undersampling			SMOTE		
Feature	1:10.39	1:4	1:6	1:8	1:4	1:6	1:8
SJT raw consensus							
all resp. opt.					×		
desirable resp. opt.					×	×	
undesirable resp. opt.							
SJT stan. consensus							
all resp. opt.							
desirable resp. opt.							
undesirable resp. opt.							
SJT dich. consensus							
all resp. opt.							
desirable resp. opt.					×		
undesirable resp. opt.							
Honesty-humility							
Emotionality					×		
Extraversion							
Agreeableness							
Conscientiousness							
Openness to experience					×		
Coaching day	×	×	×	×	×	×	×

Note. GPA = Grade Point Average SJT = Situational Judgement Test resp. opt. = response options stan. = standardised dich. = dichotomous SMOTE = Synthetic Minority Oversampling Technique

Discussion

The aim of this study was to examine the application of machine learning to the problem of selecting medical school applicants on a low base rate criterion. Specifically, we applied six algorithms and three approaches to address class imbalance in order to classify professionalism among first-year medical students based on a set of admission features. The assessment of professional behaviour included only a small number of ratings (i.e. 8.8%) which indicated that students' professional behaviour 'deserved attention'. This low base rate of unprofessional behaviour adversely affects its prediction or classification, which was indicated by the absence of significant predictors in a traditional logistic regression analysis. This confirms earlier findings that in logistic regression, the coefficient of determination (i.e. R^2) is highly dependent on the base rate (Menard, 2000). In addition, the classification performance of the machine learning algorithms in the imbalanced dataset demonstrated that almost all applicants were classified as professional, leading to a near-perfect overall classification accuracy, but an extremely low classification accuracy for the group of unprofessional students (that is, low true positive rate, TPR). Practically, poor classification performance regarding the low-base-rate class is highly undesirable, since it is typically the correct classification of this rare class (such as a rare disease, a fraudulent money transaction,

an unprofessional student) that is of interest in machine learning applications (Guo et al., 2008). In sum, our results confirm the problems recognised by other researchers relating to the low base rate of the outcome criterion and stress the importance of searching for ways to address the base rate problem.

In the present study, we used three different approaches to address the class imbalance in our dataset: undersampling, oversampling and cost-sensitive learning. Synthetically oversampling minority instances (that is, the rare class) by SMOTE turned out to be the most effective approach, as it led to the highest TPR, while not greatly reducing the true negative rate (TNR). This finding is in line with previous studies indicating that oversampling outperforms undersampling (García et al., 2012; Marqués et al., 2013). A substantial advantage of using SMOTE to oversample minority instances is that, instead of undersampling by arbitrarily removing data, it informatively enlarges the data of interest (Japkowicz & Stephen, 2002). This was evidenced by a higher TPR when using 1:4 synthetic oversampling than when using 1:4 random undersampling for most algorithms. Smaller ratios (1:6 and 1:8) resulted in a lower TPR for both resampling methods, but whereas the TPR, when using 1:6 and 1:8 *under*sampling, equals the TPR in the imbalanced dataset, 1:6 and 1:8 *over*sampling reaches a TPR well above the TPR found in the imbalanced dataset. In particular, for the KNN algorithm using $k = 1$, an increase in the average TPR (i.e. averaged across the nine different SJT versions) from 8.9% to 33.3% is achieved by the addition of eleven synthetic minority instances using 1:8 oversampling. This finding indicates that the classification accuracy regarding unprofessional behaviour among medical students may be improved by using only modest oversampling, which produces a ratio that might still be considered realistic for the unprofessionalism criterion (i.e. 1:8). Overall, controlled and informed oversampling of minority instances seems more sensible than the extensive removal of majority instances, as was done in previous retrospective case-control studies (Papadakis et al., 2004; Papadakis et al., 2005).

In addition to the resampling methods, we used cost-sensitive learning to address the class imbalance in our dataset. The ratio of misclassification costs served either the interest of the medical school applicant (i.e. misclassifying a professional applicant – a false negative – is more costly) or the interest of the medical school (i.e. misclassifying an unprofessional applicant – a false positive – is more costly). Clearly, attributing a higher cost to making false negatives resulted in a TPR of 0% for all algorithms because all instances were classified as professional, except for the *k*-nearest neighbourhood (KNN) algorithm using $k = 1$. Interestingly, the KNN algorithm using $k = 1$ was not affected by any of the misclassification cost ratios, indicating that CostSensitiveClassifier was unable to incorporate different misclassification costs in the KNN algorithm, when using one neighbour. In contrast, the KNN algorithm using $k = 3$ or $k = 5$ resulted in the highest TPR of all algorithms, when attributing a higher cost to false positives. Thus, in this study, KNN was the algorithm most affected by cost-sensitive learning, except when $k = 1$. In general, the TPR values using cost-sensitive learning are lower than when using synthetic oversampling, but are at least as high as when using random undersampling. However, prior research showed that whether oversampling outperforms cost-sensitive learning may depend on the size of the dataset (McCarthy et al., 2005). Future research should demonstrate if larger datasets lead to higher classification accuracy when using cost-sensitive learning. Better classification performance for cost-sensitive learning is desirable because resampling involves the alteration of the natural class distribution. In contrast, a disadvantage of cost-sensitive learning is that it reduces the TNR to a greater extent than the resampling methods. A limitation of both approaches is that they require the determination of the minority:majority ratio or the ratio of

misclassification costs, which often involves a somewhat arbitrary choice (Weiss & Hirsh, 2000).

Overall, the KNN and NN algorithms resulted in the highest TPR values (KNN using k = 1: 8.3% ≤ TPR ≤ 57.4% and NN: 0% ≤ TPR ≤ 45.7%, for the different approaches to address class imbalance), whereas the DT, RI and SVM algorithms led to the lowest TPR values (maximum TPR = 30.9%). It appears that the classification of unprofessional behaviour based on the included admission features could not be captured using straightforward if-then pathways using the DT or RI algorithms. Additionally, the SVM algorithm using a polynomial kernel was not able to create a hyperplane to separate the minority from the majority instances. Moreover, none of the approaches to address class imbalance affected the classification performance of the SVM algorithm. In contrast, the KNN and NN algorithms appear to provide sufficient flexibility to capture the underlying patterns in our dataset.

Feature selection – both individual feature ranking and subset evaluation – pointed out the participation in the coaching day as the most relevant feature for the classification of unprofessional behaviour in medical school. Contrary to the other admission features included in our dataset, the voluntary participation in a free-of-charge information day about the admission procedure involves an actual behaviour as opposed to test performance or self-report data. This result is in line with the finding of Yates (2011) who showed that students who fail to complete their Hepatitis B vaccination (an actual behaviour) will have more trouble to complete medical school. It seems that behavioural indicators may provide better predictors for future unprofessional behaviour than instrumental or self-reported indicators. Further research should investigate what samples of actual behaviour might serve as indicators of future performance. However, admission committees should be transparent about the behavioural indicators used for selection and should, therefore, take into account the effect of this transparency on the predictive effectiveness of these indicators.

No substantial differences in the TPR of the algorithms were found between the different SJT versions when examined in combination with the other admission features. The only noticeable difference for the KNN algorithm was a higher TPR for the SJT scores based on the desirable response options than for the SJT scores based on all response options or on the undesirable response options. In addition, two SJT scores based on the desirable response options (i.e. raw and dichotomous) were included in the feature subset when using 1:4 oversampling and one SJT score based on desirable response options (i.e. raw) was included in the feature subset when using 1:6 oversampling. Interestingly, in previous research, SJT scores based on the correct identification of 'what *not* to do' had stronger predictive validity than SJT scores based on the correct identification of 'what to do' (Elliott et al., 2011; Stemler et al., 2016). This difference was explained by the existence of higher consensus on what is considered an inappropriate response than on what is considered an appropriate response. The reason for this is that inappropriate responses usually lead to a negative outcome, whereas multiple ways exist to appropriately respond to a problem, but the chosen response depends on personal style and preference. Based on this reasoning, we argue that the ability to correctly identify undesirable response options might have a stronger cognitive loading than the ability to correctly identify desirable response options. Thus, the contradicting finding with regard to the predictive effectiveness of the SJT score, based on desirable response options, is likely explained by the fact that the set of admission features used in this study already consists of a number of cognitive predictors. Consequently, a cognitively-loaded SJT score, based on undesirable response options, increments less to the classification accuracy over and above the other admission features.

Another explanation for the absence of the SJT score, based on undesirable response options, in the feature subset is that the outcome measure of this study, that is, the rating of

students' professional behaviour, is mainly focused on the presence of desirable behaviours. An alternative outcome measure focused on the presence of undesirable behaviours, for example student records concerning unprofessional behaviours, might lead to the inclusion of the SJT score based on undesirable response options in the feature subset.

Practical implications

As machine learning is more and more often applied in the field of personnel selection (Liem et al., 2018), it will likely only be a matter of time until machine learning is incorporated in medical school admissions. The present study provides researchers and practitioners in medical school selection an introduction into machine learning and demonstrates the practical application of machine learning to real world data.

Limitations

This study is not without limitations. Firstly, despite improvements in the TPR induced by the resampling methods and cost-sensitive learning, TPR values did not surpass 74.5% (found for KNN algorithm using $k = 1$, 1:4 oversampling, raw SJT score based on desirable response options). This not so high TPR is disturbing given the importance of an early detection and prevention of problems associated with unprofessionalism among medical students (e.g. misconduct). Better performing algorithms are only achievable if considerable effort is put in the development and collection of reliable and valid input and output features. For instance, the outcome measure in this study, a 'deserves attention' rating on professional behaviour, might be suboptimal due to teachers' reluctance to fail students on professional behaviour or due to the sensitivity of professionalism evaluations to subjectivity. More nuanced and objective measures of professionalism, for instance by using a combination of tools (Van Mook et al., 2009), may improve the classification accuracy of the algorithms used in this study.

Secondly, machine learning provides new opportunities for medical school admissions, but does have its own challenges (Rowe, 2019). For instance, machine learning consists of a very broad range of algorithms to capture underlying patterns in the data, but many of these algorithms are hard for users to grasp, creating a black box that may hinder the transparency of a selection procedure. For example, the interpretation of an algorithm like neural networks is more complicated than the interpretation of a regression analysis, which can be described in terms of positive and negative associations. In addition, machine learning applications are primarily data-driven and so the theoretical understanding of the results is often omitted. Finally, if implicit bias exists in the training data that is presented to the algorithm, then this bias may be unintentionally incorporated into the algorithm, potentially leading to results at odds with the aims of a diverse student population and widening access to medical school (Chander, 2016). For example, Chander (2016) describes how an algorithm used for college admissions based on future job performance might become biased if the job market itself favours White employees. Due to these challenges of machine learning, it is important that researchers and practitioners keep a close eye on what goes in and what comes out the algorithm.

Conclusion

This is one of the first studies which used machine learning in medical school selection. We compared six algorithms of which the *k*-nearest neighbourhood and neural networks algorithms showed the most promising results with regard to the correct classification of unprofessionalism in medical students. Additionally, we applied three approaches to address the class imbalance in our dataset, due to the low occurrence of unprofessional behaviour

among medical students. Synthetically oversampling minority instances resulted in the largest improvement in the classification of unprofessional students. Regarding the admission features, participation in the coaching day was the most relevant feature for the classification of unprofessionalism, suggesting that indicators of actual behaviour may be the best predictors for future unprofessional behaviour. In addition, subset evaluation of the synthetically oversampled datasets revealed that pre university-GPA, Situational Judgement Test score and the personality traits Emotionality and Openness to experience may also be relevant to future unprofessional behaviour. In sum, this study gave a first look into how machine learning might be applied to medical school selection and paves the way for other researchers to investigate the application of machine learning for the purpose of selecting medical students.

Chapter 7

General Discussion

In admission procedures to medical schools, traditional cognitive tests are increasingly complemented by Situational Judgement Tests (SJTs) to expand the desired applicant profile with noncognitive attributes. The inclusion of noncognitive attributes in the admission process reflects their fundamental importance in the performance of medical students and doctors. The provision of high-quality care to patients and their families makes great demands on attributes such as integrity. In this thesis, integrity is defined as the presence of traits such as sincerity, fairness and modesty (Ashton & Lee, 2005) and the absence of self-centred attitudes, thoughts and beliefs (Barriga & Gibbs, 1996). Due to its relevance, medical school admissions have witnessed an increase in the use of instruments that focus on measuring integrity. In general, the promising psychometric findings on SJTs in the field of personnel selection (Clevenger et al., 2001; McDaniel et al., 2007) have been replicated in academic admission settings (Lievens, 2013; Oswald et al., 2004). Nevertheless, SJTs in medical school admissions are not as well established as tests of cognitive abilities and clinical aptitudes, as is manifested by relatively scarce research on how SJT characteristics affect the quality of an SJT in medical school admissions.

The aim of this thesis was, therefore, to examine the influence of several SJT characteristics on different quality criteria of an SJT used for the measurement of integrity among medical school applicants. The SJT characteristics included are the following: the development method, the response format, the scoring method, the response appropriateness and the response instructions. The quality criteria involved reliability, construct validity, criterion-related validity, subgroup differences, fakability and applicant perceptions. The findings for each SJT characteristic are briefly summarised in Table 1. This general discussion starts with an elaboration on two central themes of this thesis, namely the scoring method and the response appropriateness, since these two SJT characteristics are recurring topics throughout this thesis. Next, methodological and ethical complications that may arise when actually selecting and rejecting applicants based on their SJT scores are considered. The general discussion concludes with the limitations of this thesis and directions for future research.

Table 1
Short overview of the findings for each SJT characteristic examined in this thesis.

Development method	A theoretical, deductive development method combined with an empirical, inductive development method generates an SJT that has convergent and discriminant validity.
Response format	Rating formats are perceived more favourably than pick-one formats, especially by first-generation university applicants.
Scoring method	Raw consensus scoring methods have higher internal consistency, lower construct validity and obscure faking effects in comparison to standardised and dichotomous consensus scoring methods. Raw and dichotomous scores are more often included in the subset of relevant features for classifying future unprofessional behaviour than are standardised scores.
Response appropriateness	SJT scores based on 'what *not* to do' have stronger construct validity and are less affected by faking than SJT scores based on 'what to do', but are less often included in the subset of relevant features for classifying future unprofessional behaviour.
Response instructions	Should-do instructions are perceived more favourably than would-do instructions, especially by applicants of a non-Western ethnic background.

Scoring method

A recurring theme in this thesis is that conclusions on the quality of an SJT are greatly impacted by the method used for scoring the test. SJTs present applicants with dilemma-like situations together with response options that are deliberately ambiguous with regard to their appropriateness. This ambiguity advocates the use of rating, as opposed to pick-one or ranking, response formats for SJTs, because these formats allow applicants to give more nuanced responses and avoid the suggestive presence of one ultimate correct answer. The use of rating response formats is further supported by empirical findings showing that SJTs using rating formats are more strongly related to noncognitive constructs than are multiple-choice and ranking formats (Arthur et al., 2014). In addition, the results of Chapter 5 indicate that applicants have more favourable perceptions of rating response formats than of multiple-choice response formats. Despite their advantages, rating response formats do pose new difficulties in scoring SJTs because many different ways exist to quantify the similarity between an applicant's rating and the scoring key. Below, we first discuss the development of the scoring key of an SJT, followed by a deliberation on the quantification of the similarity between ratings.

The *first* step in the rational scoring of an SJT is to gather a set of ratings that can be used to score the test, thus forming the scoring key. In contrast to most previous SJTs used in medical school admissions, we did not only use Subject Matter Experts (SMEs) to develop a scoring key, but also created a scoring key based on the answers of the applicants themselves (Chapter 2). Although it may intuitively seem inappropriate to let non-experts determine the 'correct' answers to the SJT items, it must be emphasised that SJTs, like the one used in this thesis, have no 'correct' answers. The absence of definite correct answers is particularly true for rating SJTs for which an applicant does not select a single – 'correct' or 'incorrect' – answer but rather chooses a rating scale point. Additionally, high correlations have been found between expert and non-expert scoring keys themselves, and between the SJT scores based on the respective scoring keys (Legree et al., 2010; Legree et al., 2005; Motowidlo & Beier, 2010). The results of Chapter 2 also indicate that a scoring key based on SMEs does not differ from a scoring key based on the applicants themselves with respect to internal consistency reliability, ethnic subgroup differences and correlations with personality.

Moreover, for dilemmas like those described in SJT scenarios, even SMEs tend to disagree on the appropriateness of the response options. For instance, for some response options of the integrity SJT in Chapter 3, the ratings of SMEs are evenly spread across – a part of – the rating scale points. Additionally, expert-based scoring keys may be highly context-dependent. For example, in Chapter 2, a group of Dutch SMEs (i.e. individuals involved in teaching professionalism in the medical curriculum) created a scoring key for the integrity-based SJT that was developed in Scotland. A comparison of the ratings of Dutch and Scottish SMEs revealed that response options which describe the reporting of unprofessional behaviour of others to a supervisor were deemed more appropriate by Scottish SMEs than by Dutch SMEs. In such cases, a large group of non-experts might be more useful than a small group of experts, because a large number of ratings may lower the variance and, consequently, result in a more stable scoring key. Besides the generally small number of SMEs, it is often unclear who should be considered as experts. For the SJT developed in Chapter 3, the SMEs were comprised by residents in training to become general practitioners, because this group had successfully completed medical school and had followed a general training. In contrast, the group of Scottish SMEs comprised staff members from the undergraduate medical education centre of the university (Husbands et al., 2015). However, fair arguments could have been made for other individuals to serve as SMEs, for example graduate students or medical ethicists. Therefore, in the absence of an obvious group of

SMEs, we argue that it is more defensible to use a large group of non-experts as opposed to a small group of SMEs.

After establishing the set of ratings to which the applicants' ratings will be compared, the *second* step is to quantify the similarity between both sets of ratings. In Chapter 2, we used several methods to quantify this similarity based on either distance (for instance the absolute distance between an applicant's rating and the mean rating of the SME group) or endorsement (for instance the rating scale point that is endorsed by at least 50% of SMEs yields two points). The results of Chapter 2 indicated that the distance-based scoring methods showed higher internal consistency reliability and somewhat stronger correlations to personality than did the endorsement-based scoring methods. Nevertheless, distance scores are not without limitations. The main critique on distance or difference scores is that it is often not justified to create a new, unique concept by calculating the difference between two other concepts (Edwards, 2001). Other problems with difference scores involve the inability to take into account the direction of the difference (in other words, whether the applicant's rating is lower or higher relative to the scoring key) and the absolute level of the components scores underlying the difference score. That is, a maximum difference score is obtained when the applicant's rating equals the scoring key, regardless of the position of both ratings on the rating scale (Edwards, 1993). Additionally, difference scores have a number of methodological problems such as low reliability and restricted variance (Peter, Churchill, & Brown, 1993). These problems call for a closer examination and better understanding of distance-based SJT scores in future research.

A recently suggested approach that may lead to a better understanding of distance-based SJT scores is to disentangle the separate profile similarity metrics of an SJT score (Legree, Ness, Kilcullen, & Koch, 2018). Profile similarity metrics are different measures to quantify the similarity between two sets of ratings, consisting of *shape*, *elevation* and *scatter* (Cronbach & Gleser, 1953). Shape refers to the correlation between the applicant's set of ratings and the scoring key, elevation is the applicant's average position on the rating scale and scatter, or dispersion, describes the variance in the applicant's ratings (Legree et al., 2010). One problem of distance scores is that they do not differentiate between these profile similarity metrics (Legree et al., 2018). This inseparability of profile similarity metrics is undesirable because elevation and scatter are strongly influenced by applicants' response tendencies to use rating scales in a certain manner (for instance, a tendency to use a particular part of the rating scale). Response tendencies are generally irrelevant to the constructs measured by SJTs and are, therefore, viewed as a source of systematic error (McDaniel et al., 2011). In addition, elevation and scatter can be easily altered by response strategies (Legree et al., 2010). Ideally, the distance-based scores should reflect the shape of the similarity between ratings, because the shape is not affected by response tendencies.

In Chapter 2, we removed elevation and scatter from the SJT score by applying standardised and dichotomised consensus scoring methods. Standardised and dichotomous consensus scoring methods resulted in lower internal consistency reliability, caused by a reduction in scale variance. Yet, standardised consensus scores did result in slightly higher correlations to personality than the other scoring methods applied in Chapter 2, confirming the stronger validity findings found in previous research (McDaniel et al., 2011; Weng et al., 2018). Despite the potential beneficial effect of removing elevation and scatter on the construct validity of a test, Cronbach and Gleser (1953) warned researchers for eliminating elevation and scatter, because these metrics may contain valuable information. For instance, individual differences in elevation (that is, an applicant's average position on the rating scale) are probably not only a result of individual differences in response tendencies, but also of actual individual trait differences. The warning of Cronbach and Gleser (1953) may explain

the findings in Chapter 6, in which the standardised SJT score was not included in any of the feature subsets which were relevant for the classification of future unprofessional behaviour. We hypothesise that the elimination of elevation and scatter by the standardised and dichotomous scoring methods may have been too rigorous. Therefore, an interesting approach to gain more insight into distance-based SJT scores – beyond the findings presented in Chapter 2 – is to conduct a more detailed examination of the separate profile similarity metrics, as described by the above mentioned paper of Legree et al. (2018). Legree et al. (2018) showed that an optimally weighted combination of separate profile similarity metrics provides incremental validity above a distance-based SJT score in the prediction of a criterion. In sum, the scoring of rating SJTs may benefit from a more thorough investigation of the individual contribution of each profile similarity metric that underlies the distance score, as opposed to the definite elimination of one or more metrics.

Unfortunately, profile similarity metrics have their own disadvantages. For example, these metrics are still unable to take into account the direction and absolute level of applicants' ratings relative to the scoring key (Edwards, 1993). Additionally, using a difference term in a regression equation imposes a set of constraints on the equation, caused by pairing the unconstrained component variables in one difference variable (Edwards, 1994). These constraints involve assumptions that the regression coefficients of both applicant and scoring key ratings are equal in size but opposite in sign and that the coefficients of each pair of ratings are similar (Edwards, 1993). These restrictive assumptions are almost never met in practice.

Therefore, next to profile similarity metrics, another interesting method to gain more insight in distance-based SJT scores may be polynomial regression analysis. Polynomial regression analysis includes the component scores that make up the difference score and their higher-order terms as predictors, and explicitly tests the above constraints assumed by difference scores (Edwards, 2001). Polynomial regression analysis of the disentangled components of a difference score enables the investigation of more meaningful relationships with an outcome measure than solely including the difference score in traditional regression analysis (Edwards, 1994). We are not aware of previous research using polynomial regression analysis in the scoring of SJTs. Overall, both profile similarity metrics and polynomial regression analysis teach us that the use of rating SJTs in research and practice may take advantage of unravelling the separate elements of an SJT score, either in the form of similarity metrics or regression terms.

In conclusion, the studies reported in Chapter 2, 4 and 6 were carried out to find the most optimal scoring method for rating SJTs. However, based on the findings of these chapters and prior research, it appears that such a "perfect" SJT scoring method does not exist. In fact, the ideal scoring method depends on factors related to the test, context and sample. For instance, the rigorous elimination of elevation and scatter may be less relevant in low-stakes situations, where respondents may be less inclined to use response strategies (for instance, extreme responding), than in high-stakes situations. Scoring an SJT should, therefore, include a preliminary phase in which different scoring methods (differentially weighting the score elements, such as elevation, scatter and shape) are run on the data and the scoring method that results in the highest test quality is picked. An essential requirement for optimally weighting separate score elements based on empirical data is the availability of adequate criterion data, a requirement that is unfortunately not always a guaranteed certainty. Finally, to avoid that the scoring method capitalises on a specific dataset, it is recommended to cross-validate the chosen scoring method on an independent dataset.

Response appropriateness

The second recurring theme of this thesis is the distinction made between response options describing 'what to do' and response options describing 'what *not* to do'. In Chapter 3, this distinction was first made based on two theoretical models that were related to integrity, either positively (i.e. honesty-humility dimension of the HEXACO personality scale (Ashton & Lee, 2005)) or negatively (i.e. cognitive distortions of the How I Think questionnaire (Barriga & Gibbs, 1996)) related to integrity. The findings of Chapter 3 indicated that an SJT score based on *un*desirable response options that describe 'what not to do' is more strongly correlated to scores on four external integrity-related questionnaires than an SJT score based on desirable response options that describe 'what to do'. In addition, the results of Chapter 4 demonstrated a smaller faking effect for an SJT score based on *un*desirable response options than for an SJT score based on desirable response options. In contrast, Chapter 6, which focused on the predictive validity of the integrity SJT, showed that an SJT score based on desirable response options was more often included in the optimal feature subset for classifying unprofessional behaviour among first-year medical students than an SJT score based on undesirable response options.

The above findings support the notion that respondents do not only retrieve information from SJT scenarios, but also from the SJT response options themselves (Harris, Siedor, Fan, Listyg, & Carter, 2016). Rockstuhl, Ang, Ng, Lievens, and Van Dyne (2015) even argue that SJTs do not assess the judgement of situations, but that SJTs assess the judgement of response options. Stated differently, according to these researchers SJTs instruct respondents to judge the appropriateness of response options, but not to evaluate the situations in the presented scenarios (for example, the intentions, emotions and thoughts that help to make sense of a situation). Apparently, response options provide an important source of information used by applicants when responding to an SJT (Kaminski, Felfe, Schäpers, & Krumm, 2019).

The relevance of the type of response options to the quality of an SJT is demonstrated by the studies in this thesis. The types of SJT response options that can be distinguished have been described from different perspectives. For instance, Stemler et al. (2016) make a distinction based on the performance-approach and performance-avoid motivational theory of Elliot and Church (1997), which states that people are generally motivated to either approach success or to avoid failure. These authors suggest that an SJT focused on the correct identification of 'what to do' (i.e. 'approach' response options) measures a different skill (i.e. knowledge of behaviours that will lead to desirable events) than an SJT focused on the correct identification of 'what not to do' (i.e. 'avoid' response options). In addition, response options are not only distinguishable based on their professional desirability, but also based on their social desirability and plausibility (Kaminski et al., 2019).

Yet another aspect in which response options may differ is their cultural content (e.g. collectivism versus individualism), which may lead to cultural subgroup differences in responses to an SJT (Schmitt, Prasad, Ryan, Bradburn, & Nye, 2019). For instance, response options that were more strongly related to culture (as judged by an independent group of experts) showed larger differences in responses between Chinese and Caucasian American college applicants. Interestingly, in the study of Schmitt et al. (2019), the cultural content of response options had a stronger effect on subgroup differences for 'most likely' options (i.e. 'what to do') than for 'least likely' options (i.e. 'what not to do'). It appears that cultural differences are more profound in what is considered socially desirable. In sum, the response options of an SJT may differ in their expression of approach motivation, avoid motivation, professional desirability, social desirability, plausibility and cultural content.

As mentioned above, an SJT score based on response options describing 'what not to do' had stronger construct validity than an SJT score based on response options describing 'what

to do' (Chapter 3). One explanation for this difference is the existence of more consensus on what is considered to be an incorrect action than on what is considered a correct action in solving a challenging situation (Elliott et al., 2011; Stemler et al., 2016). In general, SJT response options are ambiguous in their appropriateness, but the degree of ambiguity may vary across response options. Inappropriate responses are presumably less ambiguous than appropriate responses, possibly because inappropriate responses are more obviously incorrect, whereas the correctness of appropriate responses is more dependent on respondents' personal preferences, style and culture. For instance, how to correctly solve a conflict with a fellow student is dependent on the respondent's personality and cultural background, that is, a particular appropriate response is not obviously more or less correct than another appropriate response. In contrast, inappropriate responses, such as getting aggressive with the fellow student, is obviously inappropriate regardless of an individual's personality. We argue that due to the larger consensus about the incorrectness of *in*appropriate responses, these response options can be considered knowledge-based items, which are more cognitively loaded. In contrast, since multiple *appropriate* responses exist which may result in a good outcome, the correctness of these response options is not only determined by knowledge, but also by personal taste and cultural customs. Therefore, appropriate response options may be considered behavioural-tendency items, which are less cognitively loaded.

A better understanding of the distinction between knowledge-based and behavioural-tendency items may be obtained from the theory of cognitive acuity (Leeds, 2012). Leeds (2012) defines cognitive acuity as "the capacity to detect correctness and to distinguish between differences in correctness among simultaneously presented situation-specific response options" (p. 166). In this theory, response options are perceived as emitting a correctness signal to which respondents – depending on their cognitive acuity – are more or less sensitive. The theory of cognitive acuity may shed more light on the differences in response options describing 'what to do' and 'what not to do', in the sense that undesirable response options may emit a stronger (in)correctness signal than desirable response options. The stronger correctness signal emitted by undesirable response options may point to the stronger cognitive loading of these knowledge-based response options. The premise that inappropriate response options have stronger cognitive loading than appropriate response options was supported by the smaller faking effects on an SJT score based on options describing 'what not to do' found in Chapter 4, since knowledge-based tests are presumed to be less fakable than behavioural-tendency tests.

Summarising this section, we conclude that research on SJTs should not only focus on the scenarios of the test, but also on the response options. To make optimal use of the information that is contained in the response options, SJTs should use response formats that require separate responses to each response option, such as rating formats instead of pick-one formats. Additionally, analysis at the level of response options allows the examination of applicants' knowledge about inappropriate responses as well as applicants' preferred choice of action to reach desirable outcomes. Both SJT subscores (that is, 'what to do' and 'what not to do') appear to be informative. Indeed, Chapter 3 indicated that an SJT score based on inappropriate response options provides a cleaner measurement of the construct of interest. In contrast, Chapter 6 showed that an SJT score based on appropriate response options might make a greater contribution to the classification of future unprofessional behaviour. This seemingly contradictory finding might also be explained by the hypothesis that an SJT score based on appropriate response options has a lower cognitive loading than an SJT score based on inappropriate response options. Consequently, an SJT score based on correctly identifying 'what to do' in challenging situations may have more incremental

validity over and above the other – cognitive – instruments in the admission process than an SJT scores based on correctly identifying 'what not to do' in challenging situations. Another potential explanation is that the outcome measure used in Chapter 6, that is, a rating of professional behaviour, is mainly focused on the presence of desirable behaviours, resulting in a stronger alignment to the SJT score based on appropriate response options. An outcome measure that is more focused on unprofessional behaviours, such as records on students' professional lapses, might have a stronger association with the SJT score based on inappropriate response options.

In sum, SJT research may benefit from a model describing the various attributes of SJT response options, similar to the model on SJT attributes developed by Campion et al. (2014). Important components of this model on SJT response option attributes could be performance approach and avoid motivation (Stemler et al., 2016), professional desirability, social desirability, plausibility (Kaminski et al., 2019), cultural content (Schmitt et al., 2019) and correctness signal (Leeds, 2012).

Methodological and ethical issues in the application of an integrity-based SJT

Ultimately, the development of the SJT and the thorough examination of test characteristics affecting the SJT's quality aim to improve the selection of medical students. For this purpose, admission committees have several choices how to use an SJT in their selection procedures. The first choice concerns how to weight the SJT score in relation to the other components of the selection process. For instance, admission committees may assign the same weight to the score on an SJT as to traditional components (e.g. educational achievement) of the selection process. This choice has been made for the UK Clinical Aptitude Test (UKCAT), which is one of the methods used for selection into the foundation programme in the UK (Smith & Tiffin, 2018). Admission committees could also decide to assign a lower weight to the SJT than to the cognitive elements of the selection process, as for instance is done in medical school selection in Flanders (Lievens, 2013). Interestingly, changing the relative weights of an SJT score and a cognitive ability test score may alter the socioeconomic composition of the selected students, affecting widening access for applicants of a low socioeconomic background to medical school (Lievens et al., 2016)

The second choice concerns the role of the SJT in the selection process. For example, SJT scores may be used independently from cognitive tests scores and in a non-compensatory way, for example to shortlist or screen applicants in an initial step of medical school selection (Dore et al., 2017; Patterson et al., 2009). In contrast to selecting *in* suitable applicants, SJT scores may be used to select *out* unsuitable applicants (Powis, 2015), an option that is advocated by studies indicating that SJTs have higher predictive validity at the lower than at the higher end of the performance distribution (Cousans et al., 2017). Irrespective of their position in the selection process, SJTs should be sufficiently reliable and valid to make proper selection (or rejection) decisions in a high-stakes context. Unfortunately, the actual implementation of SJTs in medical school selection may be hampered by several *methodological* issues, which are described in the following.

SJTs in general have several limitations, such as a low internal consistency reliability and medium-sized ethnic subgroup differences, as was the case for the SJT investigated in Chapter 2. In addition, SJTs may to some degree be susceptible to faking (see Chapter 4) and receive only moderately favourable applicants perceptions (see Chapter 5). Although the studies of this thesis showed that these issues may be resolved to some extent by changing the characteristics of the test (for example, rating response formats improve applicant perceptions over pick-one response format), the ultimate proof of the usefulness of any SJT is often hindered by the suboptimal measurement of the outcome to be predicted by the SJT

in question. This suboptimal measurement of the outcome is often caused by medical school admission committees' not paying thorough attention to clearly define the criterion of interest (in our case: professionalism during medical school), an issue termed the criterion problem (Ferguson et al., 2002). Partly due to the criterion problem, it is more difficult to reliably and validly measure criteria of noncognitive admission tests. For cognitive admission tests, the predictive validity can be established by the large number of exam-based grades that are readily available in medical schools' student administration systems. In contrast, noncognitive admission tests attempt to measure skills and abilities that are not captured by traditional exam data. Instead, the criteria of noncognitive admission tests (e.g. professional behaviour in medical school) are often assessed using rater-based evaluations that are more subject to bias and, therefore, less reliable and valid (Williams, Klamen, & McGaghie, 2003). Before medical schools implement admission tests for measuring noncognitive attributes, considerable effort must therefore be spent to define and subsequently measure the criterion.

Nonetheless, even if medical schools succeed in obtaining a clear definition and valid measurement of the noncognitive criterion of interest, admission committees cannot ignore another issue, namely the adverse effect of a low base rate in the noncognitive outcome on the predictive validity of an SJT. The results of Chapter 6 indicated that the much smaller class of unprofessionally behaving medical students had to be artificially oversampled – using a factor of 2.6 – to reach a ratio of four professional students to each unprofessional student, before the SJT score was included in the variable subset that had the highest classification accuracy. Yet, for most types of unprofessional behaviours, a 1:4 ratio is not realistic. Additionally, Niessen and Meijer (2016) demonstrated that a noncognitive admission test contributes less to the utility of a traditional cognitive-based admission procedure when the base rate of unsuccessful medical students is low. Furthermore, selection of medical school applicants based on attributes that have a low base rate will ultimately lead to the rejection of many suitable applicants (Colliver et al., 2007). Nevertheless, despite its low base rate, integrity remains an essential attribute as medical students and doctors low on integrity may cause serious harm and trouble for patients and colleagues dependent on their care and collaboration. Thus, although low base rates may make some attributes less appropriate for high-stakes selection than others, the relevance of integrity to the medical profession necessitates more research on analytical methods to predict outcomes that have low base rates.

The actual implementation of an SJT in medical school selection may not only be hampered by methodological issues, but possibly also by *ethical issues*, concerning the appropriateness of selecting adolescent applicants on attributes such as integrity. Reactions to a newspaper article, which unluckily labelled the integrity-based SJT developed in Chapter 3 as the 'jerk test' ('horkentest' in Dutch) (Venema, 2016), expressed these ethical issues (Hassink, 2018; Remmerswaal, 2016; Truijens, 2016). The main objection raised against integrity-based medical school admissions is that it may be unethical to select 17-year and 18-year old applicants to higher education on an attribute that they may still be developing. This is a legitimate point of critique that touches on an important difference between personnel selection and educational admissions (Hofstee, 1990). Instead of selecting applicants who are – after a short training period – immediately ready to start a job, medical school admissions involve selecting applicants for an intensive six-year educational programme. Stated differentially, the primary goal of personnel selection is economic utility, whereas educational selection should also respect applicants' rights to education (Hofstee, 1990). Indeed, some of the attributes on which applicants are selected at the beginning of the programme will develop over time as a result of the education received. The continuous development during medical education does not only apply to noncognitive attributes, but

also applies to knowledge and skills. Therefore, cognitive and noncognitive admission tests should not aim to measure the presence of fixed attributes and skills, but should instead better aim to measure individuals' potential to develop relevant attributes and skills, that is, individuals' learning potential (Hamers & Resing, 1992). In general, the selection of adolescents on cognitive potential (e.g. using a traditional exam) is accepted, whereas the selection on noncognitive potential raises concerns, most probably due to the more personal nature of noncognitive attributes. When using either cognitive or noncognitive selection instruments, admission committees should not search for ready-to-use medical doctors, but rather identify those individuals who will benefit the most from the education offered by the medical school. Subsequently, stimulating educational programmes should bridge the gap between learning potential and actual levels of desired attributes and skills (Hamers & Resing, 1992). In this light, the selection of adolescents on noncognitive potential is not more or less unethical than the selection on cognitive potential.

Overall, although medical schools' aspirations to broaden the attributes on which applicants are selected, thus endorsing the relevance of noncognitive attributes in medical student performance, are admirable, high-stakes selection into medical school should always be an evidence-based and not a fashion-based practice (Harris, Walsh, & Lammy, 2015). Hence, even though an integrity-test such as an integrity-SJT may appear face valid, clear evidence of the test's capability to reliably and validly measure integrity is essential before the test can be used for the allocation of highly-valued medical school admission spots. However, due to the methodological and ethical issues discussed above, it may be more complicated to adequately found selection decisions on integrity-based admission tests than on traditional cognitive admission tests.

Nevertheless, based on this thesis, some suggestions can be provided that may increase the usefulness of integrity-based SJTs for high-stakes selection. The greatest challenge of SJTs and other noncognitive admission instruments (e.g. personality measures) is the self-reported nature of the measurement of noncognitive attributes. Self-report measures may have limited usefulness in high-stakes selection due to their susceptibility to faking (Donovan et al., 2014), although some other researchers argue that faking may not be a problem because it is related to job performance (Ingold et al., 2014). If faking is a problem, one suggestion to increase SJTs' applicability to high-stakes selection is to reconceptualise them as measures of knowledge about the effectiveness of expressing noncognitive attributes (Lievens & Motowidlo, 2016), as knowledge cannot be faked. Additionally, the applicability of SJTs to high-stakes situations may be strengthened by using knowledge-based response instructions and response options describing 'what not to do'. Whether these suggestions will improve the use of SJTs in the high-stakes selection of medical students must be thoroughly investigated before actual selection. Until then, SJTs may serve other valuable purposes, for example as an initial preview of challenging situations that may be encountered by medical students to help applicants make informed decisions whether or not to apply to medical school (Benbassat & Baumal, 2007) or for training or assessment of noncognitive attributes during medical school (Goss et al., 2017). An alternative option is to use an SJT as a 'red flag' by offering remedial aid and additional guidance to students scoring low on the SJT (Stegers-Jager, 2018). Finally, the SJT score could be communicated only to the applicant including an advice on that applicant's suitability to medical school (Hofstee, 2005).

In the Netherlands, the complications in designing reliable and valid admission tests of noncognitive attributes has revived an earlier discussion (e.g. Hofstee, 1983) whether other admission systems may be more desirable than selection-based admissions to higher education. Because Dutch medical school applicants comprise a highly homogeneous group with regard to educational achievement, due the selective secondary school system (Stegers-

Jager, 2018), lottery-based admissions may have certain benefits. Important advantages of lottery-based admissions involve the lower costs, the provision of equal opportunities leading to an increase in diversity and the absence of personal rejection (i.e. rejection as a result of bad luck as opposed to personal failure) (Wouters, Croiset, & Kusurkar, 2018). In contrast, a recent study indicated that selection-based admissions might be more cost-effective than lottery-based admissions (Schreurs, Cleland, Muijtjens, Oude Egbrink, & Cleutjens, 2018). Other advantages of selection-based admissions involve more positive applicant perceptions, partly due to a higher perceived chance to perform (see the results of Chapter 5), and a higher efficiency in terms of lower student drop-out and better clinical performance (Urlings-Strop, Themmen, Stijnen, & Splinter, 2011). Indeed, lottery-based admissions is not only blind to applicants' demographic characteristics, but also to variances in applicants' abilities and potential (Zwick, 2017). In fact, the low efficiency of lottery-based admissions, unweighted for pre-university GPA, is often solved by introducing thresholds (e.g. applicants with higher pre-university GPAs have a higher chance of getting admitted) or exceptions (e.g. applicants with pre-university GPAs above a certain threshold are directly admitted) to the lottery system (Zwick, 2017). Unfortunately, such solutions may hamper lottery-based admissions' ability to increase the gender, ethnic and socioeconomic diversity in the admitted student population.

Clearly, both lottery-based and selection-based admissions have pros and cons and the debate on which is best should not aim for black-and-white solutions. Regarding selection-based admissions, committees should be careful not to overestimate their ability to discriminate between suitable and unsuitable applicants. For instance, rank ordering applicants based on their admission test scores may implicate a level of accuracy that is not justified (Visser, 2017). Although a selection system may be able to accurately identify high-potential or low-potential applicants, it may be incapable to accurately discriminate applicants at other parts (e.g. at the middle) of the distribution. For those applicants, a selection system will function more or less identical to a lottery system. In that case, admission committees could better refrain from selecting applicants in this area of the distribution and it would be more reasonable and transparent to actually use lottery-based admissions to allocate those admission spots. Additionally, a survey indicated that lottery-based admissions were perceived as more appropriate for rejection, whereas selection-based admissions were perceived as more appropriate for allocation (Hofstee, 1990). An intriguing change of perspective is to reframe medical school admissions not as a problem of allocating a certain number of admission spots, but as a problem of rejecting a great number of suitable applicants. From this angle, selection-based admissions of high-potential applicants – valuing and rewarding capabilities and talents – combined with lottery-based admissions of the remaining applicants appears to be the "least unacceptable" solution to the problem of medical school admissions (Hofstee, 1983).

Limitations and directions for future research

Naturally, the findings of this thesis are limited by some boundary conditions that may be addressed in future research. This section of the discussion will elaborate on four major restraints, regarding i) construct-based development, ii) response format, iii) predictive validity and iv) research context, that should be considered when interpreting the findings of this thesis.

Firstly, we used one type of construct-based development method, in which an inductive, empirical approach and a deductive, theoretical approach were combined. Specifically, we merged input of experts including current medical students with two established theoretical models (i.e. honesty-humility dimension and cognitive distortions) in order to obtain a

construct-based SJT that has realistic content. The medium-sized convergent validity with several external measures, and the discriminant validity with an unrelated measure (i.e. self-efficacy) reported in Chapter 3 support the value of this construct-based development approach. However, the uninformative results of factor analyses of the SJT revealed that the internal test structure needs to be improved. Improving this structure might be achieved by using alternative construct-based development approaches.

For instance, Olaru et al. (2019) explicitly incorporated a construct definition in the critical incident interviews that were used to develop the items of an SJT measuring the personality construct of dependability. Evaluation of the construct validity of this dependability SJT revealed a one-dimensional factor structure and medium correlations to relevant external measures. In contrast, another approach to develop a construct-based SJT does not involve the use of critical incident interviews among subject matter experts, but rather involves test developers formulating SJT items and response options that elicit different levels of one construct (Guenole et al., 2017; Lievens, 2017). Based on this approach, Oostrom et al. (2019) developed an SJT measuring the six HEXACO personality traits, resulting in a test which displayed convergent and discriminant validity and a sufficient fit of a six-factor model. Additionally, Mussel et al. (2018) developed an SJT measuring narrow personality traits, in which each scenario was relevant for a single trait, and in which they included two response options describing high levels of that trait and two response options describing low levels of that trait. This SJT had acceptable levels of construct validity. In sum, more extensive construct-based development may help to clarify the internal structure of an SJT.

Secondly, all studies of this thesis examined an SJT with a rating response format, except for Chapter 5, which compared applicant perceptions of a rating response format with a pick-one format. Rating response formats have considerable advantages in comparison to pick-one, pick-two or ranking response formats, including higher internal consistency reliability, lower correlation to cognitive ability and smaller ethnic subgroup differences (Arthur et al., 2014). In addition, as mentioned above, rating formats enable researchers to conduct analysis at the response option level instead of the scenario level.

Nonetheless, other response formats, as opposed to the traditional formats (i.e. pick-one, ranking, rating), may further advance the use of SJTs in research and practice. For instance, single-response formats consist of only one response option which forms an inherent component of the scenario to be judged on a rating scale (Motowidlo, Crook, Kell, & Naemi, 2009). For example, an item using a single response could read: "John founds out that Mary has a copy of the exam paper that will be given next week. John doesn't look at the exam paper and informs the teacher. How appropriate is this response of John?". An advantage of the single-response format is that the rating of that response option is independent of the appropriateness of other response options. In addition, there is no need to create multiple, possibly artificial, response options that might turn out to be implausible for a particular challenging situation.

Another response format that might enhance the understanding of SJTs is the too little/too much rating scale described by Vergauwe, Wille, Hofmans, Kaiser, and Fruyt (2017), where respondents have to indicate whether the expression of a particular behaviour is ideal, too much or too little. As opposed to the traditional Likert scale, the too little/too much scale allows the differentiation between optimal and extreme, suboptimal levels of behaviour and enables the detection of a curvilinear relationship between an SJT score and the criterion measure.

Finally, an open-ended response format would completely elevate the need to develop response options and untie applicants' responses from the confinement of the given response

options. Problems concerning the standardised scoring of open responses might then be solved by increased advancements in automatic text-based analysis (Banks, Woznyj, Wesslen, & Ross, 2018; Holtrop, Breda, Oostrom, De Vries, & Stooker, 2015).

In sum, future research should examine if the use of more innovative response formats, instead of the traditional response formats, improves the quality of an SJT.

Thirdly, we did not succeed to demonstrate that the integrity-based SJT had a sufficient level of predictive validity. The results of both traditional logistic regression analysis and innovative machine learning techniques in Chapter 6 indicated that the cognitive and noncognitive admission variables had limited validity in the prediction of who will behave unprofessionally as a medical student. The relatively low predictive validity is likely caused by a combination of factors related to the instrument (SJT), the attribute (integrity), the criterion (unprofessional behaviour) and the target population (adolescents). In general, SJTs have lower predictive validity (r = .26) (McDaniel et al., 2007) than measures of general mental ability (r = .51 for jobs of average complexity) (Schmidt & Hunter, 2004). Additionally, integrity is an attribute that might be more susceptible to faking than other noncognitive attributes (Alliger & Dwight, 2000) and is conceptualised in many different ways. An unclear conceptualisation might also apply to the criterion of interest. Further, the low base rate of the criterion of unprofessional behaviour might reduce the predictive validity. Lastly, even when admissions are focused on the potential to learn noncognitive skills instead of the instant presence of these skills, it may be rather difficult to discriminate between adolescents based on their potential to benefit from education in professionalism during medical school. Evidently, although the results of this thesis provide some guidelines for enhancing the quality of SJTs, improving the selection of medical school students on integrity requires admission committees to spend considerable attention to all the above mentioned factors affecting predictive validity.

Fourthly, the results presented in this thesis were obtained in a research context in which applicants were not actually selected based on their SJT scores. Even though all data were collected in a selection context and the participants of the studies were medical school applicants and not students, the administration of the SJT solely for research purposes imposes an important limitation on the generalisability of the findings. Obviously, preliminary research on the influence of SJT characteristics on the test's quality is a prerequisite before any SJT can be used for high-stakes selection. Nevertheless, future research should attempt to replicate the findings of this thesis in a selection context.

Conclusion

Situational Judgement Tests (SJTs) in medical school admissions enable the standardised measurement of noncognitive attributes among large groups of applicants. Despite the promising findings of previous research, the reliable and valid measurement of the construct of interest is not a guaranteed SJT quality. In fact, SJTs are versatile instruments that come in different shapes and forms and these varying characteristics affect the quality of the test. The implementation of SJTs in high-stakes selection situations requires the careful examination of how SJT characteristics influence the quality of the test. The studies of this thesis describe how five test characteristics influence several quality criteria of an SJT (see Table 1). The results of these studies provide guidelines to researchers and practitioners for the development and use of SJTs in medical school admissions. Well-designed SJTs may offer a valuable contribution to medical school admissions in contexts where selection ratios are low.

References

Ababneh, K. I., Hackett, R. D., & Schat, A. C. H. (2014). The role of attributions and fairness in understanding job applicant reactions to selection procedures and decisions. *Journal of Business and Psychology, 29*, 111-129.

ABIM Foundation. (2002). Medical professionalism in the new millennium: A physician charter. *Annals of Internal Medicine, 136*, 243-246.

Ahmed, H., Rhydderch, M., & Matthews, P. (2012). Can knowledge tests and situational judgement tests predict selection centre performance? *Medical Education, 46*, 777-784.

Ainsworth, M. A., & Szauter, K. M. (2018). Student response to reports of unprofessional behavior: Assessing risk of subsequent professional problems in medical school. *Medical Education Online, 23*, 1485432.

Alliger, G. M., & Dwight, S. A. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. *Educational and Psychological Measurement, 60*, 59-72.

Alpaydin, E. (2009). *Introduction to Machine Learning*. Cambridge, MA: The MIT Press.

Anglim, J., Bozic, S., Little, J., & Lievens, F. (2018). Response distortion on personality tests in applicants: comparing high-stakes to low-stakes medical settings. *Advances in Health Sciences Education, 23*, 311-321.

Arnold, L., & Stern, D. T. (2006). What is Medical Professionalism? In D. T. Stern (Ed.), *Measuring medical professionalism* (pp. 15-37). New York, NY: Oxford University Press, Inc.

Arthur, W., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology, 99*, 535-545.

Ashton, M. C., & Lee, K. (2005). Honesty-humility, the Big Five, and the five-factor model. *Journal of Personality, 73*, 1321-1354.

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*, 150-166.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly, 48*, 491-509.

Bachman, J. G., O'Malley, P. M., & Freedman-Doan, P. (2010). *Response styles revisited: Racial/ethnic and gender differences in extreme responding*. Retrieved from https://deepblue.lib.umich.edu/bitstream/handle/2027.42/137850/occ72.pdf

Baernstein, A., & Fryer-Edwards, K. (2003). Promoting reflection on professionalism: a comparison trial of educational interventions for medical students. *Academic Medicine, 78*, 742-747.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*, 3-17.

Bandura, A. (1994). Self-efficacy *Encyclopedia of human behavior*, 71-81.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology, 33*, 445-459.

Barbot, B., Haeffel, G. J., Macomber, D., Hart, L., Chapman, J., & Grigorenko, E. L. (2012). Development and validation of the Delinquency Reduction Outcome Profile (DROP) in a sample of incarcerated juveniles: A multiconstruct/multisituational scoring approach. *Psychological Assessment, 24*, 901-912.

Barlow, C. J., Morrison, S., Stephens, H. O. N., Jenkins, E., Bailey, M. J., & Pilcher, D. (2015). Unprofessional behaviour on social media by medical students. *Medical Journal of Australia, 203*, 439-439.

Barriga, A. Q., & Gibbs, J. C. (1996). Measuring cognitive distortion in antisocial youth: Development and preliminary validation of the "How I Think" questionnaire. *Aggressive Behavior, 22*, 333-343.

Barriga, A. Q., Gibbs, J. C., Potter, G. B., & Liau, A. (2001). *How I Think (HIT) questionnaire manual*. Champaign, IL: Research Press.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*, 20-29.

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

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology, 54*, 387-419.

Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment, 13*, 225-232.

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

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology, 85*, 349-360.

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.

Binder, R., Friedli, A., & Fuentes-Afflick, E. (2015). Preventing and managing unprofessionalism in medical school faculties. *Academic Medicine, 90*, 442-446.

Binet, A., & Simon, T. (1916). *The development of intelligence in children (the Binet-Simon scale)*. Baltimore: Williams & Wilkins.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*, 317-335.

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology, 62*, 229-258.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*, 185-216.

Brown, D., & Ferrill, M. J. (2009). The taxonomy of professionalism: Reframing the academic pursuit of professional development. *American Journal of Pharmaceutical Education, 73*, 68.

Campion, M. C., Ployhart, R. E., & MacKenzie Jr, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance, 27*, 283-310.

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment, 20*, 333-346.

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

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

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

Chan, D., & Schmitt, N. (2005). Situational Judgment Tests. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *The Blackwell Handbook of Personnel Selection* Malden, MA: Blackwell Publishing.

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

Chander, A. (2016). The racist algorithm. *Michigan Law Review, 115*, 1023-1045.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*, 16-28.

Chandratilake, M., McAleer, S., & Gibson, J. (2012). Cultural similarities and differences in medical professionalism: A multi-region study. *Medical Education, 46*, 257-266.

Chang, A., Boscardin, C., Chou, C. L., Loeser, H., & Hauer, K. E. (2009). Predicting failing performance on a standardized patient clinical performance examination: The importance of communication and professionalism skills deficits. *Academic Medicine, 84*, S101-S104.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003, September 22-26). *SMOTEBoost: Improving prediction of the minority class in boosting.* Paper presented at the European Conference on Principles of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia.

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

Ciraco, M., Rogalewski, M., & Weiss, G. (2005, August 21). *Improving classifier utility by altering the misclassification cost ratio.* Paper presented at the First International Workshop on Utility-based Data Mining, Chicago, IL.

Cleland, J. A., Knight, L. V., Rees, C. E., Tracey, S., & Bond, C. M. (2008). Is it me or is it them? Factors that influence the passing of underperforming students. *Medical Education, 42*, 800-809.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale: Erlbaum Associates.

Colliver, J. A., Markwell, S. J., Verhulst, S. J., & Robbs, R. S. (2007). The prognostic value of documented unprofessional behavior in medical school records for predicting and preventing subsequent medical board disciplinary action: The Papadakis studies revisited. *Teaching and Learning in Medicine, 19*, 213-215.

Cook, M. (2016). *Personnel selection: Adding value through people - A changing picture*. Malden, MA: John Wiley & Sons Ltd.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI): Professional manual.* Odessa: Psychological Assessment Resources, Inc. .

Costa, P. T., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE Handbook of Personality Theory and Assessment* (pp. 179-198). London, England: SAGE Publications Ltd.

Cousans, F., Patterson, F., Edwards, H., Walker, K., McLachlan, J. C., & Good, D. (2017). Evaluating the complementary roles of an SJT and academic assessment for entry into clinical practice. *Advances in Health Sciences Education, 22*, 401-413.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*, 456-473.

Cruess, S. R. (2006). Professionalism and medicine's social contract with society. *Clinical Orthopaedics and Related Research, 449*, 170-176.

Cruess, S. R., Johnston, S., & Cruess, R. L. (2004). " Profession": a working definition for medical educators. *Teaching and Learning in Medicine, 16*, 74-76.

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.

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

De Fruyt, F., Mervielde, I., Hoekstra, H. A., & Rolland, J.-P. (2000). Assessing adolescents' personality with the NEO PI-R. *Assessment, 7*, 329-345.

De Leng, W. E., Stegers-Jager, K. M., Born, M. P., & Themmen, A. P. N. (2018). Integrity situational judgement test for medical school selection: Judging 'what to do' versus 'what not to do'. *Medical Education, 52*, 427-437.

De Leng, W. E., Stegers-Jager, K. M., Husbands, A., Dowell, J. S., Born, M. P., & Themmen, A. P. N. (2017). Scoring method of a situational judgment test: Influence on internal consistency reliability, adverse impact and correlation with personality? *Advances in Health Sciences Education, 22*, 243-265.

De Meijer, L. A. L., Born, M. P., Van Zielst, J., & Van der Molen, H. T. (2010). Construct-driven development of a video-based situational judgment test for integrity. *European Psychologist, 15*, 229-236.

De Soete, B., Lievens, F., Oostrom, J., & Westerveld, L. (2013). Alternative predictors for dealing with the diversity–validity dilemma in personnel selection: The constructed response multimedia test. *International Journal of Selection and Assessment, 21*, 239-250.

De Vries, R. E., & Born, M. P. (2013). De Vereenvoudigde HEXACO Persoonlijkheidsvragenlijst en een additioneel interstitieel Proactiviteitsfacet. *Gedrag & Organisatie, 26*, 223-245.

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009, July 1-3). *Predicting students drop out: A case study.* Paper presented at the International Working Group on Educational Data Mining, Cordoba, Spain.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*, 498-506.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*, 78-87.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*, 81-106.

Donovan, J. J., Dwight, S. A., & Schneider, D. (2014). The impact of applicant faking on selection measures, hiring decisions, and employee performance. *Journal of Business and Psychology, 29*, 479-493.

Dore, K. L., Reiter, H. I., Eva, K. W., Krueger, S., Scriven, E., Siu, E., . . . Norman, G. R. (2009). Extending the interview to all medical school candidates - Computer-Based Multiple Sample Evaluation of Noncognitive Skills (CMSENS). *Academic Medicine, 84*, S9-S12.

Dore, K. L., Reiter, H. I., Kreuger, S., & Norman, G. R. (2017). CASPer, an online pre-interview screen for personal/professional characteristics: Prediction of national licensure scores. *Advances in Health Sciences Education, 22*, 327-336.

Duffy, T. P. (2011). The Flexner report - 100 years later. *Yale Journal of Biology and Medicine, 84*, 269-276.

Dunlop, P. D., Morrison, D. L., & Cordery, J. L. (2011). Investigating retesting effects in a personnel selection context. *International Journal of Selection and Assessment, 19*, 217-221.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*, 1-23.

Dyrbye, L. N., Massie, F. S., Eacker, A., Harper, W., Power, D., Durning, S. J., . . . Sloan, J. (2010). Relationship between burnout and professional conduct and attitudes among US medical students. *JAMA, 304*, 1173-1180.

Edwards, J. R. (1993). Problems with the use of profile similarity indices in the study of congruence in organizational research. *Personnel Psychology, 46*, 641-665.

Edwards, J. R. (1994). Regression analysis as an alternative to difference scores. *Journal of Management, 20*, 683-689.

Edwards, J. R. (2001). Ten difference score myths. *Organizational research methods, 4*, 265-287.

Eibe, F., Hall, M. A., & Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann.

Ellingson, J. E., Sackett, P. R., & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology, 92*, 386-395.

Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 72*, 218-232.

Elliott, J. G., Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., & Hoffman, N. (2011). The socially skilled teacher and the development of tacit knowledge. *British Educational Research Journal, 37*, 83-103.

Eva, K. W. (2005). Dangerous personalities. *Advances in Health Sciences Education, 10*, 275-277.

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

Fargen, K. M., Drolet, B. C., & Philibert, I. (2016). Unprofessional behaviors among tomorrow's physicians: Review of the literature with a focus on risk factors, temporal trends, and future directions. *Academic Medicine, 91*, 858-864.

Ferguson, E., James, D., & Madeley, L. (2002). Factors associated with success in medical school: Systematic review of the literature. *BMJ, 324*, 952-957.

Ferguson, E., McManus, I. C., James, D., O'Hehir, F., & Sanders, A. (2003). Pilot study of the roles of personality, references, and personal statements in relation to performance over the five years of a medical degree. *BMJ, 326*, 429-432.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London, England: SAGE Publication Ltd.

Finn, G. M., Mwandigha, L., Paton, L. W., & Tiffin, P. A. (2018). The ability of 'non-cognitive' traits to predict undergraduate performance in medical schools: A national linkage study. *BMC Medical Education, 18*, 93.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Flaskerud, J. H. (1988). Is the Likert scale format culturally biased? *Nursing Research, 37*, 158-186.

Frank, J. R., Snell, L., & Sherbino, J. (2015). *CanMEDS 2015 Physician competency framework*. Ottawa, ON: The Royal College of Physicians and Surgeons of Canada.

Fröhlich, M., Kahmann, J., & Kadmon, M. (2017). Development and psychometric examination of a German video-based situational judgment test for social competencies in medical school applicants. *International Journal of Selection and Assessment, 25*, 94-110.

García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems, 25*, 13-21.

Gardner, A. K., & Dunkin, B. J. (2017). Evaluation of validity evidence for personality, emotional intelligence, and situational judgment tests to identify successful residents. *JAMA surgery, 53*, 409-416.

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

Goffin, R. D., & Boyd, A. C. (2009). Faking and personality assessment in personnel selection: Advancing models of faking. *Canadian Psychology, 50*, 151-160.

Goffin, R. D., & Christiansen, N. D. (2003). Correcting personality tests for faking: A review of popular personality tests and an initial survey of researchers. *International Journal of Selection and Assessment, 11*, 340-344.

Gold, B., & Holodynski, M. (2015). Development and construct validation of a situational judgment test of strategic knowledge of classroom management in elementary schools. *Educational Assessment, 20*, 226-248.

Goss, B. D., Ryan, A. T., Waring, J., Judd, T., Chiavaroli, N. G., O'Brien, R. C., . . . McColl, G. J. (2017). Beyond selection: The use of situational judgement tests in the teaching and assessment of professionalism. *Academic Medicine, 92*, 780-784.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*, 930-944.

Griffin, B., & Wilson, I. G. (2012). Faking good: Self-enhancement in medical school applicants. *Medical Education, 46*, 485-490.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36*, 341-355.

Griffith, R. L., & Peterson, M. H. (2008). The failure of social desirability measures to capture applicant faking behavior. *Industrial and Organizational Psychology, 1*, 308-311.

Gruys, M. L., & Sackett, P. R. (2003). Investigating the dimensionality of counterproductive work behavior. *International Journal of Selection and Assessment, 11*, 30-42.

Guenole, N., Chernyshenko, O. S., & Weekly, J. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing, 17*, 234-252.

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008, October 18-20). *On the class imbalance problem.* Paper presented at the Fourth International Conference on Natural Computation, Jinan, China.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Hakstian, A. R., Farrell, S., & Tweed, R. G. (2002). The assessment of counterproductive tendencies by means of the California Psychological Inventory. *International Journal of Selection and Assessment, 10*, 58-86.

Hamers, J. H. M., & Resing, W. C. M. (1992). Learning potential assessment: Introduction. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment. Theoretical, methodological and practical issues.* (pp. 23-42). Lisse, the Netherlands: Swets & Zeitlinger Publishers.

Harris, Siedor, L. E., Fan, Y., Listyg, B., & Carter, N. T. (2016). In defense of the situation: An interactionist axplanation for performance on situational judgment tests. *Industrial and Organizational Psychology, 9*, 23-28.

Harris, Walsh, & Lammy. (2015). UK medical selection: lottery or meritocracy? *Clinical Medicine, 15*, 40-46.

Hassink, R. (2018, March 19). Integriteitstest - wel of niet? [Integritytest - do or don't?]. Arts en Auto. Retrieved from https://www.artsenauto.nl/integriteitstest-wel-of-niet/.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The element of statistical learning*. New York, NY: Springer.

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

He, H., & Garcia, E. A. (2008). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*, 1263-1284.

He, J., Bartram, D., Inceoglu, I., & Van de Vijver, F. J. R. (2014). Response styles and personality traits: A multilevel analysis. *Journal of Cross-Cultural Psychology, 45*, 1028-1045.

Hedlund, J., Forsythe, G. B., Horvath, J. A., Williams, W. M., Snook, S., & Sternberg, R. J. (2003). Identifying and assessing tacit knowledge: Understanding the practical intelligence of military leaders. *The Leadership Quarterly, 14*, 117-140.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9-24.

Higgins, E. T., Roney, C. J. R., Crowe, E., & Hymes, C. (1994). Ideal versus ought predilections for approach and avoidance distinct self-regulatory systems. *Journal of Personality and Social Psychology, 66*, 276-286.

Higgins, E. T., Shah, J., & Friedman, R. (1997). Emotional responses to goal attainment: strength of regulatory focus as moderator. *Journal of Personality and Social Psychology, 72*, 515-525.

Hillis, D. J., & Grigg, M. J. (2015). Professionalism and the role of medical colleges. *The Surgeon, 13*, 292-299.

Hoekstra, H. A., Ormel, J., & De Fruyt, F. (1996). *Handleiding NEO persoonlijkheidsvragenlijsten [Manual NEO personality questionnaires]*. Lisse, the Netherlands: Swets Test Services.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Thousand Oaks: Sage Publications.

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 109-127). San Francisco: Jossey-Bass Publishers.

Hofstee, W. K. B. (1990). Allocation by lot: A conceptual and empirical analysis. *Social Science Information, 29*, 745-763.

Hofstee, W. K. B. (2005). De psycholoog als detective? Kanttekeningen bij malingering- en integriteitstests [The psychologist as a detective? Comments on malingering and integrity tests]. *De Psycholoog, 40*, 670-674.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*, 1270-1285.

Holtrop, D., Born, M. P., de Vries, A., & de Vries, R. E. (2014). A matter of context: A comparison of two types of contextualized personality measures. *Personality and Individual Differences, 68*, 234-240.

Holtrop, D., Breda, W. R. J., Oostrom, J. K., De Vries, R. E., & Stooker, J. (2015, November 27). *Predicting work performance with new technology: Automatic recognition of conscientiousness in spoken text versus self-rated conscientiousness?* Paper presented at the Werkgemeenschap van onderzoekers in Arbeids- & Organisatiepsychologie, Amsterdam, the Netherlands.

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

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*, 296-309.

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *The International Journal of Aviation Psychology, 13*, 373-386.

Husbands, A., Rodgerson, M. J., Dowell, J., & Patterson, F. (2015). Evaluating the validity of an integrity-based situational judgement test for medical school admissions. *BMC Medical Education, 15*, 144.

Ingold, P. V., Kleinmann, M., König, C. J., & Melchers, K. G. (2015). Shall we continue or stop disapproving of self-presentation? Evidence on impression management and faking in a selection context and their relation to job performance. *European Journal of Work and Organizational Psychology, 24*, 420-432.

Iobst, W. F., Sherbino, J., Cate, O. T., Richardson, D. L., Dath, D., Swing, S. R., . . . Frank, J. R. (2010). Competency-based medical education in postgraduate medical education. *Medical Teacher, 32*, 651-656.

Irby, D. M., Cooke, M., & O'Brien, B. C. (2010). Calls for reform of medical education by the Carnegie Foundation for the Advancement of Teaching: 1910 and 2010. *Academic Medicine, 85*, 220-227.

Jackson, D. J. R., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrooshan, A. (2017). The internal structure of situational judgement tests reflects candidate main effects: Not dimensions or situations. *Journal of Occupational and Organizational Psychology, 90*, 1-27.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*, 371-388.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*, 429-449.

Jerant, A., Griffin, E., Rainwater, J., Henderson, M., Sousa, F., Bertakis, K. D., . . . Franks, P. (2012). Does applicant personality influence multiple mini-interview performance and medical school acceptance offers? *Academic Medicine, 87*, 1250-1259.

Jetten, J., Postmes, T., & McAuliffe, B. J. (2002). 'We're all individuals': Group norms of individualism and collectivism, levels of identification and identity threat. *European Journal of Social Psychology, 32*, 189-207.

Jha, V., McLean, M., Gibbs, T. J., & Sandars, J. (2015). Medical professionalism across cultures: A challenge for medicine and medical education. *Medical Teacher, 37*, 74-80.

Kaminski, K., Felfe, J., Schäpers, P., & Krumm, S. (2019). A closer look at response options: Is judgment in situational judgment tests a function of the desirability of response options? *International Journal of Selection and Assessment, 27*, 72-82.

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

Kanter, M. H., Nguyen, M., Klau, M. H., Spiegel, N. H., & Ambrosini, V. L. (2013). What does professionalism mean to the physician? *The Permanente Journal, 17*, 87-90.

Khaliq, A. A., Dimassi, H., Huang, C.-Y., Narine, L., & Smego Jr, R. A. (2005). Disciplinary action against physicians: Who is likely to get disciplined? *American Journal of Medicine, 118*, 773-777.

Knights, J. A., & Kennedy, B. J. (2006). Medical school selection: Screening for dysfunctional tendencies. *Medical Education, 40*, 1058-1064.

Koczwara, A., Patterson, F., Zibarras, L., Kerrin, M., Irish, B., & Wilkinson, M. (2012). Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. *Medical Education, 46*, 399-408.

König, C. J., Merz, A. S., & Trauffer, N. (2012). What is in applicants' minds when they fill out a personality test? Insights from a qualitative study. *International Journal of Selection and Assessment, 20*, 442-452.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31*, 249-268.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review, 26*, 159-190.

Kulas, J. T. (2013). Personality-based profile matching in personnel selection: Estimates of method prevalence and criterion-related validity. *Applied Psychology, 62*, 519-542.

Kulasegaram, K., Reiter, H. I., Wiesner, W., Hackett, R. D., & Norman, G. R. (2010). Non-association between Neo-5 personality tests and multiple mini-interview. *Advances in Health Sciences Education, 15*, 415-423.

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.

Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology, 96*, 202-210.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate behavioral research, 39*, 329-358.

Lee, K., Ashton, M. C., & De Vries, R. E. (2005). Predicting workplace delinquency and integrity with the HEXACO and five-factor models of personality structure. *Human Performance, 18*, 179-197.

Leeds, J. P. (2012). The theory of cognitive acuity: Extending psychophysics to the measurement of situational judgment. *Journal of Neuroscience, Psychology, and Economics, 5*, 166-181.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a Likert-based testing procedure. *Intelligence, 21*, 247-266.

Legree, P. J., & Grafton, F. C. (1995). *Evidence for an interpersonal knowledge factor: The reliability and factor structure of tests of interpersonal knowledge and general cognitive ability*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Schiences Technical Report No. 1030.

Legree, P. J., Kilcullen, R., Psotka, J., Putka, D., & Ginter, R. N. (2010). *Scoring situational judgment tests using profile similarity metrics*. Alexandria, VA: U.S. Army Institute for the Behavioral and Social Sciences Technical Report No. 1272.

Legree, P. J., Ness, A. M., Kilcullen, R. N., & Koch, A. J. (2018). *Enhancing the Validity of Rating-Based Tests*. Fort Belvoir, VA: U.S. Army Institute for the Behavioral and Social Sciences Technical Report No. 1371.

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: An international handbook* (pp. 155-179). Ashland, OH: Hogrefe & Huber Publishers.

Levashina, J., Weekley, J. A., Roulin, N., & Hauck, E. (2014). Using blatant extreme responding for detecting faking in high-stakes selection: Construct validity, relationship with general mental ability, and subgroup differences. *International Journal of Selection and Assessment, 22*, 371-383.

Libbrecht, N., & Lievens, F. (2012). Validity evidence for the situational judgment test paradigm in emotional intelligence measurement. *International Journal of Psychology, 47*, 438-447.

Libbrecht, N., Lievens, F., Carette, B., & Côté, S. (2014). Emotional intelligence predicts success in medical school. *Emotion, 14*, 64-73.

Liem, C. C. S., Langer, M., Demetriou, A., Hiemstra, A. M. F., Wicaksana, A. S., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. Van Gerven (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 197-253). Cham, Switzerland: Springer.

Lievens, F. (2013). Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests. *Medical Education, 47*, 182-189.

Lievens, F. (2017). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing, 17*, 269-276.

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., Corstjens, J., Sorrel, M. A., Abad, F. J., Olea, J., & Ponsoda, V. (2015). The cross-cultural transportability of situational judgment tests: How does a US-based integrity situational judgment test fare in Spain? *International Journal of Selection and Assessment, 23*, 361-372.

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology, 9*, 3-22.

Lievens, F., Patterson, F., Corstjens, J., Martin, S., & Nicholson, S. (2016). Widening access in selection using situational judgement tests: Evidence from the UKCAT. *Medical Education, 50*, 624-636.

Lievens, F., & Peeters, H. (2008). Impact of elaboration on responding to situational judgment test items. *International Journal of Selection and Assessment, 16*, 345-355.

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

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

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

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.

Lievens, F., & Schollaert, E. (2008). *Naar een nieuwe generatie assessment: Een open boek over situationele tests*: Barneveld, Nederland: Uitgeverij Nelissen.

Lineberry, M., Kreiter, C. D., & Bordage, G. (2013). Threats to validity in the use and interpretation of script concordance test scores. *Medical Education, 47*, 1175-1183.

Lineberry, M., Kreiter, C. D., & Bordage, G. (2014). Script concordance tests: Strong inferences about examinees require stronger evidence. *Medical Education, 48*, 452-453.

Lucieer, S. M., Stegers-Jager, K. M., Rikers, R. M. J. P., & Themmen, A. P. N. (2016). Non-cognitive selected students do not outperform lottery-admitted students in the pre-clinical stage of medical school. *Advances in Health Sciences Education, 21*, 51-61.

Luschin-Ebengreuth, M., Dimai, H. P., Ithaler, D., Neges, H. M., & Reibnegger, G. (2015). Situational judgment test as an additional tool in a medical admission test: An observational investigation. *BMC Research Notes, 8*, 81.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education, 53*, 950-965.

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

Mak-Van der Vossen, M. C., Peerdeman, S., Van Mook, W., Croiset, G., & Kusurkar, R. (2014). Assessing professional behaviour: Overcoming teachers' reluctance to fail students. *BMC Research Notes, 7*, 368.

Mak-Van der Vossen, M. C., Van Mook, W. N. K. A., Kors, J. M., Van Wieringen, W. N., Peerdeman, S. M., Croiset, G., & Kusurkar, R. A. (2016). Distinguishing three unprofessional behavior profiles of medical students using latent class analysis. *Academic Medicine, 91*, 1276-1283.

Marcus, B., Lee, K., & Ashton, M. C. (2007). Personality dimensions explaining relationships between integrity tests and counterproductive behavior: Big Five, or one in addition? *Personnel Psychology, 60*, 1-34.

Marqués, A. I., García, V., & Sánchez, J. S. (2013). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society, 64*, 1060-1070.

Mavis, B. (2001). Self-efficacy and OSCE performance among second year medical students. *Advances in Health Sciences Education, 6*, 93-102.

McCarthy, K., Zabar, B., & Weiss, G. (2005, August 21). *Does cost-sensitive learning beat sampling for classifying rare classes?* Paper presented at the First International Workshop on Utility-based Data Mining, Chicago, IL.

McCluskey, A., & Lalkhen, A. G. (2007). Statistics II: Central tendency and spread of data. *Continuing Education in Anaesthesia, Critical Care & Pain, 7*, 127-130.

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

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

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

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology, 96*, 327-336.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812-821.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*, 17-24.

Mercaldo, N. D., Lau, K. F., & Zhou, X. H. (2007). Confidence intervals for predictive values with an emphasis to case-control studies. *Statistics in Medicine, 26*, 2170-2183.

Mitchell, T. M. (1997). *Machine learning*. New York, NY: McGraw-Hill Companies, Inc.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683-729.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105-125.

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

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology, 24*, 281-288.

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

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

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*, 348-355.

Mueller-Hanson, R. A., Heggestad, E. D., & Thornton, G. C. (2006). Individual differences in impression management: An exploration of the psychological processes underlying faking. *Psychology Science, 48*, 288-312.

Mussel, P., Gatzka, T., & Hewig, J. (2018). Situational judgment tests as an alternative measure for personality assessment. *European Journal of Psychological Assessment, 34*, 328-335.

Nas, C. N., Brugman, D., & Koops, W. (2008). Measuring self-serving cognitive distortions with the "How I Think" Questionnaire. *European Journal of Psychological Assessment, 24*, 181-189.

Neal, G. E. H., Oram, R. C., & Bacon, A. J. (2018). What do students think about the situational judgment test? *Medical Teacher, 40*, 212-213.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13*, 250-260.

Niessen, A. S. M., & Meijer, R. R. (2016). Selection of medical students on the basis of non-academic skills: Is it worth the trouble? *Clinical Medicine, 16*, 339-342.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017a). Applying organizational justice theory to admission into higher education: Admission from a student perspective. *International Journal of Selection and Assessment, 25*, 72-84.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017b). Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. *Personality and Individual Differences, 106*, 183-189.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24,* 1565-1567.

Norman, G. (2015). Identifying the bad apples. *Advances in Health Sciences Education, 20*, 299-303.

O'Connell, M. S., Hartman, N. S., McDaniel, M. A., Grubb, W. L., & Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment, 15*, 19-29.

O'Neill, T. A., Lewis, R. J., Law, S. J., Larson, N., Hancock, S., Radan, J., . . . Carswell, J. J. (2017). Forced-choice pre-employment personality assessment: Construct validity and resistance to faking. *Personality and Individual Differences, 115*, 120-127.

Olaru, G., Burrus, J., MacCann, C., Zaromb, F. M., Wilhelm, O., & Roberts, R. D. (2019). Situational Judgment Tests as a method for measuring personality: Development and validity evidence for a test of Dependability. *PloS One, 14*, e0211884.

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

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*, 245-269.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660-679.

Ones, D. S., Viswesvaran, C., Schmidt, F. L., & Reiss, A. D. (1994, August 14-17). *The validity of honesty and violence scales of integrity tests in predicting violence at work.* Paper presented at the Annual Meeting of the Academy of Management, Dallas, TX.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2012). Implicit trait policies in multimedia situational judgment tests for leadership skills: Can they predict leadership behavior? *Human Performance, 25*, 335-353.

Oostrom, J. K., De Vries, R. E., & De Wit, M. (2019). Development and validation of a HEXACO situational judgment test. *Human Performance, 32*, 1-29.

Oostrom, J. K., Köbis, N. C., Ronay, R., & Cremers, M. (2017). False consensus in situational judgment tests: What would others do? *Journal of Research in Personality, 71*, 33-45.

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.

Papadakis, M. A., Hodgson, C. S., Teherani, A., & Kohatsu, N. D. (2004). Unprofessional behavior in medical school is associated with subsequent disciplinary action by a state medical board. *Academic Medicine, 79*, 244-249.

Papadakis, M. A., Teherani, A., Banach, M. A., Knettler, T. R., Rattner, S. L., Stern, D. T., . . . Hodgson, C. S. (2005). Disciplinary action by medical boards and prior behavior in medical school. *New England Journal of Medicine, 353*, 2673-2682.

Paton, L. W., Tiffin, P. A., Smith, D., Dowell, J. S., & Mwandigha, L. M. (2018). Predictors of fitness to practise declarations in UK medical undergraduates. *BMC Medical Education, 18*, 68.

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical Education, 46*, 850-868.

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.

Patterson, F., Cousans, F., Edwards, H., Rosselli, A., Nicholson, S., & Wright, B. (2017). The predictive validity of a text-based situational judgment test in undergraduate medical and dental school admissions. *Academic Medicine, 92*, 1250-1253.

Patterson, F., Ferguson, E., & Thomas, S. (2008). Using job analysis to identify core and specific competencies: Implications for selection and recruitment. *Medical Education, 42*, 1195-1204.

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education, 50*, 36-60.

Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Medical Teacher, 38*, 3-17.

Patterson, F., Zibarras, L., Carr, V., Irish, B., & Gregory, S. (2011). Evaluating candidate reactions to selection practices using organisational justice theory. *Medical Education, 45*, 289-297.

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality, 66*, 1025-1060.

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.

Pelt, D. H. M., Van der Linden, D., & Born, M. P. (2017). How emotional intelligence might get you the job: The relationship between trait emotional intelligence and faking on personality tests. *Human Performance, 31*, 33-54.

Peter, J. P., Churchill Jr, G. A., & Brown, T. J. (1993). Caution in the use of difference scores in consumer research. *Journal of Consumer Research, 19*, 655-662.

Peus, C., Braun, S., & Frey, D. (2013). Situation-based measurement of the full range of leadership model - Development and validation of a situational judgment test. *The Leadership Quarterly, 24*, 777-795.

Phyu, T. N. (2009, March 18-20). *Survey of classification techniques in data mining.* Paper presented at the International MultiConference of Engineers and Computer Scientists, Hong Kong.

Pintrich, P. R. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning.

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., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153-172.

Ployhart, R. E., & MacKenzie Jr, W. I. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Ed.), *APA Handbooks in Psychology. APA handbook of industrial and organizational psychology, Vol. 2. Selecting and developing members for the organization* (pp. 237-252). Washington, DC: American Psychological Association.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology, 56*, 733-752.

Ployhart, R. E., Ziegert, J. C., & McFarland, L. A. (2003). Understanding racial differences on cognitive ability tests in selection contexts: An integration of stereotype threat and applicant reactions research. *Human Performance, 16*, 231-259.

Powis, D. (2015). Selecting medical students: An unresolved challenge. *Medical Teacher, 37*, 252-260.

Prasad, V. (2011). Are we treating professionalism professionally? Medical school behavior as predictors of future outcomes. *Teaching and Learning in Medicine, 23*, 337-341.

Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: A review. *International Journal of Computing and Business Research, 5*, 1-29.

Remmerswaal, W. (2016, December 5). Verdeelde reacties op 'horkentest' Erasmus [Divided responses to 'jerk test' Erasmus]. *Algemeen Dagblad*. Retrieved from

https://www.ad.nl/binnenland/verdeelde-reacties-op-horkentest-erasmus~a24f1c65/.

Rennie, S. C., & Crosby, J. R. (2001). Are “tomorrow's doctors” honest? Questionnaire study exploring medical students' attitudes and reported behaviour on academic misconduct. *BMJ, 322*, 274-275.

Robie, C., Risavy, S. D., Holtrop, D., & Born, M. P. (2017). Fully contextualized, frequency-based personality measurement: A replication and extension. *Journal of Research in Personality, 70*, 56-65.

Robinson, S. L., & Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal, 38*, 555-572.

Robson, S. M., Jones, A., & Abraham, J. (2007). Personality, faking, and convergent validity: A warning concerning warning statements. *Human Performance, 21*, 89-106.

Rockstuhl, T., Ang, S., Ng, K.-Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgment tests: Evidence from intercultural multimedia SJTs. *Journal of Applied Psychology, 100*, 464-480.

Rosenfeld, B., Sands, S. A., & Gorp, W. G. V. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology, 15*, 349-359.

Rosman, T., Mayer, A.-K., & Krampen, G. (2015). Measuring psychology students’ information-seeking skills in a situational judgment test format: Construction and validation of the PIKE-P test. *European Journal of Psychological Assessment, 32*, 220-229.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634-644.

Roulin, N., Krings, F., & Binggeli, S. (2016). A dynamic model of applicant faking. *Organizational Psychology Review, 6*, 145-170.

Rowe, M. (2019). An introduction to machine learning for clinicians. *Academic Medicine*. doi: 10.1097/ACM.0000000000002792

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

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

Ryan, A. M., Sacco, J. M., McFarland, L. A., & Kriska, S. D. (2000). Applicant self-selection: Correlates of withdrawal from a multiple hurdle process. *Journal of Applied Psychology, 85*, 163-179.

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.

Schleicher, D. J., Venkataramani, V., Morgeson, F. P., & Campion, M. A. (2006). So you didn't get the job… now what do you think? Examining opportunity-to-perform fairness perceptions. *Personnel Psychology, 59*, 559-590.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*, 162-173.

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

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 135-155). New York, NY: Lawrence Erlbaum Associates.

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 noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology, 94*, 1479-1497.

Schmitt, N., Prasad, J. J., Ryan, A. M., Bradburn, J. C., & Nye, C. D. (2019). Culture as a determinant of option choice in a situational judgement test: A new look. *Journal of Occupational and Organizational Psychology, 92*, 330-351.

Schreurs, B., Derous, E., Proost, K., Notelaers, G., & Witte, K. D. (2008). Applicant selection expectations: Validating a multidimensional measure in the military. *International Journal of Selection and Assessment, 16*, 170-176.

Schreurs, S., Cleland, J., Muijtjens, A. M. M., Oude Egbrink, M. G. A., & Cleutjens, K. (2018). Does selection pay off? A cost–benefit comparison of medical school selection and lottery systems. *Medical Education, 52*, 1240-1248.

Schripsema, N. R., Van Trigt, A. M., Borleffs, J. C. C., & Cohen-Schotanus, J. (2017). Impact of vocational interests, previous academic experience, gender and age on situational judgement test performance. *Advances in Health Sciences Education, 22*, 521-532.

Shrank, W. H., Reed, V. A., & Jernstedt, C. (2004). Fostering professionalism in medical education. *Journal of General Internal Medicine, 19*, 887-892.

Smith, D. T., & Tiffin, P. A. (2018). Evaluating the validity of the selection measures used for the UK's foundation medical training programme: A national cohort study. *BMJ open, 8*, e021918.

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

Statistics Netherlands. Wat verstaat het CBS onder een allochtoon? [How does Statistic Netherlands define a person from an ethnic minority background?]. Retrieved from https://www.cbs.nl/nl-nl/faq/specifiek/wat-verstaat-het-cbs-onder-een-allochtoon-.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797-811.

Stegers-Jager, K. M. (2018). Lessons learned from 15 years of non-grades-based selection for medical school. *Medical Education, 52*, 86-95.

Stegers-Jager, K. M., Cohen-Schotanus, J., & Themmen, A. P. N. (2012). Motivation, learning strategies, participation and medical school performance. *Medical Education, 46*, 678-688.

Stegers-Jager, K. M., Steyerberg, E. W., Cohen-Schotanus, J., & Themmen, A. P. N. (2012). Ethnic disparities in undergraduate pre-clinical and clinical performance. *Medical Education, 46*, 575-585.

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

Steiner, D. D., & Gilliland, S. W. (1996). Fairness reactions to personnel selection techniques in France and the United States. *Journal of Applied Psychology, 81*, 134-141.

Stemler, S. E., Aggarwal, V., & Nithyanand, S. (2016). Knowing what NOT to do is a critical job skill: Evidence from 10 different scoring methods. *International Journal of Selection and Assessment, 24*, 229-245.

Stern, D. T., Frohna, A. Z., & Gruppen, L. D. (2005). The prediction of professional behaviour. *Medical Education, 39*, 75-82.

Sternberg, R. J., & Collaborators, T. R. P. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence, 34*, 321-350.

Sternberg, R. J., Wagner, R. K., & Okagaki, L. (1993). Practical intelligence: The nature and role of tacit knowledge in work and at school. In J. M. Puckett & H. W. Reese (Eds.), *Mechanisms of everyday cognition* (pp. 205-227). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99-103.

Stumpf, H., & Stanley, J. C. (1996). Gender-related differences on the College Board's Advanced Placement and Achievement tests, 1982–1992. *Journal of Educational Psychology, 88*, 353-364.

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence, 23*, 687-719.

Swick, H. M. (2000). Toward a normative definition of medical professionalism. *Academic Medicine, 75*, 612-616.

Tiffin, P. A., & Carter, M. (2015, May). Understanding the measurement model of the UKCAT situational judgement test: Summary report. *UCAT Consortium.* Retrieved from https://www.ucat.ac.uk/media/1183/understanding-the-measurement-model-of-the-ukcat-sjt.pdf.

Truijens, A. (2016, December 10). 'Horktest' voor artsen in de dop is griezelig en heilloos ['Jerktest' for future medical doctors is creepy and disastrous]. *Volkskrant*. Retrieved from https://www.volkskrant.nl/nieuws-achtergrond/aleid-truijens-horktest-voor-artsen-in-de-dop-is-griezelig-en-heilloos~b478a233/.

Uggerslev, K. L., Fassina, N. E., & Kraichy, D. (2012). Recruiting through the stages: A meta-analytic test of predictors of applicant attraction at different stages of the recruiting process. *Personnel Psychology, 65*, 597-660.

Urlings-Strop, L. C., Stijnen, T., Themmen, A. P. N., & Splinter, T. A. W. (2009). Selection of medical students: A controlled experiment. *Medical Education, 43*, 175-183.

Urlings-Strop, L. C., Themmen, A. P. N., Stijnen, T., & Splinter, T. A. W. (2011). Selected medical students achieve better than lottery-admitted students during clerkships. *Medical Education, 45*, 1032-1040.

Van Hooft, E. A. J., & Born, M. P. (2012). Intentional response distortion on personality tests: using eye-tracking to understand response processes when faking. *Journal of Applied Psychology, 97*, 301-316.

Van Iddekinge, C. H., & Arnold, J. D. (2017). Retaking employment tests: What we know and what we still need to know. *Annual Review of Organizational Psychology and Organizational Behavior, 4*, 445-471.

Van Mook, W. N. K. A., Gorter, S. L., De Grave, W. S., Van Luijk, S. J., Wass, V., Zwaveling, J. H., . . . Van Der Vleuten, C. P. M. (2010). Bad apples spoil the barrel: Addressing unprofessional behaviour. *Medical Teacher, 32*, 891-898.

Van Mook, W. N. K. A., Gorter, S. L., O'Sullivan, H., Wass, V., Schuwirth, L. W., & Van der Vleuten, C. P. M. (2009). Approaches to professional behaviour assessment: Tools in the professionalism toolbox. *European Journal of Internal Medicine, 20*, e153-e157.

Vasilopoulos, N. L., Reilly, R. R., & Leaman, J. A. (2000). The influence of job familiarity and impression management on self-report measure scale scores and response latencies. *Journal of Applied Psychology, 85*, 50-64.

Venema, J. (2016, December 3). Erasmus MC ontwikkelt 'horkentest' om botte arts te weren [Erasmus MC develops 'jerk test' to keep out brusque medical doctor]. *Algemeen Dagblad*. Retrieved from https://www.ad.nl/rotterdam/erasmus-mc-ontwikkelt-horkentest-om-botte-arts-te-weren~a683ce14/.

Vergauwe, J., Wille, B., Hofmans, J., Kaiser, R. B., & Fruyt, F. D. (2017). The too little/too much scale: A new rating format for detecting curvilinear effects. *Organizational Research Methods, 20*, 518-544.

Visser, K. (2017, August 23). Kantel de selectie in het hoger onderwijs [Topple selective admission in higher education]. *ScienceGuide*. Retrieved from https://www.scienceguide.nl/2017/08/kantel-de-selectie-in-het-hoger-onderwijs/.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197-210.

Walsh, S., Arnold, B., Pickwell-Smith, B., & Summers, B. (2016). What kind of doctor would you like me to be? *Clinical Teacher, 13*, 98-101.

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

Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement, and subgroup differences. *Human Performance, 17*, 433-461.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement and application* (pp. 157-182). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Weiss, G. M., & Hirsh, H. (2000). *Learning to predict extremely rare events*. AAAI Technical Report WS-00-05.

Weiss, G. M., & Provost, F. (2001). *The effect of class distribution on classifier learning: An empirical study*. Technical Report Rutgers University.

Weng, Q. D., Yang, H., Lievens, F., & McDaniel, M. A. (2018). Optimizing the validity of situational judgment tests: The importance of scoring methods. *Journal of Vocational Behavior, 104*, 199-209.

Westring, A. J. F., Oswald, F. L., Schmitt, N., Drzakowski, S., Imus, A., Kim, B., & Shivpuri, S. (2009). Estimating trait and situational variance in a situational judgment test. *Human Performance, 22*, 44-63.

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

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.

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15*, 270-292.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann.

Wouters, A., Croiset, G., & Kusurkar, R. A. (2018). Selection and lottery in medical school admissions: who gains and who loses? *MedEdPublish*. doi: https://doi.org/10.15694/mep.2018.0000271.1

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*, 377-392.

Yates, J. (2011). Development of a 'toolkit' to identify medical students at risk of failure to thrive on the course: An exploratory retrospective case study. *BMC Medical Education, 11*, 95.

Yates, J., & James, D. (2006). Predicting the “strugglers”: A case-control study of students at Nottingham University Medical School. *BMJ, 332*, 1009-1013.

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research, 5*, 1205-1224.

Zhang, H. (2004, May 12-14). *The optimality of naive Bayes.* Paper presented at the Seventeenth International Artificial Intelligence Research Society Conference, Miami Beach, FL.

Ziegler, M. (2015). “F*** You, I Won’t Do What You Told Me!” - Response biases as threats to psychological assessment. *European Journal of Psychological Assessment, 31*, 153-158.

Zwick, R. (2017). *Who gets in? Strategies for fair and effective college admissions*. Cambridge, MA: Harvard University Press.

Appendices

Appendix 2A

Example Scenario

Michael questions Sarah, a fellow medical student about extreme and provocative comments about individuals' sexual preferences on her Facebook page. Sarah argues she should be free to express her personal views. She also insists that her personal views have no bearing on her performance as a medical student or patient care.

How appropriate are each of the following responses by Michael in this situation?

1. Advise Sarah to remove all controversial comments from her Facebook page
2. Alert Facebook that Sarah's page contains potentially inappropriate content as they could remove it
3. Ask Sarah to ensure her privacy settings are restricted so her page is inaccessible to patients or the general public
4. Inform a member of staff about Sarah's Facebook comments
5. Withhold advice to Sarah as her views do not affect patient care or performance as a medical student

Appendix 2B

Correlation between the SJT scores resulting from the 28 different scoring methods. All correlations are significant. The numbers in the table correspond to the scoring methods in Table 2 to 4 of Chapter 2.

Method	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
1.														
2.	.95													
3.	.94	.99												
4.	.98	.93	.92											
5.	.95	.96	.95	.97										
6.	.95	.95	.96	.97	.99									
7.	.91	.83	.84	.89	.85	.86								
8.	.86	.83	.84	.85	.84	.84	.93							
9.	.86	.85	.86	.85	.84	.85	.92	.98						
10.	.91	.84	.85	.92	.88	.89	.98	.92	.91					
11.	.87	.84	.84	.89	.87	.88	.92	.96	.95	.95				
12.	.87	.85	.86	.89	.88	.89	.91	.95	.96	.95	.99			
13.	.34	.35	.37	.34	.35	.36	.26	.30	.32	.28	.31	.32		
14.	.35	.38	.39	.35	.37	.38	.26	.28	.31	.27	.30	.32	.98	
15.	.34	.35	.37	.35	.37	.38	.25	.29	.31	.28	.32	.33	.96	.94
16.	.34	.37	.38	.36	.38	.39	.25	.29	.31	.28	.31	.33	.95	.95
17.	.31	.32	.34	.31	.33	.34	.29	.33	.34	.31	.34	.35	.88	.84
18.	.31	.32	.34	.31	.32	.34	.29	.34	.36	.30	.35	.36	.87	.82
19.	.30	.31	.34	.31	.33	.35	.27	.30	.32	.30	.34	.35	.87	.84
20.	.31	.32	.34	.32	.33	.35	.27	.32	.34	.30	.35	.36	.87	.83
21.	-.58	-.62	-.61	-.57	-.58	-.58	-.31	-.37	-.39	-.34	-.39	-.41	-.36	-.38
22.	-.59	-.63	-.61	-.57	-.58	-.58	-.31	-.37	-.39	-.34	-.39	-.41	-.36	-.38
23.	-.51	-.52	-.56	-.51	-.51	-.54	-.49	-.63	-.63	-.49	-.61	-.61	-.39	-.36
24.	-.52	-.53	-.56	-.51	-.51	-.54	-.50	-.63	-.63	-.50	-.61	-.61	-.39	-.36
25.	-.88	-.91	-.93	-.84	-.84	-.85	-.75	-.77	-.80	-.74	-.76	-.77	-.37	-.38
26.	-.73	-.74	-.77	-.72	-.72	-.75	-.73	-.80	-.80	-.75	-.79	-.79	-.37	-.36
27.	-.90	-.91	-.92	-.86	-.86	-.87	-.78	-.80	-.82	-.77	-.79	-.80	-.40	-.41
28.	-.82	-.81	-.83	-.82	-.81	-.83	-.87	-.91	-.91	-.87	-.90	-.89	-.35	-.34

Appendix 2B continued

Method	15.	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.
1.													
2.													
3.													
4.													
5.													
6.													
7.													
8.													
9.													
10.													
11.													
12.													
13.													
14.													
15.													
16.	.99												
17.	.87	.85											
18.	.85	.83	.97										
19.	.91	.89	.96	.93									
20.	.91	.88	.95	.95	.99								
21.	-.37	-.38	-.29	-.29	-.32	-.32							
22.	-.37	-.38	-.30	-.30	-.32	-.32	1						
23.	-.38	-.37	-.42	-.44	-.40	-.42	.61	.60					
24.	-.39	-.37	-.43	-.44	-.41	-.42	.60	.60	1				
25.	-.36	-.36	-.33	-.33	-.33	-.34	.67	.67	.59	.58			
26.	-.37	-.37	-.41	-.41	-.41	-.41	.56	.55	.69	.68	.77		
27.	-.39	-.39	-.36	-.37	-.36	-.37	.68	.68	.64	.63	.97	.81	
28.	-.35	-.35	-.39	-.39	-.38	-.38	.49	.47	.69	.68	.82	.94	.85

Appendix 3A

Five example scenarios used in the integrity situational judgement test.

Scenario 1								
Michelle, Tim and Sarah have been given a group assignment. The group meets for the first time to discuss how to approach the assignment. During this meeting, it becomes clear that Sarah has already completed the assignment. Sarah explains that she had some spare time in which she completed the group assignment, but that she doesn't mind to share her work with Michelle and Tim. Tim says he is satisfied with this deal. *Judge for each of the following response options how appropriate they would be for Michelle.*								
		Very Inappropriate						Very appropriate
1.	Accept Sarah's offer because that is the most easy thing to do in this case.	1	2	3	4	5	6	
2.	Accept Sarah's offer because it would be a waste to carry out the same assignment twice.	1	2	3	4	5	6	
3.	Thank Sarah for the offer but propose to split up the assignment among themselves.	1	2	3	4	5	6	
4.	Together with Tim, complement Sarah's work to ensure your contribution to the group assignment.	1	2	3	4	5	6	

Scenario 2								
Steven and Robin are close friends. They are taking an exam. When Steven looks up he notices that Robin is using a smartphone which he has hidden in the sleeve of his sweater. *Judge for each of the following response options how appropriate they would be for Steven.*								
		Very Inappropriate						Very appropriate
1.	Do nothing since it is the proctors' responsibility to monitor the exam.	1	2	3	4	5	6	
2.	Meet Robin after the exam and tell him that you saw him using his smartphone.	1	2	3	4	5	6	
3.	Inform the proctors that Robin is violating the rules.	1	2	3	4	5	6	
4.	Do nothing because we all cheat sometimes.	1	2	3	4	5	6	

Scenario 3							
Stacey overhears two fellow students talk about yesterday's clinical skills course. During this course, students practice physical examinations on each other. The two students gossip about Maarten, a fellow student who has a neaves (a birth mark) on a large part of his body. *Judge for each of the following response options how appropriate they would be for Stacey.*							
		Very Inappropriate					Very appropriate
1.	Act as if she does not hear her fellow students because she doesn't care.	1	2	3	4	5	6
2.	Tell her fellow students that she can hear them gossiping about Maarten.	1	2	3	4	5	6
3.	Tell her fellow students that they should stop gossiping because they are violating the confidentiality rules.	1	2	3	4	5	6
4.	Say nothing because her fellow students would view her as a bore if she did.	1	2	3	4	5	6

Scenario 4							
Eva finds out that she has to hand in an essay tomorrow. Eva forgot to note the deadline in her calendar and has not yet started with the assignment. The essay is an important part as this has to be graded as 'sufficient' to pass the course. *Judge for each of the following response options how appropriate they would be for Eva.*							
		Very Inappropriate					Very appropriate
1.	Email her teacher to tell honestly that she forgot the deadline and ask for a postponement.	1	2	3	4	5	6
2.	Resit the essay, because she should have written down the assignment in her calendar.	1	2	3	4	5	6
3.	Make sure to have something on paper and hope for the best.	1	2	3	4	5	6
4.	Ask for a postponement because the teacher had not posted a reminder of the deadline.	1	2	3	4	5	6

Scenario 5							
Farid has started medical school. He joined a student association because he is new in town. The student association organizes a welcome week for new members. This welcome week, however, overlaps with a few mandatory lectures. *Judge for each of the following response options how appropriate they would be for Farid.*							
		Very Inappropriate					Very appropriate
1.	Explain the situation to the teacher and ask if he can make up for the lectures at a different moment.	1	2	3	4	5	6
2.	Skip the mandatory lectures and go to the welcome week.	1	2	3	4	5	6
3.	Call in sick for the lectures since other students will probably do the same.	1	2	3	4	5	6
4.	Try to find a solution together with the student association and the medical school.	1	2	3	4	5	6

Appendix 3B

Intraclass correlation coefficients (ICCs) including 95% confidence interval for the GP residents on all SJT items and the HH-based and CD-based SJT items.

Version	Total	HH-based	CD-based
I	.73 [.66; .80]	.37 [.26; .51]	.44 [.33; .58]
II	.65 [.57; .73]	.28 [.19; .41]	.40 [.29; .54]
III	.73 [.66; .79]	.19 [.12; .30]	.35 [.25; .49]

Note. HH = honesty-humility CD = cognitive distortions 95% confidence interval between square brackets

Appendix 3C

Number of items, rating scale and previously established internal consistency reliability of the measures used to examine the construct validity.

Measure	# items	rating scale	α
HEXACO-SPI honesty-humility	16	1: *strongly disagree* - 5: *strongly agree*	.78
HIT questionnaire	54	1: *disagree strongly* - 6: *agree strongly*	.78- .90
ICB student-related items	25	1: *never even considered it* - 6: *did it three or more times*	.88
Workplace deviance measure	17	1: *never* - 7: *daily*	.78-.81
MSLQ Self-efficacy subscale	8	1: *not at all true of me* - 7: *very true of me*	.93

Note. HEXACO-SPI = HEXACO Simplified Personality Inventory HIT = How I Think ICB = Inventory Counterproductive Behaviour MSLQ = Motivated Strategies of Learning Questionnaire

Appendix 3D

Intercorrelations between the SJT response option categories, honesty-humility facets and cognitive distortion categories.

		Integrity SJT						
		Sc	Bo	Mi	As	Si	Fa	Mo
Integrity SJT	Sc							
	Bo	**.63**						
	Mi	**.67**	**.67**					
	As	**.59**	**.56**	**.62**				
	Si	**.67**	**.62**	**.58**	**.57**			
	Fa	**.68**	**.67**	**.70**	**.55**	**.68**		
	Mo	**.60**	**.57**	**.57**	**.52**	**.67**	**.68**	
Integrity-rel. measures	Sc	**-.34**	**-.35**	**-.36**	**-.37**	**-.24**	**-.29**	**-.23**
	Bo	**-.24**	**-.26**	**-.21**	-.25	**-.16**	**-.17**	-.09
	Mi	**-.31**	**-.34**	-.31	-.34	**-.23**	**-.30**	**-.24**
	As	**-.29**	**-.30**	-.21	-.27	**-.18**	-.18	-.18
	Si	**.31**	**.34**	**.25**	**.30**	**.23**	**.25**	**.23**
	Fa	**.26**	**.31**	**.28**	**.31**	**.18**	.24	**.16**
	Mo	**.17**	**.22**	**.22**	**.12**	**.18**	**.17**	**.15**

Appendix 3D continued

		Integrity-related measures						
		Sc	Bo	Mi	As	Si	Fa	Mo
Integrity SJT	Sc							
	Bo							
	Mi							
	As							
	Si							
	Fa							
	Mo							
Integrity-rel. measures	Sc	.74/.73						
	Bo	**.68**	.75/.68					
	Mi	**.72**	**.71**	.67/.66				
	As	**.65**	**.69**	**.73**	.65/.65			
	Si	**-.41**	**-.30**	-.37	-.25	.63/.63		
	Fa	**-.59**	**-.51**	**-.54**	**-.47**	**.51**	.58/.56	
	Mo	**-.41**	**-.27**	**-.38**	**-.33**	**.32**	**.36**	.48/.45

Note. Sc = Self-centeredness Bo = Blaming others Mi = Minimizing As = Assuming the Worst Si = Sincerity Fa = Fairness Mo = Modesty Correlation coefficients are meta-analytically merged across version A and B Bold coefficients represent significant correlations ($p < .01$, two-tailed) Internal consistency reliabilities are on the diagonal line (version A/version B)

Appendix 4A

Example scenario 1							
John finds out that Mary has a copy of the exam paper that will be given next week. She tells him that she has already sold the exam to some fellow students and asks him if he also wants to look at the exam paper. *Judge for each of the following response options how appropriate they would be for John.*							
		Very inappropriate					Very appropriate
1.	Look at the exam paper because everyone would do that.	1	2	3	4	5	6
2.	Don't look at the exam since he is not entitled to do so.	1	2	3	4	5	6
3.	Look at the exam and tell no one you did.	1	2	3	4	5	6
4.	Don't look at the exam and inform the teacher.	1	2	3	4	5	6

Example scenario 2							
Michelle, Tim and Sarah have been given a group assignment. The group meets for the first time to discuss how to approach the assignment. During this meeting, it becomes clear that Sarah has already completed the assignment. Sarah explains that she had some spare time in which she completed the group assignment, but that she doesn't mind to share her work with Michelle and Tim. Tim says he is satisfied with this deal. *Judge for each of the following response options how appropriate they would be for Michelle.*							
		Very inappropriate					Very appropriate
1.	Accept Sarah's offer because that is the most easy thing to do in this case.	1	2	3	4	5	6
2.	Accept Sarah's offer because it would be a waste to carry out the same assignment twice.	1	2	3	4	5	6
3.	Thank Sarah for the offer but propose to split up the assignment among themselves.	1	2	3	4	5	6
4.	Together with Tim, complement Sarah's work to ensure your contribution to the group assignment.	1	2	3	4	5	6

Appendix 4B

Means, standard deviation, range and intercorrelations between scores obtained during the selection procedure and the SJT scores on both time points and for both scoring methods for the group of applicants that provided data on both test administrations (N = 317).

	M	SD	Range	4	5	6
1. Gender	0.26	0.44	0 - 1[1]			
2. Dutch ethnicity	0.60	0.49	0 - 1			
3. 1st generation HE	0.28	0.45	0 - 1			
4. Age (on T2)	18.75	2.46	17.00 - 44.68			
5. pu-GPA	6.70	0.43	5.67 - 7.58[2]	-.10		
6. ECA score	54.20	20.04	5.00 - 150.00	**-.30**	.01	
Cognitive tests						
7. Mathematics	33.53	6.82	16.00 - 48.00	**-.21**	**.30**	.04
8. Logical reasoning	54.93	7.50	34.00 - 76.00	-.03	.14	.09
9. Lecture	49.77	9.67	22.00 - 72.00	**-.26**	**.28**	**.16**
Raw consensus						
10. T1 SJT score	121.94	7.14	89.04 - 137.40	-.10	.08	-.02
11. T2 SJT score	120.22	7.32	91.68 - 140.00	-.04	.07	-.10
12. T1-T2 difference	1.72	7.99	-32.26 - 23.62	-.05	.01	.07
Stan. consensus						
13. T1 SJT score	84.31	4.42	61.66 - 90.51	.03	.04	-.06
14. T2 SJT score	85.33	4.11	59.31 - 90.27	0	.09	-.09
15. T1-T2 difference	-1.02	4.09	-25.71 - 15.42	.03	-.04	.03
Dich. consensus						
16. T1 SJT score	36.38	2.98	22.00 - 40.00	.01	.07	-.08
17. T2 SJT score	37.21	2.93	21.00 - 40.00	-.01	.11	**-.12**
18. T1-T2 difference	-0.83	2.96	-15.00 - 12.00	.02	-.03	.04
Construct validity						
19. Honesty-humility	56.25	6.93	16.00 - 72.00	.04	-.02	-.11

Appendix 4B continued

	7	8	9	10	11	12	13	14	15
1. Gender									
2. Dutch ethnicity									
3. 1st generation HE									
4. Age (on T2)									
5. pu-GPA									
6. ECA score									
Cognitive tests									
7. Mathematics									
8. Logical reasoning	**.35**								
9. Lecture	**.35**	**.27**							
Raw consensus									
10. T1 SJT score	**.12**	**.18**	**.16**						
11. T2 SJT score	.10	**.13**	**.16**	**.39**					
12. T1-T2 difference	.02	.04	0	**.54**	**-.57**				
Stan. consensus									
13. T1 SJT score	.03	**.18**	.08	**.53**	.08	**.41**			
14. T2 SJT score	.06	**.14**	.08	**.27**	**.33**	-.06	**.54**		
15. T1-T2 difference	-.03	.05	.01	**.30**	**-.25**	**.50**	**.54**	**-.42**	
Dich. consensus									
16. T1 SJT score	.09	**.17**	**.13**	**.59**	**.15**	**.39**	**.92**	**.48**	**.51**
17. T2 SJT score	.11	**.13**	.11	.32	**.42**	-.10	**.50**	**.88**	**-.34**
18. T1-T2 difference	-.01	.04	.02	**.28**	**-.26**	**.49**	**.42**	**-.39**	**.85**
Construct validity									
19. Honesty-humility	-.06	-.03	-.10	-.01	-.12	.10	**.17**	**.24**	-.03

Appendix 4B continued

	16	17	18
1. Gender			
2. Dutch ethnicity			
3. 1st generation HE			
4. Age (on T2)			
5. pu-GPA			
6. ECA score			
Cognitive tests			
7. Mathematics			
8. Logical reasoning			
9. Lecture			
Raw consensus			
10. T1 SJT score			
11. T2 SJT score			
12. T1-T2 difference			
Stan. consensus			
13. T1 SJT score			
14. T2 SJT score			
15. T1-T2 difference			
Dich. consensus			
16. T1 SJT score			
17. T2 SJT score	**.50**		
18. T1-T2 difference	**.51**	**-.49**	
Construct validity			
19. Honesty-humility	.13	**.25**	-.10

Note. M = Mean SD = Standard deviation [1]0 = female; 1 = male HE = Higher education pu-GPA = pre-university Grade Point Average [2]GPA ranges from 1 (low performance) to 10 (high performance) ECA = Extracurricular activities Stan. = Standardised Dich. = Dichotomous T1 = selection orientation day (low stakes) T2 = selection testing day (high stakes) Bold coefficients indicate a significant correlation ($p < .05$, two-tailed) Honesty-humility data was available for 171 applicants

Appendix 4C

Distribution of the judgements of the Subject Matter Experts across the rating scales (1: very inappropriate – 6: very appropriate) of the SJT items.

		1	2	3	4	5	6
S1.1	UD	78.3	17.4	4.3	0	0	0
S1.2	D	0	0	0	13	26.1	60.9
S1.3	UD	73.9	26.1	0	0	0	0
S1.4	D	0	0	4.3	8.7	30.4	56.5
S2.1	UD	0	33.3	61.1	5.6	0	0
S2.2	D	0	0	0	44.4	55.6	0
S2.3	UD	5.6	33.3	27.8	27.8	5.6	0
S2.4	D	0	0	5.6	16.7	38.9	38.9
S3.1	D	11.1	5.6	5.6	27.8	27.8	22.2
S3.2	UD	16.7	50	16.7	11.1	5.6	0
S3.3	UD	16.7	50	33.3	0	0	0
S3.4	D	0	5.9	11.8	35.3	23.5	23.5
S4.1	D	0	4.5	13.6	45.5	31.8	4.5
S4.2	UD	27.3	50	13.6	4.5	4.5	0
S4.3	D	0	0	0	4.5	31.8	63.6
S4.4	UD	18.2	36.4	36.4	0	9.1	0
S5.1	UD	54.5	22.7	13.6	4.5	4.5	0
S5.2	D	0	0	0	13.6	31.8	54.5
S5.3	UD	59.1	18.2	13.6	0	4.5	4.5
S5.4	D	0	4.5	4.5	22.7	22.7	45.5
S6.1	UD	5.6	22.2	22.2	27.8	11.1	11.1
S6.2	UD	11.1	27.8	44.4	5.6	0	11.1
S6.3	D	11.1	0	5.6	50	33.3	0
S6.4	D	11.1	22.2	22.2	27.8	16.7	0
S7.1	D	0	4.5	0	0	22.7	72.7
S7.2	UD	63.6	27.3	4.5	0	4.5	0
S7.3	D	0	0	0	18.2	36.4	45.5
S7.4	UD	40.9	22.7	22.7	4.5	9.1	0
S8.1	D	4.5	0	13.6	27.3	27.3	27.3
S8.2	D	4.5	0	0	18.2	36.4	40.9
S8.3	UD	40.9	27.3	18.2	9.1	4.5	0
S8.4	UD	68.1	18.2	13.6	0	0	0
S9.1	UD	22.2	27.8	16.7	33.3	0	0
S9.2	D	0	0	0	5.6	38.9	55.6
S9.3	UD	16.7	27.8	44.4	11.1	0	0
S9.4	D	0	0	5.6	11.1	33.3	50
S10.1	D	0	0	0	22.7	22.7	54.5
S10.2	UD	68.2	13.6	13.6	0	4.5	0
S10.3	UD	68.2	22.7	4.5	0	4.5	0
S10.4	D	0	0	0	4.5	18.2	77.3

Note. D = Desirable response option UD = Undesirable response option

Appendix 5A

Applicant perception items

Label	Item
Perceived predictive validity	How would you rate the effectiveness of a Situational Judgement Test for identifying qualified people for medical school? (1: *very ineffective* – 7: *very effective*)
Perceived fairness	If you would not be admitted based on a Situational Judgement Test, what would you think of the fairness of this procedure? (1: *very unfair* – 7: *very fair*)
Face validity	A Situational Judgement Test is a logical test for identifying qualified applicants for medical school. (1: *strongly disagree* – 7: *strongly agree*)
Applicant differentiation	A Situational Judgement Test measures an individual's important qualities, differentiating them from others. (1: *strongly disagree* – 7: *strongly agree*)
Study relatedness	A person who scores well on a Situational Judgement Test will be a good medical student. (1: *strongly disagree* – 7: *strongly agree*)
Chance to perform	I could really show my skills and abilities through a Situational Judgement Test. (1: *strongly disagree* – 7: *strongly agree*)
Ease of cheating	It is easy to cheat or fake on a Situational Judgement Test. (1: strongly disagree – 7: *strongly agree*)

Appendix 5B

Means (and standard deviations) for process favourability and the other applicant perception items for the four SJT versions.

	Should do rating	Should do pick-one	Would do rating	Would do pick-one
Process favourability	4.45 (1.20)	4.35 (1.31)	4.47 (1.17)	4.28 (1.43)
Face validity	4.39 (1.14)	4.38 (1.41)	4.33 (1.33)	4.22 (1.62)
Applicant differentiation	4.45 (1.59)	3.86 (1.49)	4.15 (1.45)	4.02 (1.96)
Study relatedness	3.74 (1.39)	3.45 (1.34)	3.70 (1.28)	3.43 (1.48)
Chance to perform	3.84 (1.57)	3.41 (1.60)	3.76 (1.49)	3.43 (1.64)
Ease of cheating	4.70 (1.85)	5.15 (1.84)	5.20 (1.80)	5.45 (1.78)

Means (and standard deviations) for process favourability and the other applicant perception items for the four SJT versions for men and women.

	Should do rating		Should do pick-one		Would do rating		Would do pick-one	
	Men	Women	Men	Women	Men	Women	Men	Women
Process favourability	4.29 (1.36)	4.49 (1.16)	4.16 (1.35)	4.43 (1.30)	4.86 (0.93)	4.32 (1.23)	4.10 (1.66)	4.34 (1.37)
Face validity	4.14 (1.71)	4.46 (1.26)	4.43 (1.24)	4.36 (1.47)	4.69 (1.26)	4.19 (1.34)	3.70 (1.84)	4.36 (1.54)
Applicant differentiation	4.05 (2.04)	4.56 (1.26)	3.71 (1.57)	3.91 (1.46)	4.73 (1.34)	3.93 (1.44)	3.55 (1.88)	4.15 (1.62)
Study relatedness	3.95 (1.43)	3.68 (1.83)	3.21 (1.41)	3.54 (1.32)	4.00 (1.41)	3.58 (1.21)	2.95 (1.64)	3.56 (1.41)
Chance to perform	4.10 (1.64)	3.76 (1.55)	3.00 (1.56)	3.55 (1.60)	3.96 (1.51)	3.69 (1.48)	3.50 (1.82)	3.41 (1.60)
Ease of cheating	4.86 (1.96)	4.66 (1.84)	5.32 (1.84)	5.09 (1.85)	4.85 (1.97)	5.34 (1.72)	5.95 (1.64)	5.32 (1.80)

Means (and standard deviations) for process favourability and the other applicant perception items for the four SJT versions for first-generation university students and non-first-generation university students.

	Should do rating		Should do pick-one		Would do rating		Would do pick-one	
	1st gen.		1st gen.		1st gen.		1st gen.	
	yes	no	yes	no	yes	no	yes	no
Process favourability	4.79 (0.95)	4.25 (1.26)	3.87 (1.48)	4.58 (1.14)	4.61 (1.04)	4.39 (1.25)	4.39 (1.45)	4.21 (1.47)
Face validity	4.93 (1.12)	4.05 (1.42)	4.05 (1.31)	4.52 (1.40)	4.74 (1.06)	4.12 (1.40)	4.26 (1.73)	4.11 (1.60)
Applicant differentiation	4.46 (1.45)	4.35 (1.70)	3.11 (1.29)	4.16 (1.45)	4.41 (1.55)	4.03 (1.36)	4.11 (1.88)	3.96 (1.68)
Study relatedness	3.96 (1.32)	3.60 (1.36)	3.21 (1.13)	3.60 (1.35)	3.89 (1.16)	3.69 (1.31)	3.53 (1.68)	3.36 (1.46)
Chance to perform	4.29 (1.41)	3.58 (1.56)	3.11 (1.63)	3.59 (1.57)	3.89 (1.42)	3.64 (1.51)	3.37 (1.92)	3.39 (1.57)
Ease of cheating	4.43 (1.45)	4.92 (1.93)	5.21 (2.10)	5.19 (1.76)	5.26 (1.53)	5.17 (1.90)	5.53 (1.84)	5.49 (1.78)

Note. 1st gen. = first-generation university student

Means (and standard deviations) for process favourability and the other applicant perception items for the four SJT versions for applicant of a Dutch, non-Western and Western ethnic background.

	Should do rating			Should do pick-one		
	D	NW	W	D	NW	W
Process favourability	4.43 (1.17)	4.81 (0.99)	3.64 (1.38)	4.41 (1.21)	4.29 (1.29)	5.00 (1.52)
Face validity	4.20 (1.43)	5.05 (1.07)	3.64 (1.21)	4.38 (1.29)	4.47 (1.68)	4.67 (1.51)
Applicant differentiation	4.39 (1.65)	4.81 (1.44)	3.55 (1.57)	3.82 (1.43)	4.11 (1.52)	4.50 (1.87)
Study relatedness	3.75 (1.27)	4.05 (1.40)	2.91 (1.45)	3.40 (1.19)	3.79 (1.55)	3.83 (1.72)
Chance to perform	3.70 (1.60)	4.48 (1.21)	3.09 (1.45)	3.44 (1.55)	3.63 (1.67)	3.50 (1.87)
Ease of cheating	4.89 (1.83)	4.52 (1.75)	4.55 (1.86)	5.21 (1.87)	5.32 (1.70)	4.83 (2.14)
	Would do rating			Should do pick-one		
	D	NW	W	D	NW	W
Process favourability	4.43 (1.18)	4.21 (1.21)	5.19 (1.07)	4.37 (1.47)	4.03 (1.46)	3.75 (1.19)
Face validity	4.37 (1.33)	4.06 (1.35)	4.50 (1.41)	4.15 (1.63)	4.28 (1.78)	3.75 (0.96)
Applicant differentiation	4.00 (1.32)	4.18 (1.85)	5.25 (0.71)	3.97 (1.71)	4.22 (1.90)	3.25 (1.26)
Study relatedness	3.65 (1.22)	3.94 (1.44)	4.13 (1.25)	3.48 (1.54)	3.17 (1.47)	3.00 (1.16)
Chance to perform	3.65 (1.48)	3.88 (1.50)	3.88 (1.55)	3.47 (1.66)	3.22 (1.73)	3.00 (1.16)
Ease of cheating	5.10 (1.83)	5.53 (1.46)	5.25 (2.19)	5.61 (1.76)	5.17 (2.04)	5.00 (0.82)

Note. D = Dutch ethnic background NW = Non-Western ethnic background W = Western ethnic background (not Dutch)

Appendix 6A

Performance for the different machine learning algorithms and the different approaches to address class imbalance for the different SJT versions calculated with either the raw, standardised or dichotomous consensus scoring method and for either all, only desirable or only undesirable response options.

Table 6A1

True positive rate, true negative rate and area under the curve for different machine learning algorithms and different approaches to address class imbalance caused by the low base rate of unprofessional behaviour for the SJT score calculated with the raw consensus method, based on all response options.

			Random undersampling			SMOTE oversampling			Cost-sensitive $C_{FN} : C_{FP}$			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
KNN (k = 1)	TPR (%)	8.3	13.9	16.7	16.7	57.4	50.0	36.2	8.3	8.3	8.3	8.3
	TNR (%)	90.4	75.0	75.0	74.7	62.8	69.8	75.1	90.4	90.4	90.4	90.4
	AUC	.49	.44	.46	.47	.61	.60	.56	.53	.53	.46	.46
KNN (k = 3)	TPR (%)	0	11.1	5.6	5.6	53.2	24.2	14.9	0	0	30.6	30.6
	TNR (%)	99.2	87.5	89.8	96.2	63.6	85.6	87.2	100	100	62.8	62.8
	AUC	.47	.42	.34	.42	.60	.58	.53	.48	.48	.47	.47
KNN (k = 5)	TPR (%)	2.8	11.1	2.8	0	48.9	12.9	10.6	0	0	8.3	44.4
	TNR (%)	100	86.1	95.4	99.3	63.6	91.7	94.1	100	100	91.2	43.9
	AUC	.45	.39	.34	.41	.61	.61	.51	.46	.46	.43	.46
DT	TPR (%)	0	5.6	0	0	30.9	3.2	0	0	0	8.3	11.1
	TNR (%)	100	96.5	100	100	90.1	98.4	100	100	100	94.4	92.0
	AUC	.46	.39	.46	.46	.70	.50	.47	.46	.46	.45	.38
RI	TPR (%)	0	5.6	0	2.8	27.7	4.8	4.3	0	0	16.7	8.3
	TNR (%)	100	90.3	95.4	99.3	90.4	97.3	97.6	100	100	94.4	92.2
	AUC	.39	.45	.36	.44	.64	.52	.57	.46	.46	.46	.47

Table 6A1continued

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
NN	TPR (%)	2.8	13.9	5.6	5.6	45.7	25.8	17.0	0	0	5.6	13.9
	TNR (%)	93.9	79.9	90.3	91.3	84.8	89.6	93.3	100	99.7	92.0	88.8
	AUC	.50	.42	.46	.50	.68	.61	.60	.47	.49	.51	.49
NB	TPR (%)	0	11.1	2.8	2.8	45.7	30.6	6.4	0	0	11.1	22.2
	TNR (%)	98.9	77.1	95.8	97.6	82.6	89.8	95.5	99.7	99.7	88.2	77.0
	AUC	.45	.42	.39	.42	.69	.60	.53	.45	.45	.45	.45
SVM	TPR (%)	0	0	0	0	0	0	0	0	0	0	0
	TNR (%)	100	100	100	100	100	100	100	100	100	100	100
	AUC	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Note. KNN = *k*-nearest neighbourhood DT = Decision tree RI = Rule induction NN = Neural network NB = Naive Bayes SVM = Support Vector Machine TPR = True positive rate TNR = True negative rate AUC = Area under the curve SMOTE = Synthetic Minority Oversampling Technique C_{FP} = cost of a false positive C_{FN} = cost of a false negative

Table 6A2

True positive rate, true negative rate and area under the curve for different machine learning algorithms and different approaches to address class imbalance caused by the low base rate of unprofessional behaviour for the SJT score calculated with the raw consensus method, based on desirable response options.

			Random undersampling			SMOTE oversampling			Cost-sensitive $C_{FN} : C_{FP}$			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
KNN (k = 1)	TPR (%)	13.9	27.8	11.1	11.1	74.5	46.8	38.3	13.9	13.9	13.9	13.9
	TNR (%)	82.1	56.9	77.3	78.1	59.4	71.9	80.2	82.1	82.1	82.1	82.1
	AUC	.48	.42	.43	.48	.66	.61	.59	.50	.50	.47	.47
KNN (k = 3)	TPR (%)	2.8	25.0	8.3	5.6	64.9	35.5	19.1	0	0	38.9	38.9
	TNR (%)	97.9	67.4	90.7	94.8	62.3	77.5	91.4	100	100	60.4	60.4
	AUC	.50	.44	.48	.46	.67	.61	.61	.52	.51	.52	.51
KNN (k = 5)	TPR (%)	0	11.1	5.6	0	56.4	25.8	6.4	0	0	11.1	58.3
	TNR (%)	100	85.4	95.4	98.3	65.0	86.9	97.1	100	100	89.3	41.2
	AUC	.50	.38	.45	.45	.65	.59	.54	.52	.52	.49	.51
DT	TPR (%)	0	0	0	0	31.9	21.0	4.3	0	0	0	11.1
	TNR (%)	100	97.9	100	100	90.4	97.1	98.9	100	100	95.2	90.9
	AUC	.46	.42	.46	.46	.67	.62	.55	.46	.46	.44	.42
RI	TPR (%)	2.8	2.8	2.8	2.8	37.2	21.0	10.6	0	0	13.9	8.3
	TNR (%)	99.5	93.1	98.1	97.9	94.1	93.0	97.1	100	100	94.9	88.5
	AUC	.45	.41	.47	.43	.71	.59	.56	.46	.46	.45	.45

Table 6A2 continued

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
NN	TPR (%)	5.6	16.7	2.8	8.3	41.5	24.2	23.4	0	0	22.2	11.1
	TNR (%)	95.7	77.8	85.6	90.6	87.2	93.0	94.7	100	99.7	88.2	88.8
	AUC	.53	.51	.44	.47	.68	.61	.55	.48	.49	.59	.50
NB	TPR (%)	0	0	0	0	47.9	30.6	8.5	0	0	5.6	16.7
	TNR (%)	98.9	94.4	96.8	95.5	83.4	91.4	96.0	99.7	99.7	89.3	77.3
	AUC	.41	.42	.39	.40	.69	.61	.51	.41	.41	.41	.41
SVM	TPR (%)	0	0	0	0	0	0	0	0	0	0	0
	TNR (%)	100	100	100	100	100	100	100	100	100	100	100
	AUC	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Note. KNN = *k*-nearest neighbourhood DT = Decision tree RI = Rule induction NN = Neural network NB = Naive Bayes SVM = Support Vector Machine TPR = True positive rate TNR = True negative rate AUC = Area under the curve SMOTE = Synthetic Minority Oversampling Technique C_{FP} = cost of a false positive C_{FN} = cost of a false negative

Table 6A3

True positive rate, true negative rate and area under the curve for different machine learning algorithms and different approaches to address class imbalance caused by the low base rate of unprofessional behaviour for the SJT score calculated with the raw consensus method, based on undesirable response options.

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
KNN (k = 1)	TPR (%)	8.3	16.7	5.6	2.8	54.3	45.2	25.5	8.3	8.3	8.3	8.3
	TNR (%)	93.9	68.8	90.7	91.7	69.3	81.0	90.6	93.9	93.9	93.9	93.9
	AUC	.51	.43	.47	.48	.62	.64	.57	.54	.54	.48	.48
KNN (k = 3)	TPR (%)	2.8	22.2	2.8	5.6	55.3	40.3	17.0	0	0	25.0	25.0
	TNR (%)	98.7	81.9	95.8	97.2	67.4	86.6	95.2	100	100	69.5	69.5
	AUC	.48	.47	.51	.48	.63	.64	.61	.50	.50	.47	.47
KNN (k = 5)	TPR (%)	2.8	11.1	2.8	2.8	55.3	22.6	8.5	0	0	8.3	38.9
	TNR (%)	99.7	88.9	98.6	99.3	68.7	90.9	95.7	100	100	92.0	47.6
	AUC	.44	.42	.45	.49	.62	.66	.55	.46	.46	.42	.45
DT	TPR (%)	0	0	0	0	33.0	12.9	6.4	0	0	5.6	5.6
	TNR (%)	100	100	100	100	93.3	97.1	98.1	100	100	93.6	88.8
	AUC	.46	.46	.46	.46	.68	.61	.64	.46	.46	.43	.40
RI	TPR (%)	0	2.8	0	0	34.0	19.4	10.6	0	0	8.3	5.6
	TNR (%)	100	94.4	98.1	97.9	92.0	96.3	96.8	100	100	92.0	93.0
	AUC	.37	.41	.41	.43	.70	.65	.56	.46	.46	.42	.48

Table 6A3 continued

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
NN	TPR (%)	2.8	13.9	5.6	5.6	40.4	25.8	8.5	0	0	2.8	8.3
	TNR (%)	95.2	80.6	85.2	91.3	81.3	89.6	92.0	100	99.7	92.0	87.7
	AUC	.48	.44	.46	.53	.66	.57	.57	.47	.47	.40	.50
NB	TPR (%)	0	8.3	5.6	2.8	45.7	27.4	8.5	0	0	5.6	16.7
	TNR (%)	98.9	94.4	96.3	94.4	82.4	89.3	95.5	99.7	99.7	87.7	74.9
	AUC	.42	.43	.41	.40	.68	.61	.51	.42	.42	.42	.42
SVM	TPR (%)	0	0	0	0	0	0	0	0	0	0	0
	TNR (%)	100	100	100	100	100	100	100	100	100	100	100
	AUC	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Note. KNN = *k*-nearest neighbourhood DT = Decision tree RI = Rule induction NN = Neural network NB = Naive Bayes SVM = Support Vector Machine TPR = True positive rate TNR = True negative rate AUC = Area under the curve SMOTE = Synthetic Minority Oversampling Technique C_{FP} = cost of a false positive C_{FN} = cost of a false negative

Table 6A4

True positive rate, true negative rate and area under the curve for different machine learning algorithms and different approaches to address class imbalance caused by the low base rate of unprofessional behaviour for the SJT score calculated with the standardised consensus method, based on all response options.

			Random undersampling			SMOTE oversampling			Cost-sensitive $C_{FN} : C_{FP}$			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
KNN (k = 1)	TPR (%)	8.3	22.2	11.1	5.6	52.1	45.2	23.4	8.3	8.3	8.3	8.3
	TNR (%)	85.6	63.2	87.0	85.1	71.9	77.0	79.9	85.6	85.6	85.6	85.6
	AUC	.47	.43	.50	.48	.63	.61	.52	.50	.50	.44	.44
KNN (k = 3)	TPR (%)	2.8	22.2	8.3	8.3	51.1	27.4	12.8	0	0	25.0	25.0
	TNR (%)	98.7	76.4	92.1	97.6	81.3	85.8	95.7	100	100	63.6	63.6
	AUC	.45	.46	.50	.46	.67	.61	.54	.46	.46	.46	.46
KNN (k = 5)	TPR (%)	2.8	11.1	2.8	2.8	46.8	19.4	6.4	0	0	5.6	55.6
	TNR (%)	99.7	82.6	95.8	99.3	82.1	89.8	96.5	100	100	90.4	41.4
	AUC	.48	.41	.49	.53	.67	.58	.54	.50	.50	.46	.52
DT	TPR (%)	0	0	0	0	34.0	8.1	6.4	0	0	8.3	8.3
	TNR (%)	100	100	100	100	90.9	97.6	99.5	100	100	93.6	89.8
	AUC	.46	.46	.46	.46	.70	.57	.55	.46	.46	.44	.40
RI	TPR (%)	2.8	2.8	0	2.8	35.1	9.7	8.5	0	0	13.9	11.1
	TNR (%)	99.7	95.1	98.1	97.6	91.2	97.3	96.8	100	100	92.5	88.5
	AUC	.38	.46	.40	.42	.68	.60	.57	.46	.46	.46	.43

Table 6A4 continued

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
NN	TPR (%)	5.6	11.1	5.6	11.1	40.4	24.2	17.0	0	0	13.9	8.3
	TNR (%)	95.5	80.6	88.4	91.3	85.8	91.7	93.0	100	99.5	89.8	86.4
	AUC	.49	.46	.48	.55	.66	.63	.59	.46	.49	.48	.47
NB	TPR (%)	0	5.6	2.8	2.8	50.0	21.0	4.3	0	0	8.3	16.7
	TNR (%)	97.9	94.4	95.8	94.8	84.0	90.6	95.2	99.2	99.2	88.8	77.5
	AUC	.40	.40	.37	.37	.69	.58	.48	.40	.40	.40	.40
SVM	TPR (%)	0	0	0	0	0	0	0	0	0	0	0
	TNR (%)	100	100	100	100	100	100	100	100	100	100	100
	AUC	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Note. KNN = *k*-nearest neighbourhood DT = Decision tree RI = Rule induction NN = Neural network NB = Naive Bayes SVM = Support Vector Machine TPR = True positive rate TNR = True negative rate AUC = Area under the curve SMOTE = Synthetic Minority Oversampling Technique C_{FP} = cost of a false positive C_{FN} = cost of a false negative

Table 6A5

True positive rate, true negative rate and area under the curve for different machine learning algorithms and different approaches to address class imbalance caused by the low base rate of unprofessional behaviour for the SJT score calculated with the standardised consensus method, based on desirable response options.

			Random undersampling			SMOTE oversampling			Cost-sensitive $C_{FN} : C_{FP}$			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
KNN (k = 1)	TPR (%)	11.1	25.0	13.9	8.3	63.8	46.8	40.4	11.1	11.1	11.1	11.1
	TNR (%)	82.6	52.1	77.3	77.1	61.2	73.3	74.6	82.6	82.6	82.6	82.6
	AUC	.47	.39	.46	.44	.61	.60	.58	.49	.49	.45	.45
KNN (k = 3)	TPR (%)	2.8	19.4	13.9	13.9	62.8	32.3	14.9	0	0	33.3	33.3
	TNR (%)	97.3	72.9	89.4	92.4	66.0	84.2	88.5	100	100	58.6	58.6
	AUC	.46	.44	.51	.45	.66	.61	.57	.48	.48	.50	.49
KNN (k = 5)	TPR (%)	2.8	8.3	2.8	2.8	56.4	24.2	10.6	0	0	13.9	55.6
	TNR (%)	99.5	81.9	98.1	97.9	71.4	86.4	92.5	100	100	87.2	40.6
	AUC	.49	.39	.48	.51	.66	.62	.55	.50	.50	.48	.53
DT	TPR (%)	0	0	0	0	34.0	9.7	2.1	0	0	2.8	5.6
	TNR (%)	100	100	100	100	91.2	97.9	98.7	100	100	92.0	88.5
	AUC	.46	.46	.46	.46	.72	.63	.52	.46	.46	.37	.34
RI	TPR (%)	2.8	2.8	0	2.8	35.1	14.5	14.9	0	0	5.6	8.3
	TNR (%)	99.7	94.4	98.1	97.9	92.0	94.4	96.8	100	100	90.1	88.2
	AUC	.39	.45	.40	.43	.75	.59	.59	.46	.46	.38	.43

Table 6A5 continued

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
NN	TPR (%)	5.6	11.1	19.4	8.3	42.6	24.2	17.0	0	0	11.1	8.3
	TNR (%)	96.5	79.9	89.4	92.0	88.0	91.2	89.8	100	99.7	89.8	89.0
	AUC	.51	.43	.50	.52	.68	.62	.51	.47	.49	.52	.52
NB	TPR (%)	2.8	5.6	2.8	2.8	46.8	30.6	10.6	0	0	8.3	16.7
	TNR (%)	97.9	94.4	96.3	95.8	85.3	90.4	95.7	98.9	98.9	89.3	77.8
	AUC	.40	.40	.37	.37	.70	.60	.48	.40	.40	.40	.40
SVM	TPR (%)	0	0	0	0	0	0	0	0	0	0	0
	TNR (%)	100	100	100	100	100	100	100	100	100	100	100
	AUC	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Note. KNN = *k*-nearest neighbourhood DT = Decision tree RI = Rule induction NN = Neural network NB = Naive Bayes SVM = Support Vector Machine TPR = True positive rate TNR = True negative rate AUC = Area under the curve SMOTE = Synthetic Minority Oversampling Technique C_{FP} = cost of a false positive C_{FN} = cost of a false negative

Table 6A6

True positive rate, true negative rate and area under the curve for different machine learning algorithms and different approaches to address class imbalance caused by the low base rate of unprofessional behaviour for the SJT score calculated with the standardised consensus method, based on undesirable response options.

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
KNN (k = 1)	TPR (%)	8.3	16.7	13.9	11.1	54.3	37.1	27.7	8.3	8.3	8.3	8.3
	TNR (%)	92.2	69.4	90.3	91.0	59.6	85.6	87.4	92.2	92.2	92.2	92.2
	AUC	.50	.43	.53	.53	.56	.63	.56	.54	.54	.47	.47
KNN (k = 3)	TPR (%)	2.8	22.2	8.3	5.6	53.2	29.0	14.9	0	0	27.8	27.8
	TNR (%)	98.9	81.9	93.1	97.6	64.2	88.8	93.6	100	100	69.0	69.0
	AUC	.49	.47	.48	.47	.61	.59	.59	.50	.50	.49	.50
KNN (k = 5)	TPR (%)	2.8	11.1	2.8	2.8	47.9	17.7	10.6	0	0	8.3	44.4
	TNR (%)	99.2	84.0	96.8	98.6	69.5	92.8	96.8	100	100	90.9	44.1
	AUC	.45	.42	.44	.49	.61	.61	.53	.46	.46	.42	.46
DT	TPR (%)	0	0	0	0	42.6	22.6	2.1	0	0	5.6	5.6
	TNR (%)	100	100	100	100	92.5	97.6	98.7	100	100	91.2	88.5
	AUC	.46	.46	.46	.46	.73	.60	.52	.46	.46	.43	.36
RI	TPR (%)	2.8	5.6	0	2.8	41.5	19.4	8.5	0	0	11.1	5.6
	TNR (%)	99.7	93.1	98.1	97.9	90.9	96.8	97.3	100	100	92.8	89.3
	AUC	.39	.48	.40	.42	.69	.60	.53	.46	.46	.41	.44

Table 6A6 continued

			Random undersampling			SMOTE oversampling			Cost-sensitive $C_{FN} : C_{FP}$			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
NN	TPR (%)	5.6	5.6	5.6	11.1	47.9	24.2	17.0	0	0	11.1	5.6
	TNR (%)	94.9	81.3	85.6	91.7	85.8	89.3	92.5	100	99.5	90.4	89.0
	AUC	.47	.43	.46	.54	.70	.58	.55	.46	.49	.48	.42
NB	TPR (%)	0	5.6	2.8	2.8	48.9	29.0	8.5	0	0	5.6	16.7
	TNR (%)	98.1	94.4	96.3	95.5	82.6	90.1	95.5	99.5	99.2	88.2	76.7
	AUC	.39	.39	.36	.36	.69	.60	.49	.39	.39	.39	.39
SVM	TPR (%)	0	0	0	0	0	0	0	0	0	0	0
	TNR (%)	100	100	100	100	100	100	100	100	100	100	100
	AUC	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Note. KNN = *k*-nearest neighbourhood DT = Decision tree RI = Rule induction NN = Neural network NB = Naive Bayes SVM = Support Vector Machine TPR = True positive rate TNR = True negative rate AUC = Area under the curve SMOTE = Synthetic Minority Oversampling Technique C_{FP} = cost of a false positive C_{FN} = cost of a false negative

Table 6A7

True positive rate, true negative rate and area under the curve for different machine learning algorithms and different approaches to address class imbalance caused by the low base rate of unprofessional behaviour for the SJT score calculated with the dichotomous consensus method, based on all response options.

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
KNN (k = 1)	TPR (%)	5.6	22.2	13.9	11.1	55.3	35.5	36.2	5.6	5.6	5.6	5.6
	TNR (%)	90.1	66.0	86.1	88.2	56.7	82.1	84.5	90.1	90.1	90.1	90.1
	AUC	.48	.44	.51	.52	.55	.59	.60	.51	.51	.45	.45
KNN (k = 3)	TPR (%)	2.8	27.8	8.3	5.6	57.4	25.8	17.0	0	0	27.8	27.8
	TNR (%)	98.7	74.3	93.1	95.8	61.0	86.9	92.8	100	100	61.5	61.5
	AUC	.45	.48	.54	.48	.59	.55	.59	.47	.47	.47	.47
KNN (k = 5)	TPR (%)	2.8	16.7	2.8	2.8	51.1	21.0	8.5	0	0	11.1	63.9
	TNR (%)	98.9	80.6	94.4	98.3	65.5	93.0	94.9	100	100	88.2	40.4
	AUC	.52	.42	.51	.52	.63	.61	.58	.54	.54	.50	.53
DT	TPR (%)	0	0	0	0	40.4	25.8	0	0	0	8.3	11.1
	TNR (%)	100	100	100	100	93.0	96.5	100	100	100	91.4	87.7
	AUC	.46	.46	.46	.46	.70	.60	.47	.46	.46	.44	.39
RI	TPR (%)	0	2.8	0	2.8	35.1	25.8	6.4	0	0	5.6	11.1
	TNR (%)	100	94.4	98.6	97.6	90.4	93.6	96.0	100	100	93.3	85.0
	AUC	.39	.44	.44	.42	.67	.57	.53	.46	.46	.42	.45

Table 6A7 continued

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
NN	TPR (%)	2.8	8.3	11.1	8.3	48.9	24.2	23.4	0	0	11.1	13.9
	TNR (%)	95.2	77.8	86.6	94.1	84.8	87.4	94.1	100	99.7	89.8	89.3
	AUC	.51	.45	.48	.52	.70	.60	.54	.45	.48	.54	.50
NB	TPR (%)	0	8.3	2.8	2.8	50.0	30.6	6.4	0	0	5.6	11.1
	TNR (%)	98.7	93.1	95.8	95.5	83.2	90.6	96.0	99.7	99.7	89.0	75.1
	AUC	.38	.39	.36	.36	.70	.60	.49	.38	.38	.39	.39
SVM	TPR (%)	0	0	0	0	0	0	0	0	0	0	0
	TNR (%)	100	100	100	100	100	100	100	100	100	100	100
	AUC	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Note. KNN = *k*-nearest neighbourhood DT = Decision tree RI = Rule induction NN = Neural network NB = Naive Bayes SVM = Support Vector Machine TPR = True positive rate TNR = True negative rate AUC = Area under the curve SMOTE = Synthetic Minority Oversampling Technique C_{FP} = cost of a false positive C_{FN} = cost of a false negative

Table 6A8

True positive rate, true negative rate and area under the curve for different machine learning algorithms and different approaches to address class imbalance caused by the low base rate of unprofessional behaviour for the SJT score calculated with the dichotomous consensus method, based on desirable response options.

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
KNN (k = 1)	TPR (%)	11.1	30.6	11.1	16.7	70.2	46.8	40.4	11.1	11.1	11.1	11.1
	TNR (%)	84.5	63.2	82.4	82.6	55.6	72.5	74.3	84.5	84.5	84.5	84.5
	AUC	.48	.47	.47	.51	.62	.61	.57	.50	.50	.46	.46
KNN (k = 3)	TPR (%)	5.6	25.0	16.7	8.3	67.0	40.3	25.5	0	0	47.2	47.2
	TNR (%)	95.5	69.4	91.2	91.7	58.8	81.0	84.5	100	100	55.3	55.3
	AUC	.52	.44	.51	.48	.63	.65	.59	.53	.53	.56	.55
KNN (k = 5)	TPR (%)	5.6	13.9	11.1	11.1	61.7	35.5	19.1	0	0	19.4	58.3
	TNR (%)	98.4	82.6	97.7	96.9	64.7	83.7	88.2	100	100	82.6	37.7
	AUC	.49	.39	.49	.48	.64	.64	.53	.51	.51	.48	.52
DT	TPR (%)	0	0	0	0	34.0	12.9	2.1	0	0	5.6	8.3
	TNR (%)	100	100	100	100	92.8	96.3	98.9	100	100	90.6	88.5
	AUC	.46	.46	.46	.46	.71	.56	.53	.46	.46	.38	.41
RI	TPR (%)	0	2.8	0	0	35.1	24.2	10.6	0	0	0	2.8
	TNR (%)	100	93.8	98.6	97.9	93.0	95.5	96.8	100	100	93.3	92.2
	AUC	.43	.46	.44	.42	.71	.60	.60	.46	.46	.37	.43

Table 6A8 continued

			Random undersampling			SMOTE oversampling			Cost-sensitive $C_{FN} : C_{FP}$			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
NN	TPR (%)	8.3	11.1	13.9	8.3	45.7	29.0	27.7	0	0	8.3	11.1
	TNR (%)	96.3	83.3	87.5	92.0	84.0	89.6	92.2	100	99.7	89.9	89.8
	AUC	.55	.45	.53	.51	.68	.67	.56	.47	.49	.49	.49
NB	TPR (%)	0	2.8	2.8	0	45.7	25.8	8.5	0	0	8.3	16.7
	TNR (%)	98.1	94.4	96.8	95.1	82.6	91.4	96.5	99.7	99.5	88.8	76.5
	AUC	.40	.39	.37	.36	.68	.60	.50	.40	.40	.40	.40
SVM	TPR (%)	0	0	0	0	0	0	0	0	0	0	0
	TNR (%)	100	100	100	100	100	100	100	100	100	100	100
	AUC	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Note. KNN = *k*-nearest neighbourhood DT = Decision tree RI = Rule induction NN = Neural network NB = Naive Bayes SVM = Support Vector Machine TPR = True positive rate TNR = True negative rate AUC = Area under the curve SMOTE = Synthetic Minority Oversampling Technique C_{FP} = cost of a false positive C_{FN} = cost of a false negative

Table 6A9

True positive rate, true negative rate and area under the curve for different machine learning algorithms and different approaches to address class imbalance caused by the low base rate of unprofessional behaviour for the SJT score calculated with the dichotomous consensus method, based on undesirable response options.

			Random undersampling			SMOTE oversampling			Cost-sensitive C_{FN} : C_{FP}			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
KNN (k = 1)	TPR (%)	5.6	22.2	8.3	8.3	47.9	33.9	31.9	5.6	5.6	5.6	5.6
	TNR (%)	94.4	69.4	90.7	93.1	67.1	88.0	88.8	94.4	94.4	94.4	94.4
	AUC	.50	.46	.50	.53	.58	.61	.59	.54	.54	.46	.46
KNN (k = 3)	TPR (%)	2.8	27.8	8.3	5.6	57.4	24.2	17.0	0	0	22.2	22.2
	TNR (%)	98.9	81.3	94.4	97.2	66.3	88.5	94.1	100	100	68.2	68.2
	AUC	.46	.49	.53	.49	.62	.58	.59	.48	.48	.44	.45
KNN (k = 5)	TPR (%)	2.8	8.3	5.6	2.8	45.7	16.1	6.4	0	0	11.1	52.8
	TNR (%)	99.7	87.5	96.3	99.0	69.0	93.9	96.3	100	100	91.2	46.3
	AUC	.50	.43	.49	.54	.62	.62	.56	.52	.52	.48	.52
DT	TPR (%)	0	0	0	0	40.4	12.9	0	0	0	2.8	8.3
	TNR (%)	100	100	100	100	91.4	96.8	100	100	100	92.2	86.9
	AUC	.46	.46	.46	.46	.71	.58	.47	.46	.46	.40	.39
RI	TPR (%)	0	5.6	0	2.8	29.8	17.7	4.3	0	0	8.3	5.6
	TNR (%)	99.7	94.4	98.1	97.6	94.4	94.1	96.8	100	100	92.0	91.2
	AUC	.38	.45	.41	.43	.69	.64	.49	.46	.46	.42	.45

Table 6A9 continued

			Random undersampling			SMOTE oversampling			Cost-sensitive $C_{FN} : C_{FP}$			
		~1:10	1:4	1:6	1:8	1:4	1:6	1:8	5:1	3:1	1:3	1:5
NN	TPR (%)	0	5.6	5.6	8.3	46.8	25.8	21.3	0	0	13.9	16.7
	TNR (%)	93.9	77.1	88.9	94.8	84.8	88.2	94.7	100	99.7	89.8	89.0
	AUC	.49	.41	.45	.46	.70	.60	.59	.45	.46	.53	.50
NB	TPR (%)	0	8.3	5.6	2.8	45.7	32.3	10.6	0	0	8.3	11.1
	TNR (%)	98.7	93.1	95.4	95.5	82.4	89.6	95.5	99.7	99.7	88.8	73.5
	AUC	.38	.41	.38	.36	.68	.60	.50	.38	.38	.38	.38
SVM	TPR (%)	0	0	0	0	0	0	0	0	0	0	0
	TNR (%)	100	100	100	100	100	100	100	100	100	100	100
	AUC	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50

Note. KNN = *k*-nearest neighbourhood DT = Decision tree RI = Rule induction NN = Neural network NB = Naive Bayes SVM = Support Vector Machine TPR = True positive rate TNR = True negative rate AUC = Area under the curve SMOTE = Synthetic Minority Oversampling Technique C_{FP} = cost of a false positive C_{FN} = cost of a false negative

Appendix 6B

Results of individual feature ranking for the different resampling methods for the different SJT versions calculated with either the raw, standardised or dichotomous consensus scoring method and for either all, only desirable or only undesirable response options.

Table 6B1
First, second and third position in the individual feature ranking for the different resampling methods for the SJT score calculated with the raw consensus method, based on all response options.

		Undersampling			SMOTE		
Feature	~1:10	1:4	1:6	1:8	1:4	1:6	1:8
Pre-university GPA					2	1	
Extracurricular activities							
Logical reasoning							
Scientific reading	3	3	3	3		3	3
Anatomy							
Mathematics							
Curriculum sample							
SJT					3		
Honesty-humility							
Emotionality	2	2	2	2			2
Extraversion							
Agreeableness							
Conscientiousness							
Openness to experience							
Coaching day	1	1	1	1	1	2	1

Note. GPA = Grade Point Average SJT = Situational Judgement Test SMOTE = Synthetic Minority Oversampling Technique

Table 6B2

First, second and third position in the individual feature ranking for the different resampling methods for the SJT score calculated with the raw consensus method, based on desirable response options.

		Undersampling			SMOTE		
Feature	~1:10	1:4	1:6	1:8	1:4	1:6	1:8
Pre-university GPA					3	1	
Extracurricular activities							
Logical reasoning							
Scientific reading	3	3	3	3		3	3
Anatomy							
Mathematics							
Curriculum sample							
SJT							
Honesty-humility							
Emotionality	2	2	2	2			2
Extraversion							
Agreeableness							
Conscientiousness							
Openness to experience					2		
Coaching day	1	1	1	1	1	2	1

Note. GPA = Grade Point Average SJT = Situational Judgement Test SMOTE = Synthetic Minority Oversampling Technique

Table 6B3

First, second and third position in the individual feature ranking for the different resampling methods for the SJT score calculated with the raw consensus method, based on undesirable response options.

		Undersampling			SMOTE		
Feature	~1:10	1:4	1:6	1:8	1:4	1:6	1:8
Pre-university GPA					1	1	
Extracurricular activities							
Logical reasoning							
Scientific reading	3	3	3	3		3	3
Anatomy							
Mathematics							
Curriculum sample							
SJT							
Honesty-humility							
Emotionality	2	2	2	2			2
Extraversion					2		
Agreeableness							
Conscientiousness							
Openness to experience							
Coaching day	1	1	1	1	3	2	1

Note. GPA = Grade Point Average SJT = Situational Judgement Test SMOTE = Synthetic Minority Oversampling Technique

Table 6B4

First, second and third position in the individual feature ranking for the different resampling methods for the SJT score calculated with the standardised consensus method, based on all response options.

		Undersampling			SMOTE		
Feature	~1:10	1:4	1:6	1:8	1:4	1:6	1:8
Pre-university GPA					3	1	
Extracurricular activities							
Logical reasoning							
Scientific reading	3	3	3	3		3	3
Anatomy							
Mathematics							
Curriculum sample							
SJT							
Honesty-humility							
Emotionality	2	2	2	2			2
Extraversion							
Agreeableness							
Conscientiousness							
Openness to experience					1		
Coaching day	1	1	1	1	2	2	1

Note. GPA = Grade Point Average SJT = Situational Judgement Test SMOTE = Synthetic Minority Oversampling Technique

Table 6B5

First, second and third position in the individual feature ranking for the different resampling methods for the SJT score calculated with the standardised consensus method, based on desirable response options.

		Undersampling			SMOTE		
Feature	~1:10	1:4	1:6	1:8	1:4	1:6	1:8
Pre-university GPA					2	1	
Extracurricular activities							
Logical reasoning							
Scientific reading	3	3	3	3		3	3
Anatomy							
Mathematics							
Curriculum sample							
SJT							
Honesty-humility							
Emotionality	2	2	2	2			2
Extraversion					3		
Agreeableness							
Conscientiousness							
Openness to experience							
Coaching day	1	1	1	1	1	2	1

Note. GPA = Grade Point Average SJT = Situational Judgement Test SMOTE = Synthetic Minority Oversampling Technique

Table 6B6

First, second and third position in the individual feature ranking for the different resampling methods for the SJT score calculated with the standardised consensus method, based on undesirable response options.

		Undersampling			SMOTE		
Feature	~1:10	1:4	1:6	1:8	1:4	1:6	1:8
Pre-university GPA					3	2	
Extracurricular activities							
Logical reasoning							
Scientific reading	3	3	3	3			3
Anatomy							
Mathematics							
Curriculum sample							
SJT							
Honesty-humility							
Emotionality	2	2	2	2			2
Extraversion							
Agreeableness							
Conscientiousness							
Openness to experience					1	1	
Coaching day	1	1	1	1	2	3	1

Note. GPA = Grade Point Average SJT = Situational Judgement Test SMOTE = Synthetic Minority Oversampling Technique

Table 6B7

First, second and third position in the individual feature ranking for the different resampling methods for the SJT score calculated with the dichotomous consensus method, based on all response options.

		Undersampling			SMOTE		
Feature	~1:10	1:4	1:6	1:8	1:4	1:6	1:8
Pre-university GPA					3	1	
Extracurricular activities							
Logical reasoning							
Scientific reading	3	3	3	3		3	3
Anatomy							
Mathematics							
Curriculum sample							
SJT							
Honesty-humility							
Emotionality	2	2	2	2			2
Extraversion					1		
Agreeableness							
Conscientiousness							
Openness to experience							
Coaching day	1	1	1	1	2	2	1

Note. GPA = Grade Point Average SJT = Situational Judgement Test SMOTE = Synthetic Minority Oversampling Technique

Table 6B8

First, second and third position in the individual feature ranking for the different resampling methods for the SJT score calculated with the dichotomous consensus method, based on desirable response options.

		Undersampling			SMOTE		
Feature	~1:10	1:4	1:6	1:8	1:4	1:6	1:8
Pre-university GPA					3	1	
Extracurricular activities							
Logical reasoning							
Scientific reading	3	3	3	3		3	3
Anatomy							
Mathematics							
Curriculum sample							
SJT							
Honesty-humility							
Emotionality	2	2	2	2			2
Extraversion							
Agreeableness							
Conscientiousness							
Openness to experience					1		
Coaching day	1	1	1	1	2	2	1

Note. GPA = Grade Point Average SJT = Situational Judgement Test SMOTE = Synthetic Minority Oversampling Technique

Table 6B9

First, second and third position in the individual feature ranking for the different resampling methods for the SJT score calculated with the dichotomous consensus method, based on undesirable response options.

		Undersampling			SMOTE		
Feature	~1:10	1:4	1:6	1:8	1:4	1:6	1:8
Pre-university GPA					2	2	
Extracurricular activities							
Logical reasoning							
Scientific reading	3	3	3	3			3
Anatomy							
Mathematics							
Curriculum sample							
SJT							
Honesty-humility							
Emotionality	2	2	2	2			2
Extraversion					3		
Agreeableness							
Conscientiousness							
Openness to experience						1	
Coaching day	1	1	1	1	1	3	1

Note. GPA = Grade Point Average SJT = Situational Judgement Test SMOTE = Synthetic Minority Oversampling Technique

Appendix 6C

True positive rate (% TPR) for each approach to address class imbalance (x-axis) and for each SJT version (separate lines) per machine learning algorithm (separate graphs).

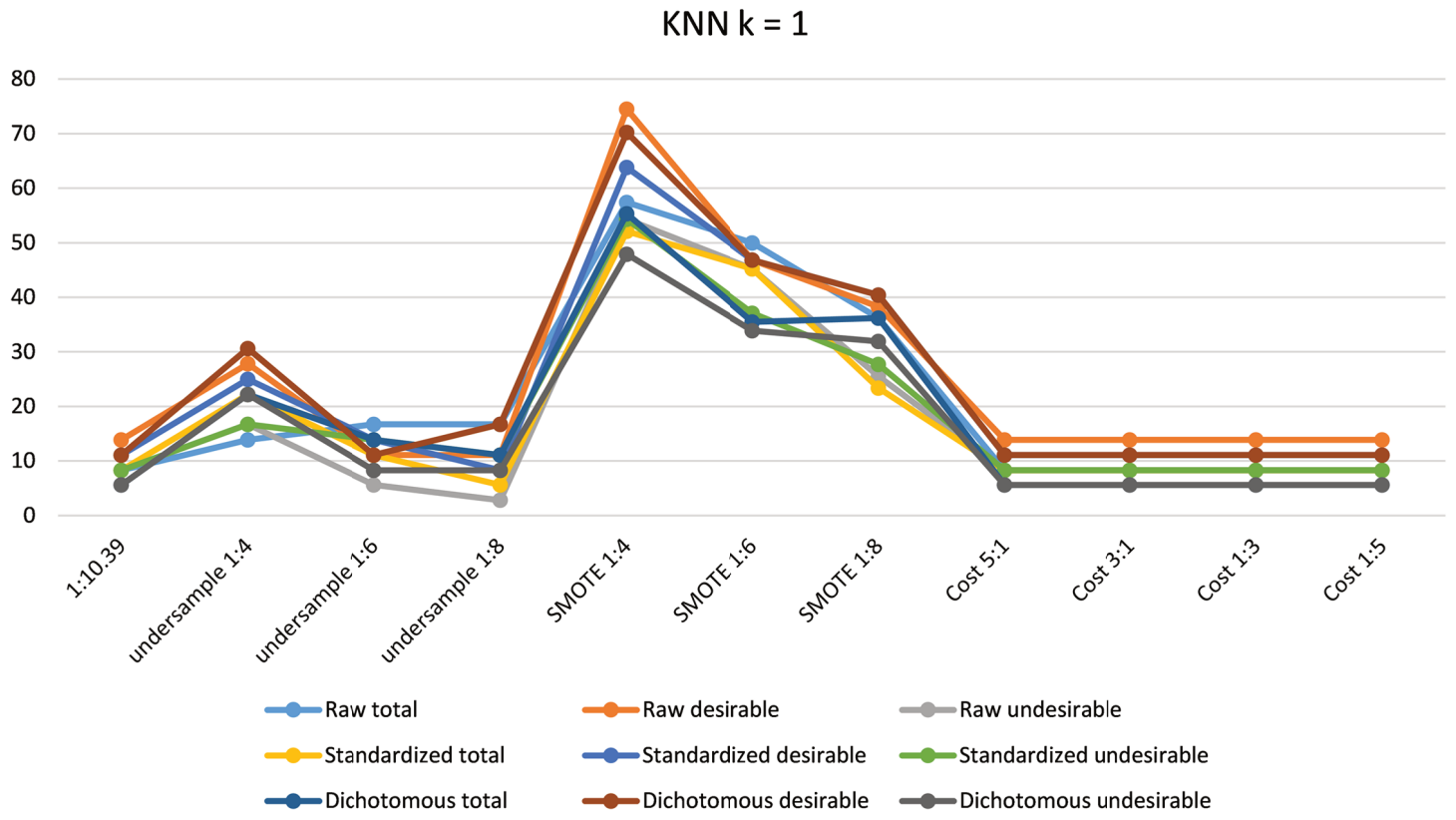

Figure 6C1. True positive rate (% TPR) for each approach to address class imbalance (x-axis) and for each SJT version (separate lines) for the *k*-nearest neighbourhood (KNN) algorithm using $k = 1$.

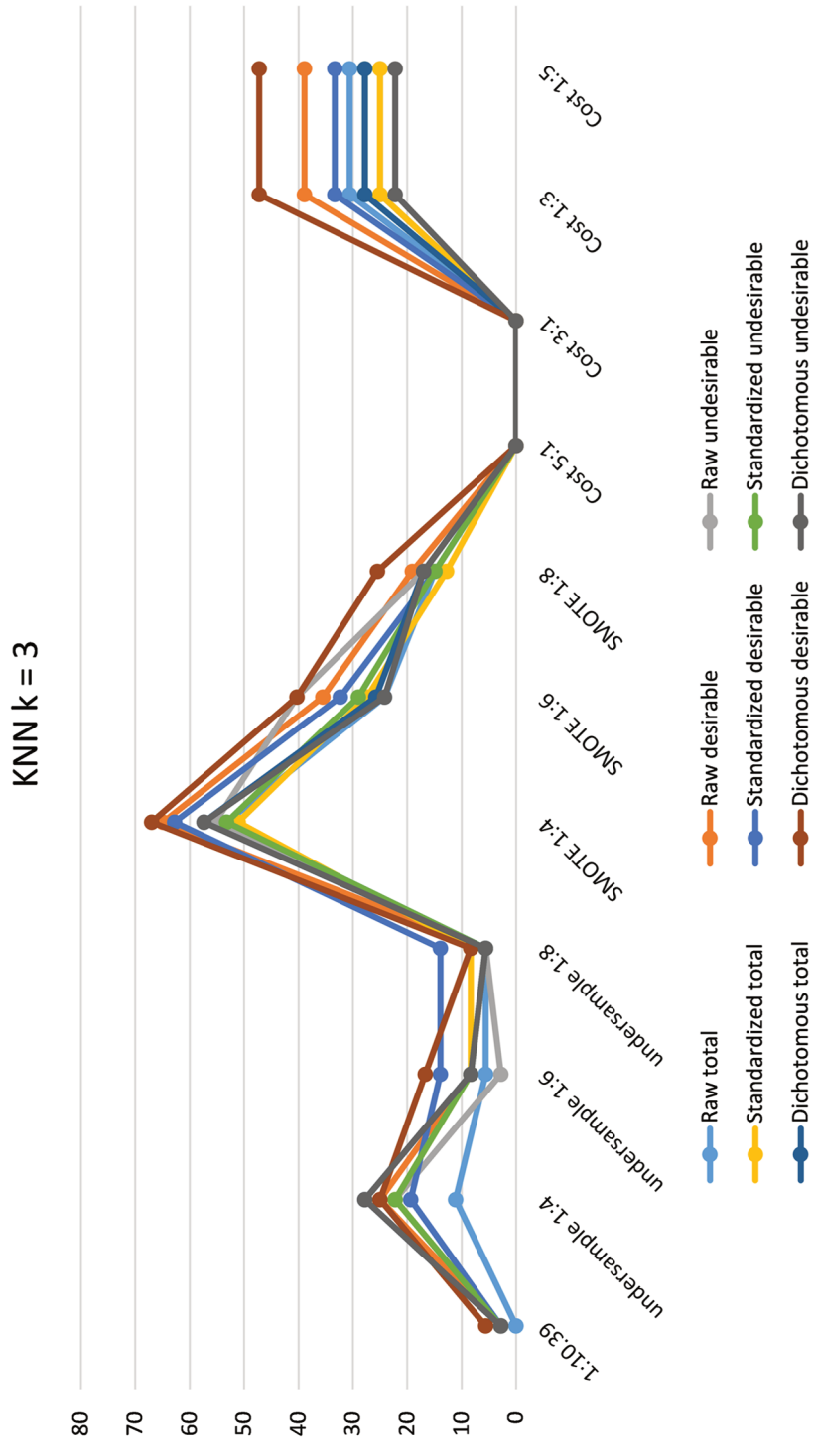

Figure 6C2. True positive rate (% TPR) for each approach to address class imbalance (x-axis) and for each SJT version (separate lines) for the *k*-nearest neighbourhood (KNN) algorithm using $k = 3$.

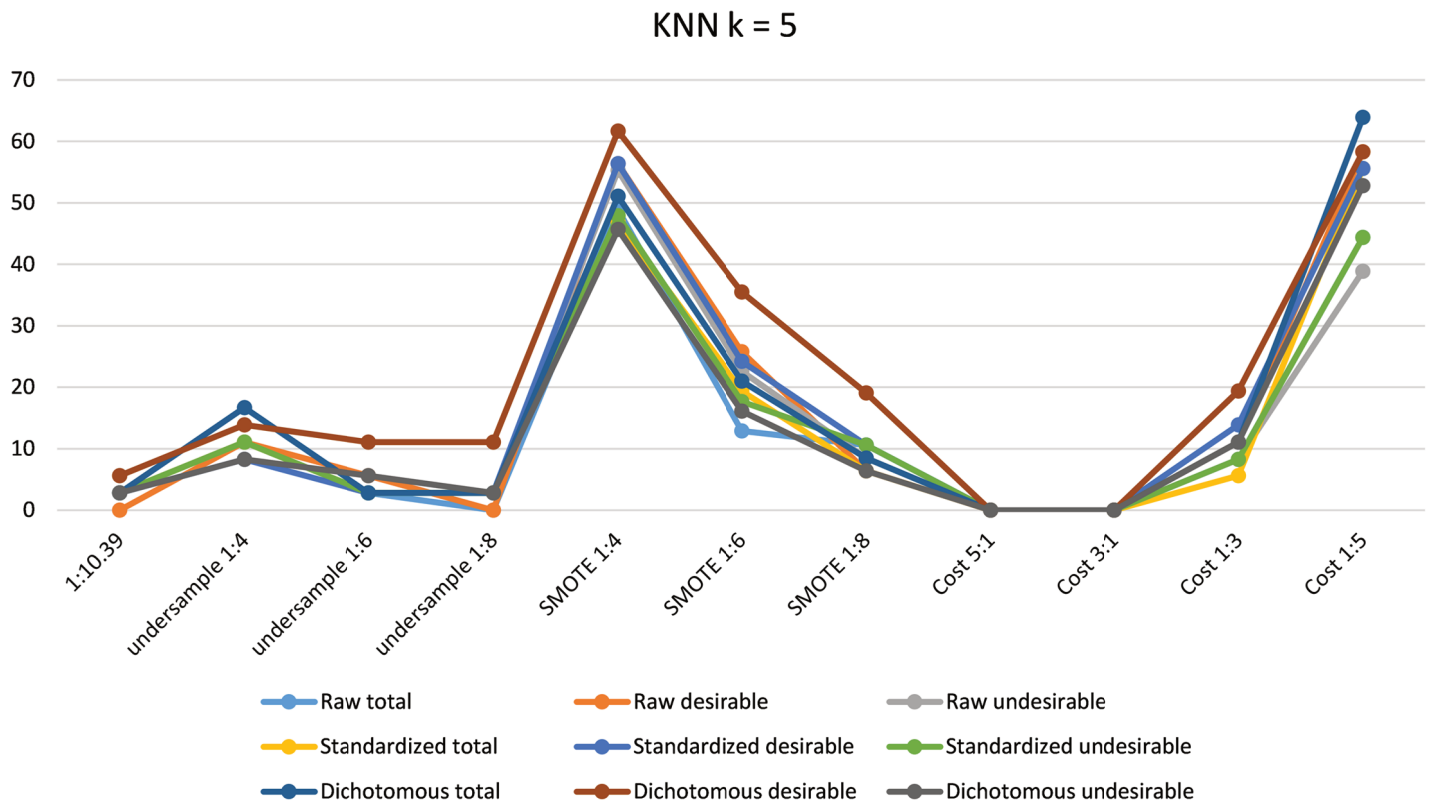

Figure 6C3. True positive rate (% TPR) for each approach to address class imbalance (x-axis) and for each SJT version (separate lines) for the *k*-nearest neighbourhood (KNN) algorithm using $k = 5$.

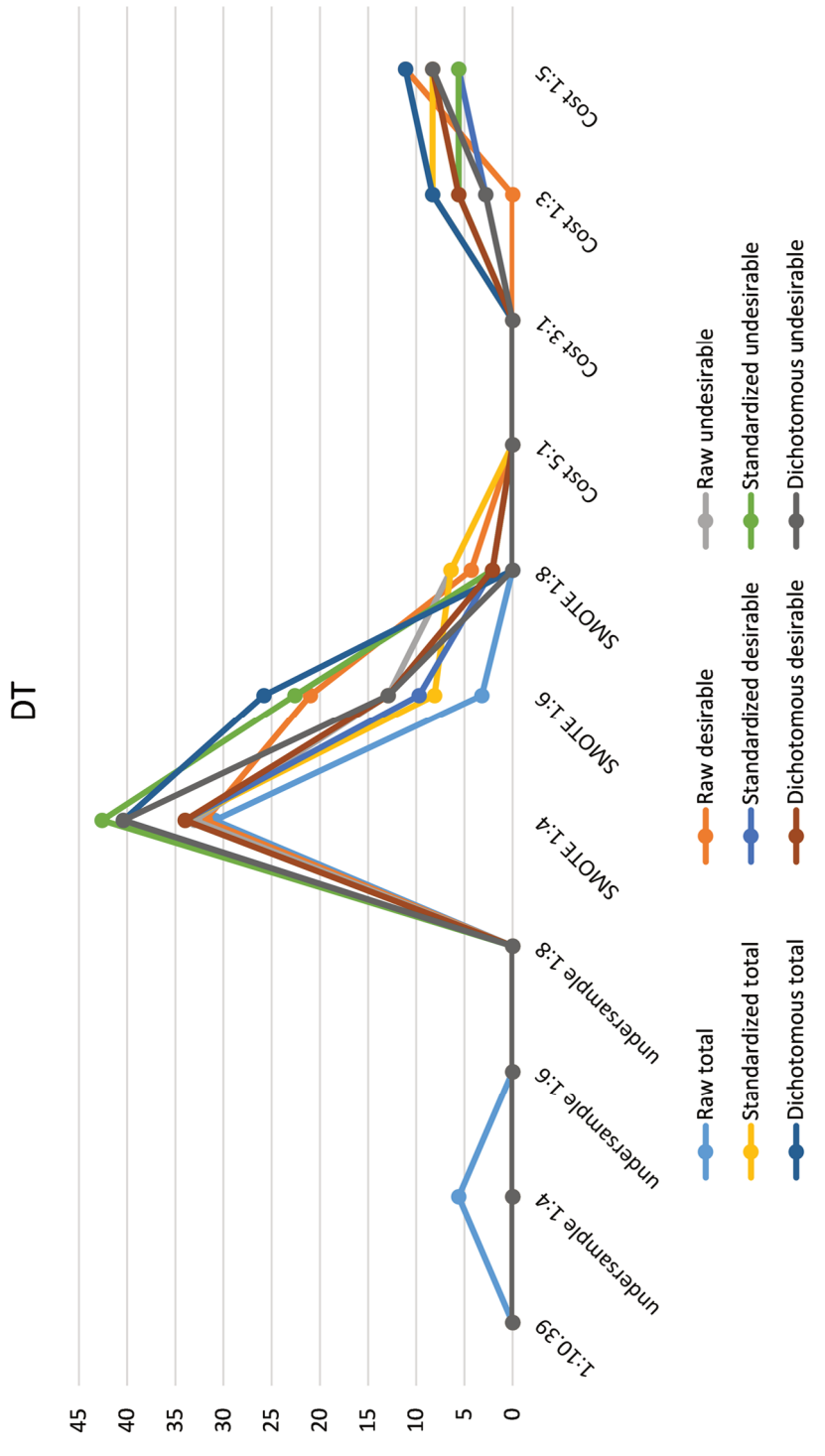

Figure 6C4. True positive rate (% TPR) for each approach to address class imbalance (x-axis) and for each SJT version (separate lines) for the decision tree (DT) algorithm.

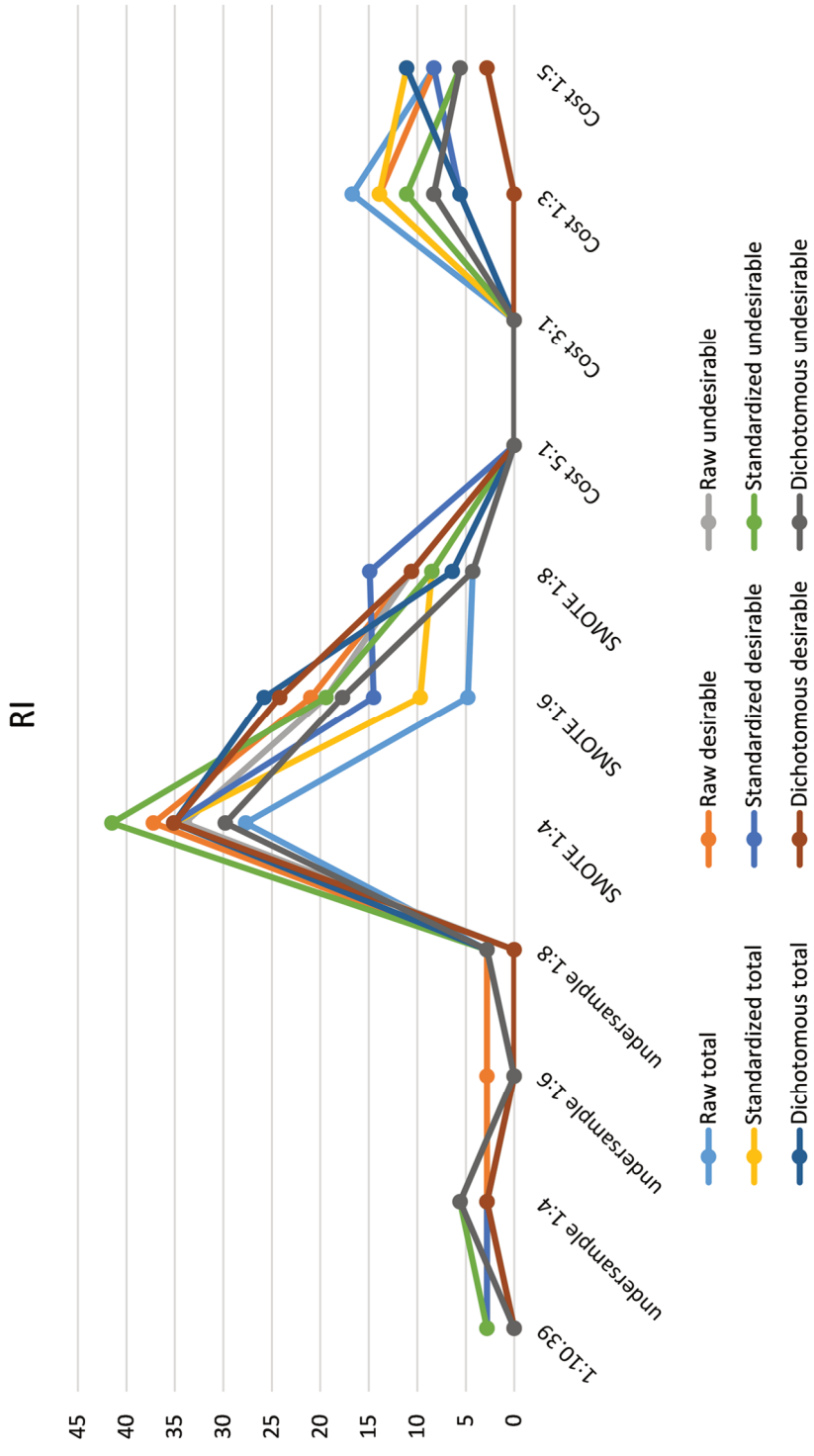

Figure 6C5. True positive rate (% TPR) for each approach to address class imbalance (x-axis) and for each SJT version (separate lines) for the rule induction (RI) algorithm.

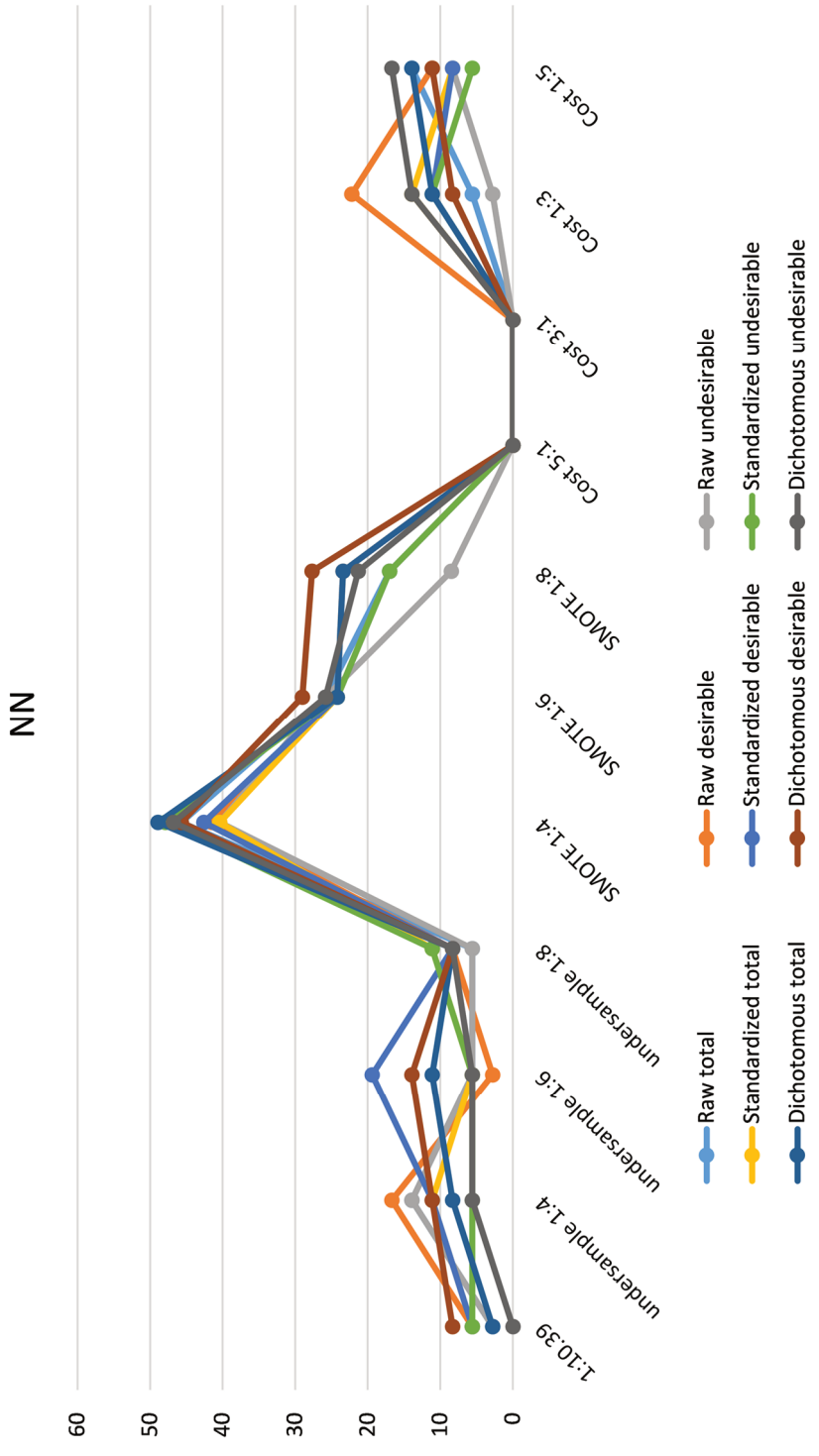

Figure 6C6. True positive rate (% TPR) for each approach to address class imbalance (x-axis) and for each SJT version (separate lines) for the neural networks (NN) algorithm.

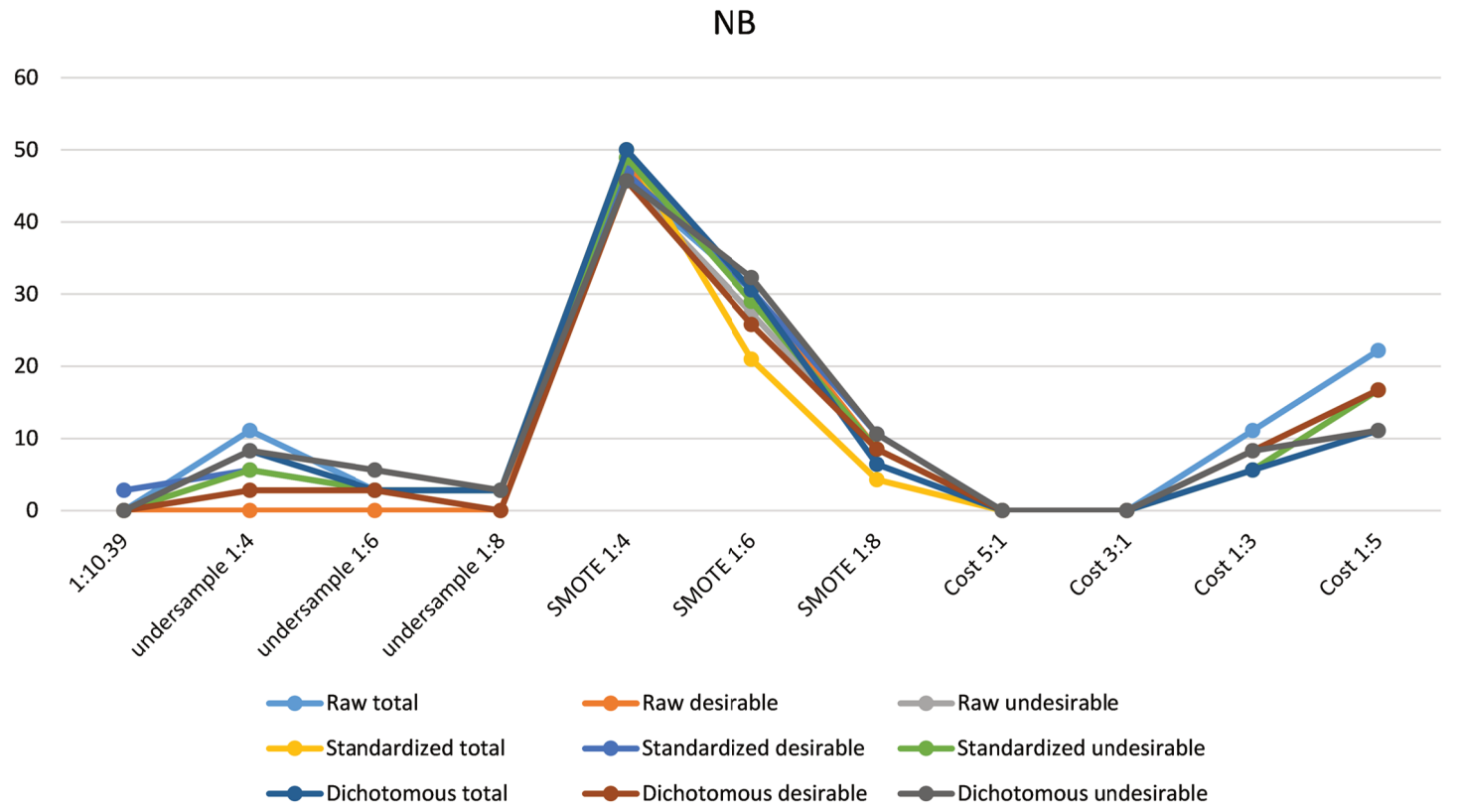

Figure 6C7. True positive rate (% TPR) for each approach to address class imbalance (x-axis) and for each SJT version (separate lines) for the naive Bayes (NB) algorithm.

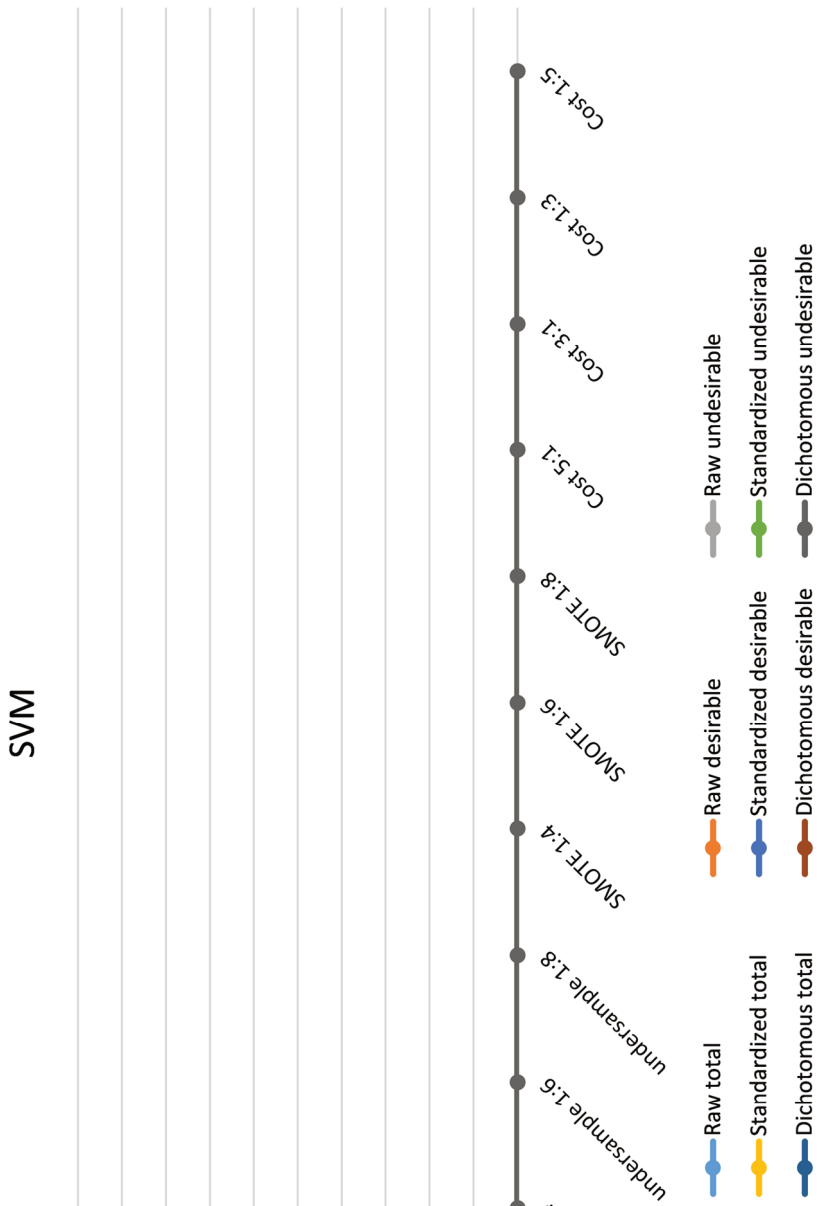

Figure 6C8. True positive rate (% TPR) for each approach to address class imbalance (x-axis) and for each SJT version (separate lines) for the support vector machine (SVM) algorithm.

Appendix 6D

True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) per approach to address class imbalance (separate graphs).

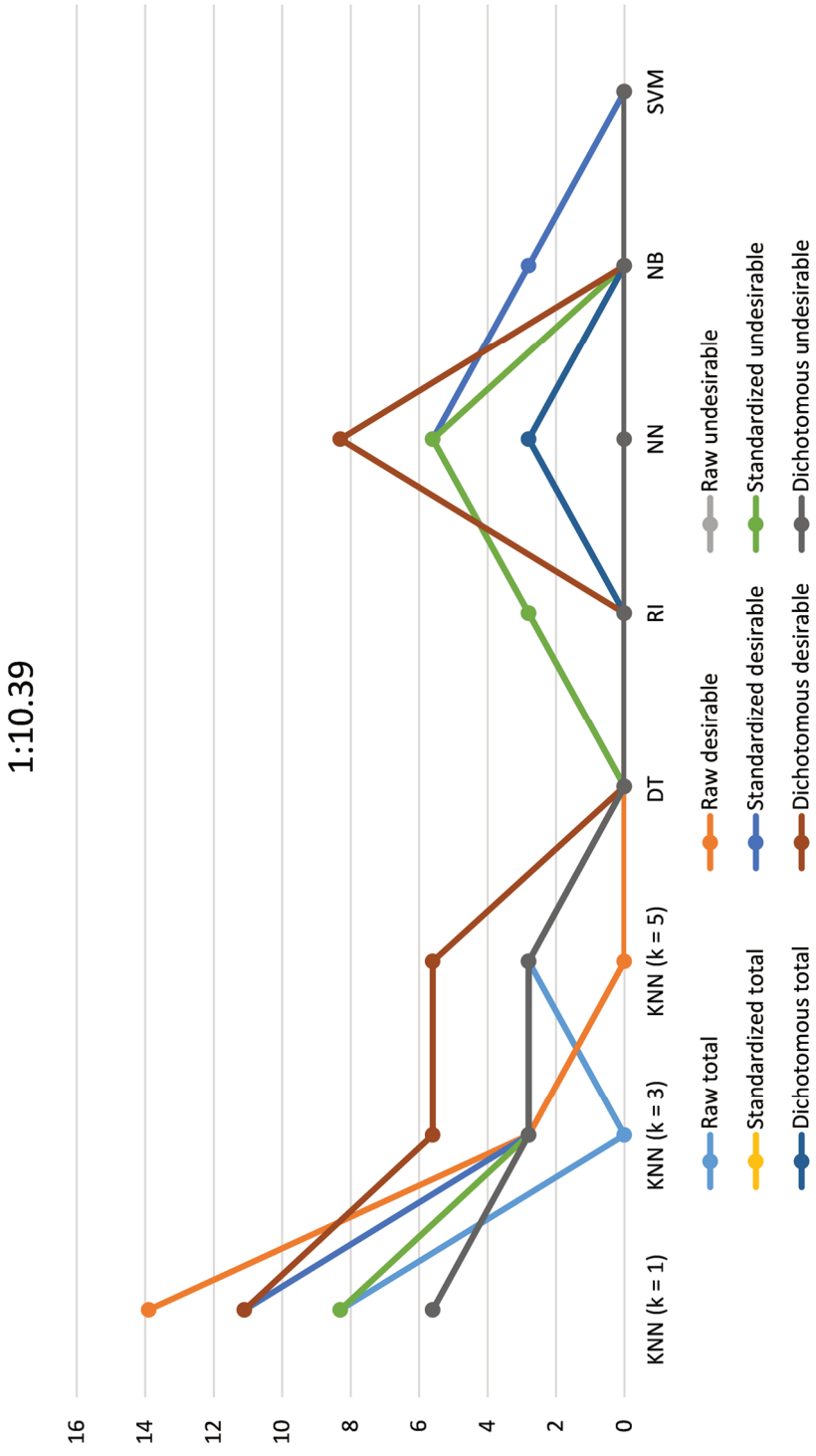

Figure 6D1. True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) for the imbalanced dataset.

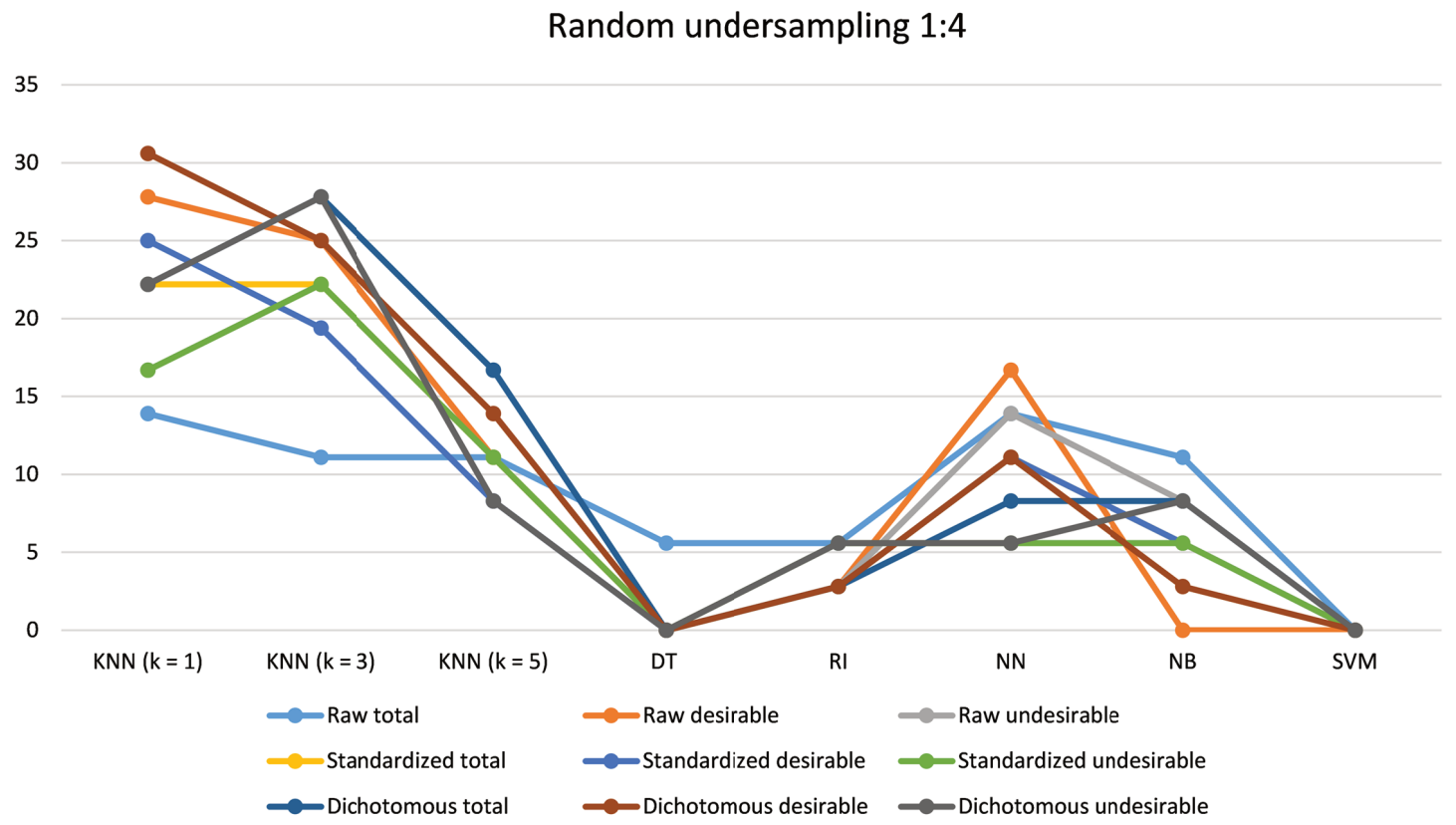

Figure 6D2. True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) for random undersampling 1:4.

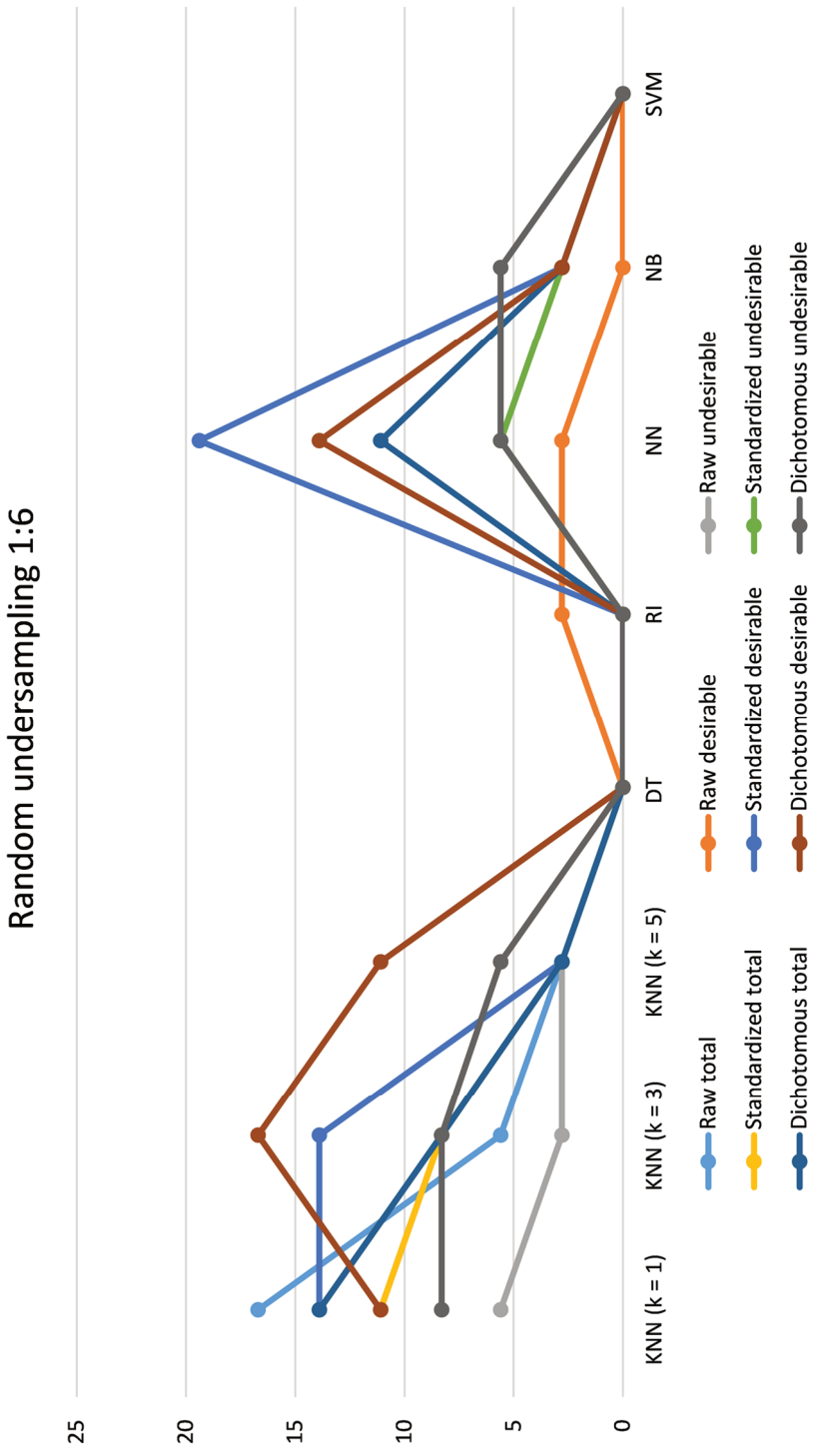

Figure 6D3. True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) for random undersampling 1:6.

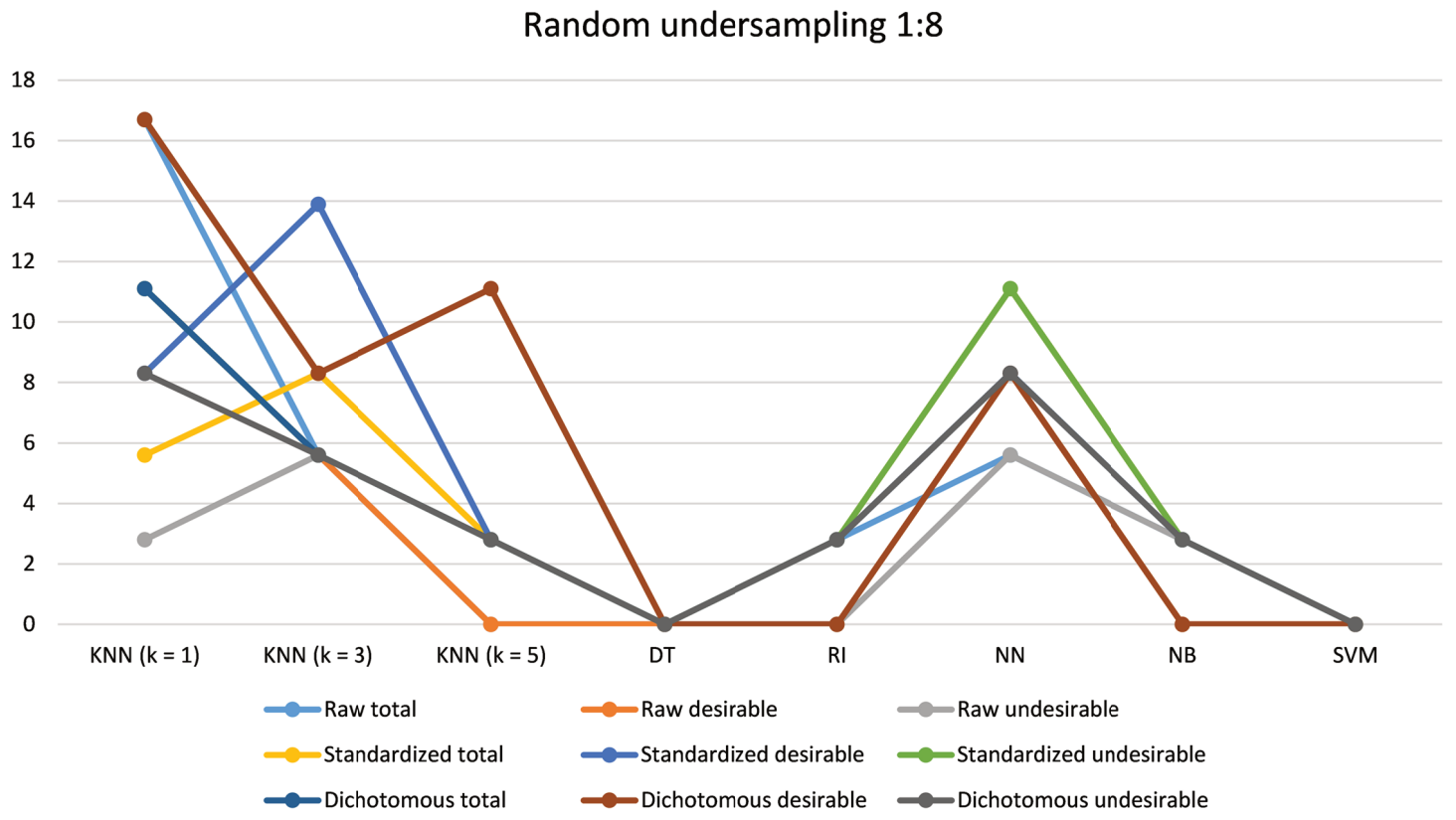

Figure 6D4. True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) for random undersampling 1:8.

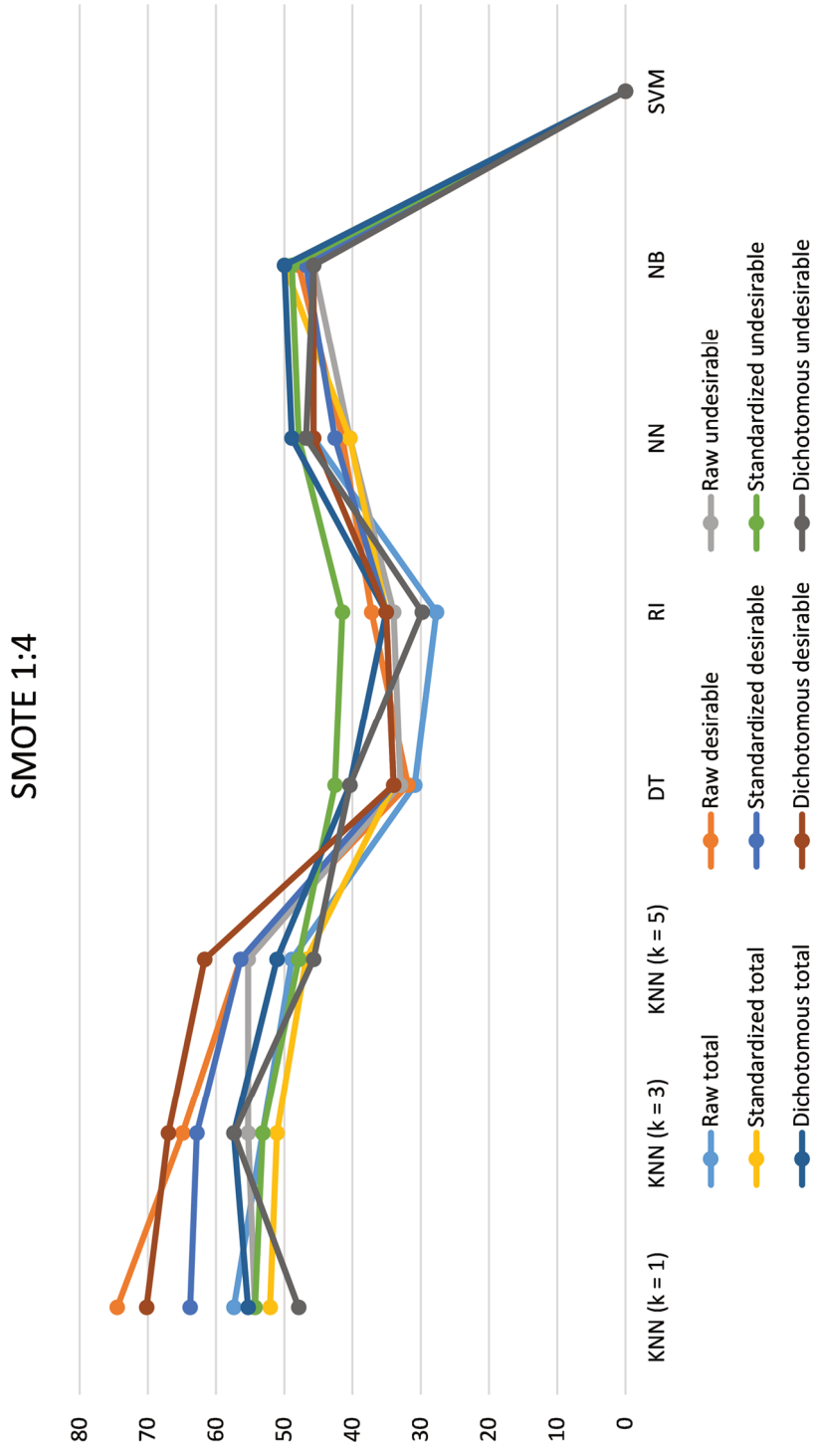

Figure 6D5. True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) for synthethic minority oversampling (SMOTE) 1:4.

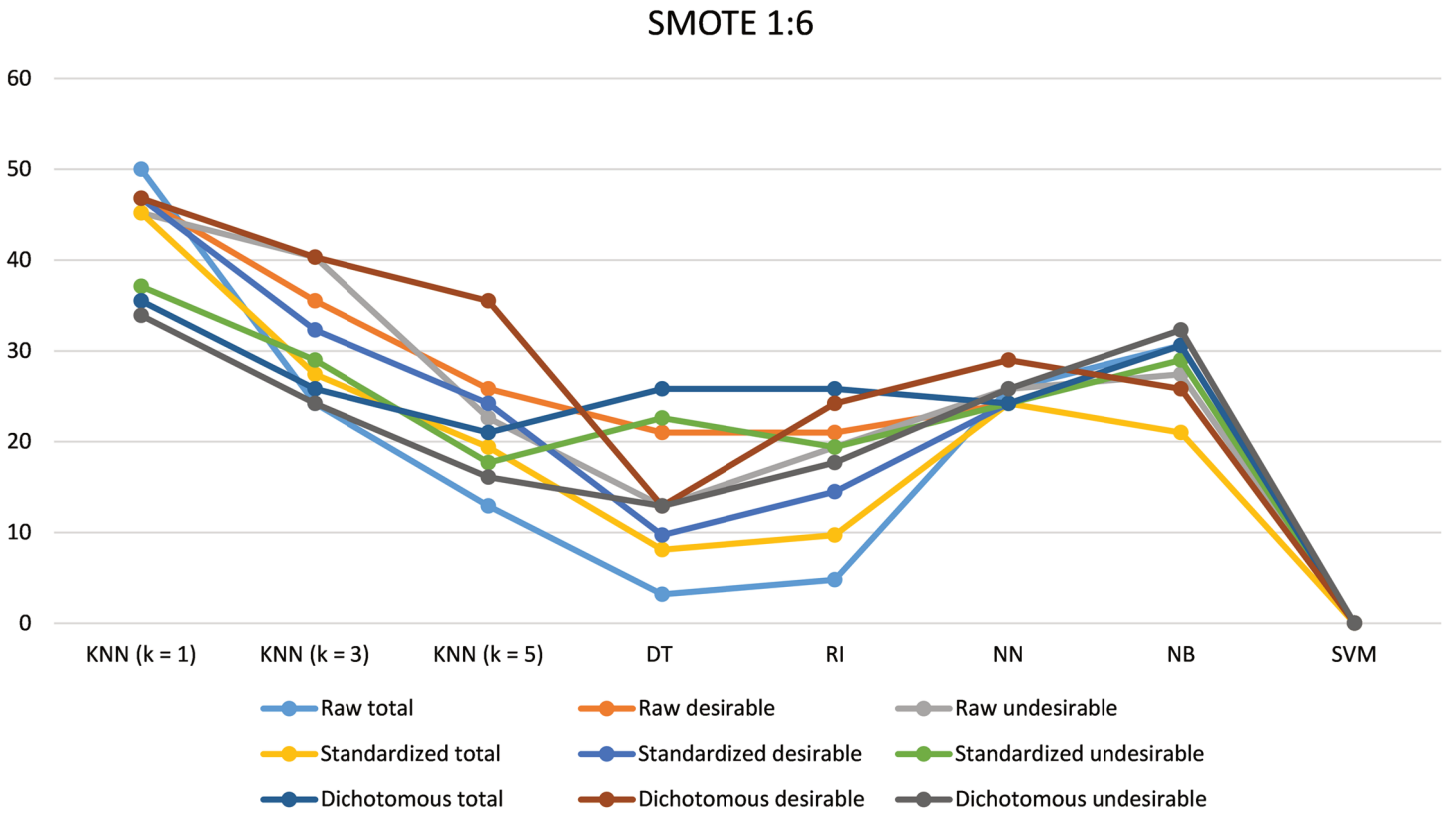

Figure 6D6. True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) for synthethic minority oversampling (SMOTE) 1:6.

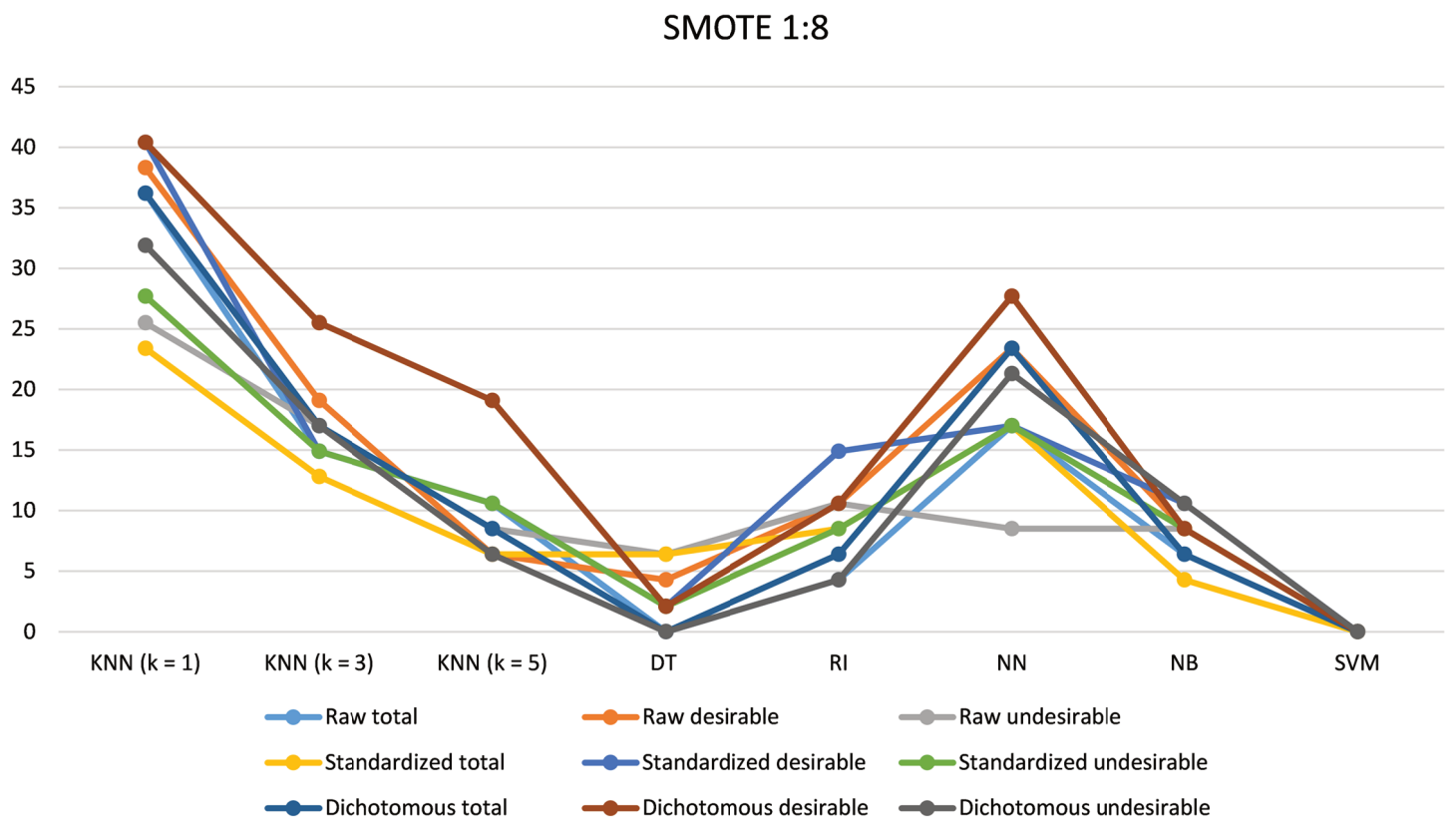

Figure 6D7. True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) for synthethic minority oversampling (SMOTE) 1:8.

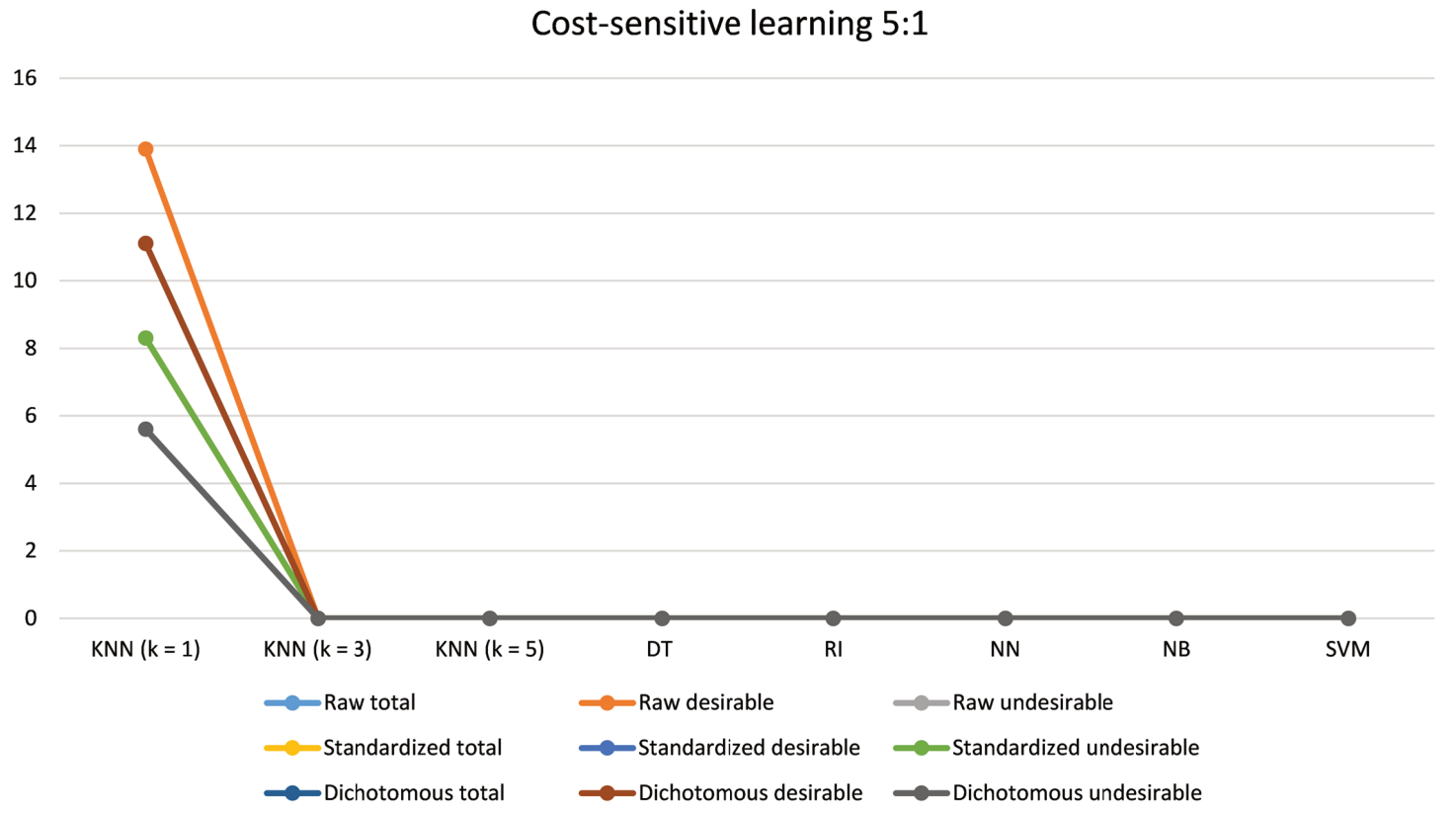

Figure 6D8. True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) for cost-sensitive learning $C_{False\ Negative} : C_{False\ Positive} = 5:1$.

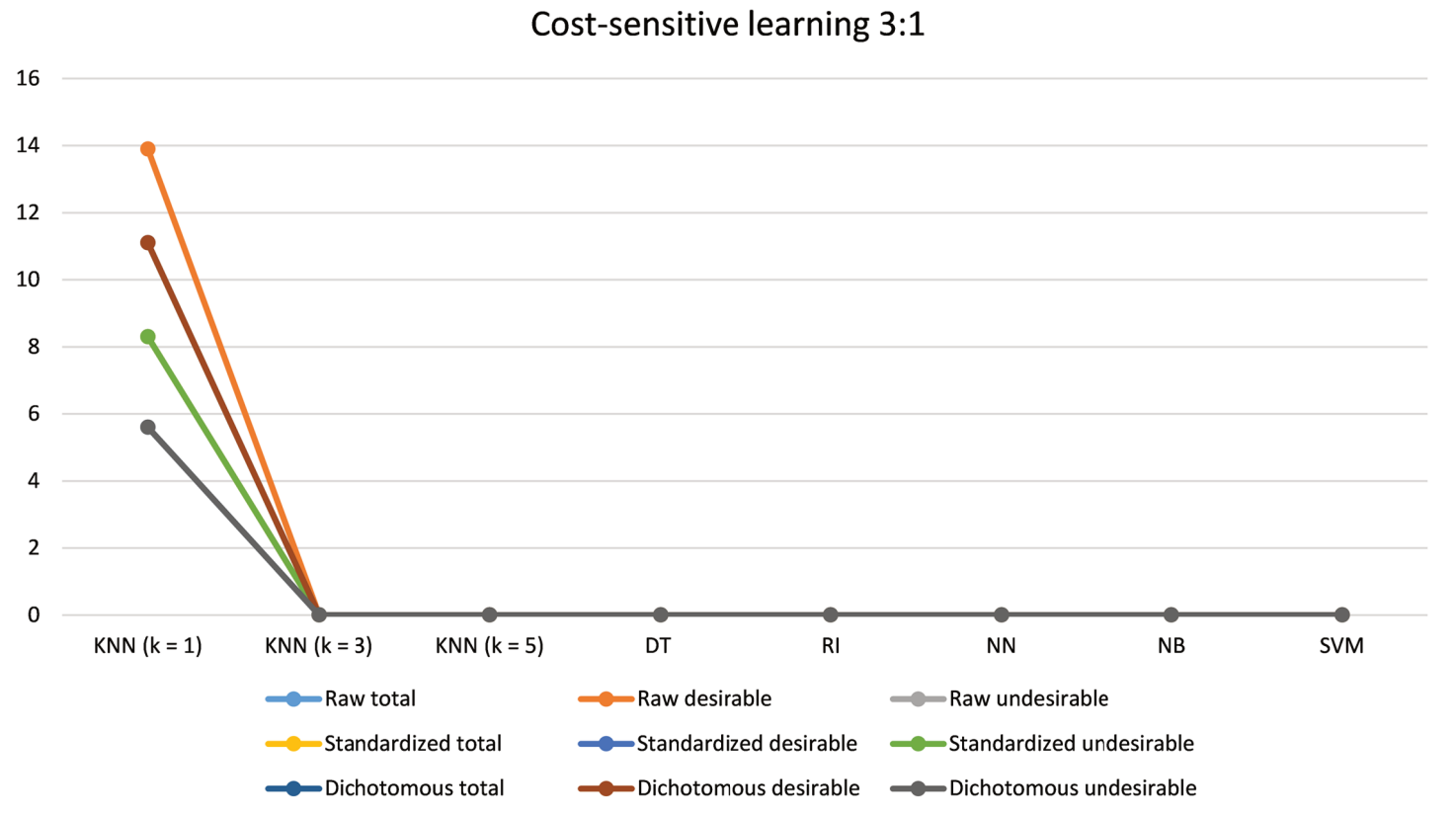

Figure 6D9. True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) for cost-sensitive learning $C_{\text{False Negative}} : C_{\text{False Positive}} = 3:1$.

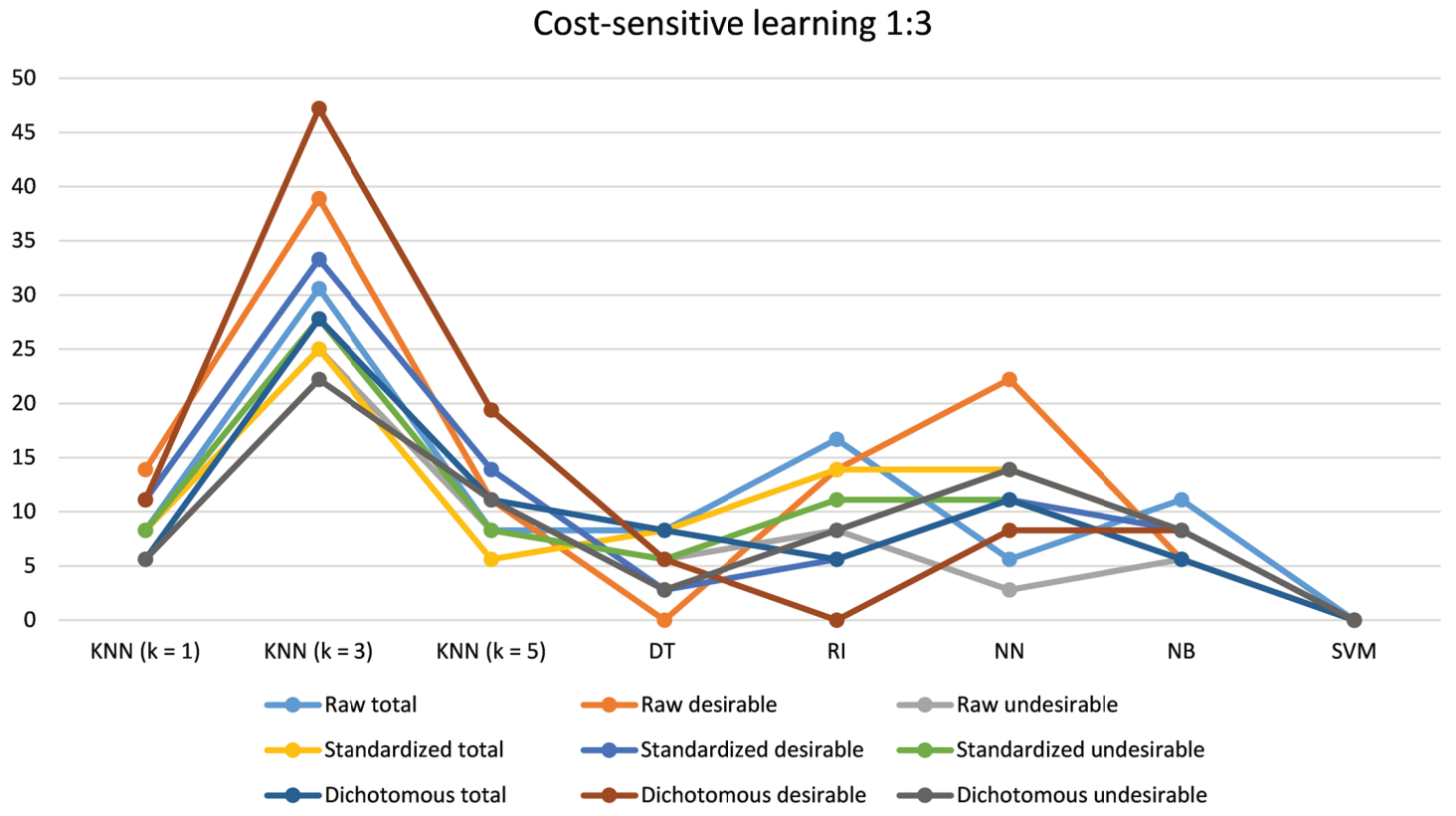

Figure 6D10. True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) for cost-sensitive learning $C_{\text{False Negative}} : C_{\text{False Positive}} = 1:3$.

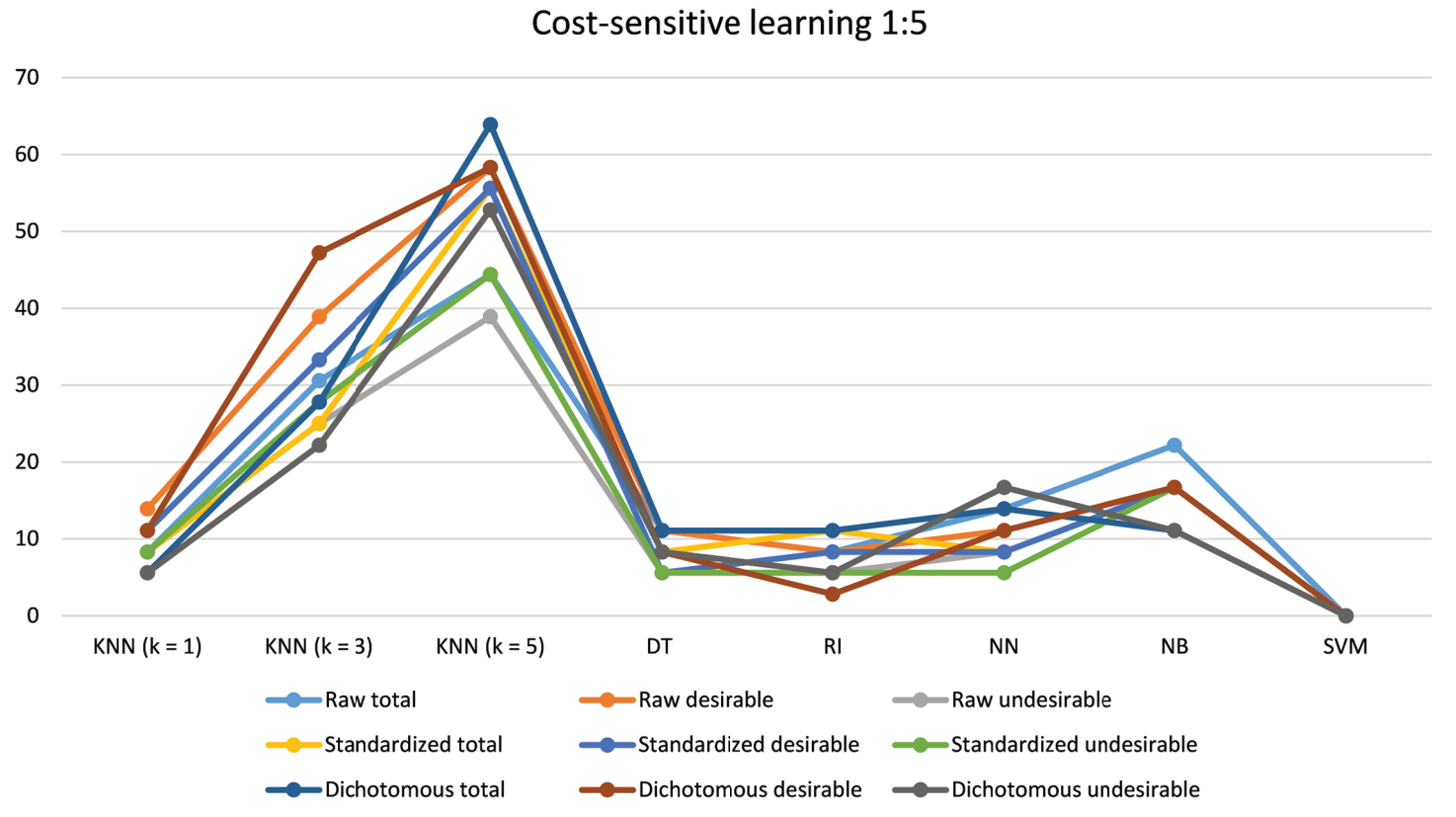

Figure 6D11. True positive rate (% TPR) for each machine learning algorithm (x-axis) and for each SJT version (separate lines) for cost-sensitive learning $C_{False\ Negative} : C_{False\ Positive} = 1:5$.

Summaries, Dankwoord, Publications and presentations, Curriculum Vitae, PhD Portfolio

Summary

More and more medical schools aim not only to select their students on traditional tests of knowledge and cognitive abilities, but also on tests of noncognitive attributes, such as integrity. The aim to measure noncognitive attributes among large groups of applicants in a standardised manner has led to the introduction of the Situational Judgement Test (SJT) in medical school admissions. An SJT consists of scenarios describing challenging situations in a relevant context (in this case: medical school), followed by a number of possible responses to those situations. The task of the applicant is to judge the appropriateness of these response options.

SJTs come in various shapes and forms and changing the characteristics of the test may alter the quality of an SJT. Previous research on SJTs in medical school admissions has paid limited attention to the influence of SJT characteristics on the quality of the test. However, the application of SJTs in high-stakes selection situations necessitates careful examination of which characteristics positively influence the quality of the test in order to improve the assessment of applicants' noncognitive attributes. Therefore, the goal of this thesis is to examine the influence of various test characteristics on a number of quality criteria of an SJT measuring integrity for medical school admissions. The General Introduction (**Chapter 1**) describes the findings of previous studies regarding the test characteristics and quality criteria of SJTs. The SJT characteristics included are the following: the development method, the response format, the scoring method, the response appropriateness and the response instructions. The test quality criteria involved are reliability, construct validity, criterion-related validity, subgroup differences (based on ethnic and socioeconomic background), fakability and applicant perceptions.

The study described in **Chapter 2** investigated an existing integrity-based SJT that was developed in Scotland. This SJT consisted of ten scenarios, each followed by five response options, that had to be judged on a four-point rating scale ranging from very inappropriate to very appropriate. The Scottish SJT was translated to Dutch and administered to two cohorts of medical school applicants. Initially, we used the same scoring method that was applied to the original SJT. This scoring method was based on the ratings of a group of Subject Matter Experts (SMEs), in our case, a group of 13 individuals involved in the education and assessment of professionalism in medical school. In particular, according to the Scottish method, item scores were determined by the number of SMEs that endorsed the ratings scale points. However, a search of the SJT literature revealed that many different methods exist to quantify the similarity between an applicant's set of ratings and the scoring key. Therefore, the goal of the study in Chapter 2 was to examine the influence of the SJT characteristic, scoring method, on three quality criteria: internal consistency reliability, subgroup differences between applicants of a Dutch and non-Western background and correlation to personality (that is, construct validity). By altering four scoring method aspects, i.e. the way of controlling for systematic error, the type of reference group, the type of distance and the type of central tendency statistic, twenty-eight scoring methods were formulated. The results of Chapter 2 indicated that the type of scoring method used has a strong influence on internal consistency reliability and subgroup differences. The way that a scoring method controls for systematic error had the largest influence on both quality criteria. This systematic error is often caused by individual differences in rating scale use (for example, only use the extremes or the middle of rating scales). These results suggest that rather than arbitrarily choosing one out of many methods to score an SJT, researchers and practitioners should thoroughly examine which scoring method leads to the highest test quality before using an SJT for high-

stakes selection. Additionally, the quality criteria of the test were comparable irrespective of the reference group (i.e. the individuals of whom the ratings formed the basis for the scoring key) used, suggesting that the group of applicants themselves might comprise an equally useful reference group as a group of SMEs.

The study in **Chapter 3** describes the development of an SJT tailored to the context of the Erasmus MC Medical School, since the Scottish SJT received some critique regarding its applicability to the Dutch context. The goal of this study was to develop a construct-based SJT that has realistic content to measure applicants' integrity. This was done by combining an empirical, inductive development approach with a theoretical, deductive development approach. The inductive approach involved critical incident interviews with SMEs (i.e. individuals involved in the assessment of medical students' professional behaviour) and the input of medical students and staff to develop realistic scenarios and response options. The deductive approach based the response options of the SJT on two integrity-related theoretical models. In particular, three facets of the HEXACO personality dimension honesty-humility (positively related to integrity) were used to write desirable response options. In addition, four categories of cognitive distortions, i.e. inaccurate thinking styles that may lead to antisocial behaviour (negatively related to integrity), were used to write undesirable response options. The resulting Integrity SJT consisted of 57 scenarios, each followed by four response options that had to be judged on a six-point rating scale from very inappropriate to very appropriate. The influence of the SJT characteristics, development method and response appropriateness, on the construct validity was examined in Chapter 3. The Integrity SJT had significant and medium-sized correlations with four external integrity-related measures, indicating convergent validity. Additionally, the Integrity SJT was not correlated to a measure of self-efficacy, indicating discriminant validity. Interestingly, the convergent validity was stronger for an SJT score based on undesirable response options (based on cognitive distortions) than for an SJT score based on desirable response options (based on honesty-humility). Higher consensus on what is considered inappropriate than appropriate in challenging situations might explain this difference in construct validity. The results of Chapter 3 indicate that established theoretical models provide useful guidance in the development of a construct-based SJT and that it may be promising to focus SJT development on the ability to correctly identify undesirable response options as inappropriate (i.e. what *not* to do).

The study in **Chapter 4** addresses a quality criteria of the SJT that is frequently called into question in high-stakes selection situations, that is, the susceptibility to faking. Faking is defined as the deliberate distortion of responses in order to make a better impression and increase the chances of getting admitted. Faking was investigated by administering the same ten SJT scenarios to the same applicants twice. The first administration (T1) was in a low-stakes situation, namely during a voluntary coaching day where applicants received information on the admission procedure and no selection took place. The second administration (T2) was in a high-stakes selection testing day situation, where applicants were administered three cognitive admission tests on which they were selected. Due to this difference in stakes, it was expected that applicants were more motivated to fake during the second administration. Faking was operationalised as a T1-T2 increase in the use of extreme rating scale points and in the SJT score. The SJT characteristics examined in Chapter 4 were the scoring method and response appropriateness. Three scoring methods were used that differed in the way of controlling for systematic error caused by individual differences in response tendencies (for example, only use the extremes or the middle of rating scales). The results of this study demonstrated that applicants used more extreme rating scale points on T2 than on T1, indicating that applicants respond differently to an SJT when the stakes are

higher. Whether a T1-T2 increase in extreme responding is associated with a T1-T2 increase in SJT score depends on the method used for scoring the SJT. An SJT score that controlled for systematic error increased from T1 to T2 with an increase in extreme responding. In contrast, an SJT score *not* controlling for response tendencies decreased from T1 to T2 with an increase in extreme responding. These findings suggest that a faking effect might be obscured when scoring an SJT using a method that does not control for systematic error caused by individual differences in response tendencies. Additionally, a stronger faking effect was found for an SJT score based on desirable response options than for an SJT score based on undesirable response options. This result indicates that the capacity to recognise what one should *not* to do in challenging situations might be harder to fake than the capacity to recognise what one should do in challenging situations.

The study described in **Chapter 5** shifts from the perspective of the admission committee to the perspective of the applicant, by examining the quality criteria applicant perceptions. A voluntary online survey was administered after the selection testing day but before applicants received the admission decision. The survey asked applicants about their perceptions of eleven admission methods, including an SJT. Applicant perceptions were assessed using an overall favourability score and five items concerning factors that may affect the perceived favourability of admission methods. The online survey included two example SJT items that varied in both their response instructions – either asking respondents what they should do or would do in the presented scenarios – and response format – either asking respondents to pick one of the response options or separately rate each response option using a rating scale. Manipulating these two SJT characteristics, response instructions and response format, resulted in four different SJT versions that were randomly assigned to the respondents. The results indicated that rating formats received more positive applicant perceptions than pick-one formats and would-do instructions were perceived as easier to cheat than should-do instructions. In addition, subgroup differences in applicant perceptions of an SJT were investigated and demonstrated no significant differences between subgroups based on gender, ethnic and socioeconomic background. However, significant interactions between the demographic subgroup variables and the SJT characteristics suggest that subgroups might differ in their preference for certain SJT characteristics. For instance, applicants from a low socioeconomic background, but not applicants from a high socioeconomic background, had more positive perceptions of rating formats than of pick-one formats.

The study in **Chapter 6** focuses on the ability of the Integrity SJT to predict future unprofessional behaviour in first-year medical students. A methodological difficulty in the prediction of unprofessional behaviour is the low prevalence of unprofessional behaviour among medical students (in our case: 8.8%), a difficulty termed the base rate problem. Although the occurrence of unprofessional behaviour is rare, professional lapses of medical students and doctors may have serious consequences for individuals dependent on their care and collaboration. In order to examine the predictive validity of the Integrity SJT in relation to this rare outcome, the study in Chapter 6 addresses the base rate problem using novel techniques from the field of machine learning. In machine learning, a computer is presented with a set of training data of which it automatically "learns" the underlying patterns, which can be captured using different algorithms. Machine learning is often applied to imbalanced datasets (for example, in the prediction of rare diseases) and, therefore, provides various methods to address the base rate problem. Six algorithms and three approaches to address the base rate problem were applied to classify unprofessionalism in first-year medical students based on several cognitive and noncognitive admission variables. These variables included pre-university grades, extracurricular activities, cognitive test scores, personality test scores, SJT score and voluntary participation in a coaching day. The base rate problem in our dataset

was reflected by low classification accuracy for the group of unprofessional students (cases) and high classification accuracy for the group of professional students (controls). The most effective method to increase the classification accuracy for the cases was the synthetic minority oversampling technique (SMOTE). This technique increases the number of cases based on the characteristics of the existing cases. This finding implies that the controlled oversampling of cases appears to be a more effective solution to the base rate problem than the extensive removal of controls. Evaluation of the contribution of each admission variable to the classification of unprofessional behaviour demonstrated that the voluntary participation in the coaching day was the most valuable variable in the classification of the outcome variable, suggesting that behavioural indicators may be the best predictors of future (unprofessional) behaviour. Regarding the Integrity SJT, the SJT characteristics that were examined in this study were the scoring method and response appropriateness. No substantial differences between the scoring methods in the classification accuracy were found. An SJT score based on desirable response options appeared more valuable in the classification of future unprofessionalism than an SJT score based on undesirable response options. This is possibly explained by the focus of the outcome variable on professional behaviours as opposed to unprofessional behaviours.

The General Discussion (**Chapter 7**) elaborates on three main themes of this thesis. First, it is discussed how profile similarity metrics and polynomial regression analysis may further improve the understanding of SJT scoring methods. Second, different categorisation systems for the different types of SJT response options are considered. Third, a deliberation of methodological and ethical issues that may complicate the application of an integrity-based SJT for medical school admissions is provided. The general discussion concludes with a review of the limitations of this thesis, including directions for future research.

Summary (in Dutch)

Steeds meer geneeskundeopleidingen lijken hun studenten niet alleen te willen selecteren op traditionele tests gericht op kennis en cognitieve vaardigheden, maar ook op tests gericht op niet-cognitieve eigenschappen, zoals integriteit. De doelstelling om niet-cognitieve kenmerken bij grote groepen kandidaten op een gestandaardiseerde manier te meten heeft geleid tot de invoering van situationele beoordelingstests (in het Engels: *Situational Judgement Tests*; SJTs) in toelatingsprocedures voor de geneeskundeopleiding. Een SJT bestaat uit scenario's die lastige situaties beschrijven in een relevante context (in dit geval: de geneeskundeopleiding), gevolgd door een aantal mogelijke reacties op die situaties. De taak van de kandidaat is om de gepastheid van deze responsopties te beoordelen.

SJTs komen voor in verschillende vormen en maten en het aanpassen van de kenmerken van de test beïnvloedt mogelijk de kwaliteit van een SJT. Eerder onderzoek naar SJTs binnen toelatingsprocedures tot de geneeskundeopleiding heeft weinig aandacht besteed aan de invloed van SJT-kenmerken op de kwaliteit van de test. Het gebruik van SJTs in selectiesituaties, waarbij er voor kandidaten veel op het spel staat (in het Engels: *high-stakes situaties*), vereist echter zorgvuldig onderzoek naar welke testkenmerken de kwaliteit van de SJT positief beïnvloeden om zo de meting van niet-cognitieve eigenschappen van kandidaten te verbeteren. Het doel van dit proefschrift is daarom het onderzoeken van de invloed van verschillende testkenmerken op een aantal kwaliteitscriteria van een SJT, gericht op het meten van integriteit voor de toelating tot de geneeskundestudie. De Algemene Inleiding (**Hoofdstuk 1**) beschrijft de bevindingen van eerdere studies met betrekking tot de testkenmerken en kwaliteitscriteria van SJTs. De volgende SJT-kenmerken komen achtereenvolgens aan bod: de ontwikkelmethode, het responsformat, de scoringsmethode, de geschiktheid van de responsopties en de responsinstructies. De kwaliteitscriteria betreffen de interne consistentiebetrouwbaarheid, de constructvaliditeit, de criteriumgerelateerde validiteit, subgroepverschillen (op basis van etnische en socio-economische achtergrond), *faken* en kandidaatpercepties.

De studie beschreven in **Hoofdstuk 2** bestudeerde een bestaande integriteit-gebaseerde SJT die is ontwikkeld in Schotland. Deze SJT bestaat uit tien scenario's, elk gevolgd door vijf responsopties, die beoordeeld moeten worden op een vier-punt beoordelingsschaal lopend van zeer ongepast tot zeer gepast. De Schotse SJT is vertaald naar het Nederlands en werd afgenomen bij twee cohorten kandidaten voor de geneeskundeopleiding. In eerste instantie gebruikten wij dezelfde scoringsmethode als bij de oorspronkelijke SJT. Deze scoringsmethode is gebaseerd op de beoordelingen van een groep inhoudelijke experts (in het Engels: *Subject Matter Experts*; SME's), in ons geval een groep van 13 personen, die betrokken zijn bij het onderwijs en de beoordeling van professionaliteit in de geneeskundeopleiding. De score op een item wordt volgens de Schotse methode bepaald door het aantal SME's dat voor iedere beoordelingsschaalpunt kiest. Uit de SJT-literatuur blijkt echter dat er veel verschillende methoden bestaan om de overeenkomst tussen de beoordelingen van een kandidaat en de scoringssleutel te kwantificeren. Het doel van de studie in Hoofdstuk 2 was daarom om de invloed van het SJT-kenmerk scoringsmethode te onderzoeken op drie kwaliteitscriteria: de interne consistentiebetrouwbaarheid, subgroepverschillen tussen kandidaten met een Nederlandse en een niet-Westerse achtergrond en het verband met persoonlijkheid (dat wil zeggen, constructvaliditeit). Door het variëren van vier scoringsmethodeaspecten, namelijk de manier van controleren voor systematische fouten, het soort referentiegroep, het soort afstand en het soort centrale tendentiestatistiek, werden achtentwintig scoringsmethoden gevormd. De resultaten van Hoofdstuk 2 lieten zien dat de scoringsmethode een sterke invloed heeft op de interne

consistentiebetrouwbaarheid en op de scoreverschillen tussen subgroepen. De manier waarop een scoringsmethode controleert voor systematische fouten had de grootste invloed op beide kwaliteitscriteria. Deze systematische fouten worden veelal veroorzaakt door individuele verschillen in het gebruik van beoordelingsschalen (bijvoorbeeld het gebruik van alleen de uitersten of alleen het midden van beoordelingsschalen). Deze resultaten impliceren dat onderzoekers en testgebruikers, in plaats van het willekeurig kiezen van één van de vele methoden om een SJT te scoren, nauwkeurig moeten onderzoeken welke scoringsmethode leidt tot de hoogste testkwaliteit, voordat een SJT wordt gebruikt voor *high-stakes* selectie. Daarnaast waren de uitkomsten op de kwaliteitscriteria van de SJT vergelijkbaar ongeacht welke referentiegroep (dat wil zeggen, de individuen van wie de beoordelingen de basis vormen voor de scoringssleutel) werd gebruikt, wat erop wijst dat de groep kandidaten zelf een even bruikbare referentiegroep zou kunnen vormen als een groep SME's.

De studie in **Hoofdstuk 3** beschrijft de ontwikkeling van een SJT die is toegespitst op de context van de geneeskundeopleiding van het Erasmus MC, aangezien de Schotse SJT werd bekritiseerd met betrekking tot de toepasbaarheid in de Nederlandse context. Het doel van deze studie was het ontwikkelen van een constructgebaseerde SJT met een realistische inhoud voor het meten van de integriteit van kandidaten. Dit werd gedaan door een empirische, inductieve ontwikkelmethode te combineren met een theoretische, deductieve ontwikkelmethode. De inductieve methode bestond uit kritieke incidenten interviews met SME's (namelijk, personen die betrokken zijn bij de beoordeling van professioneel gedrag van geneeskundestudenten) en de input van geneeskundestudenten en medisch personeel voor de ontwikkeling van realistische scenario's en responsopties. De deductieve methode baseerde de responsopties van de SJT op twee integriteitgerelateerde theoretische modellen. Drie facetten van de HEXACO persoonlijkheidsdimensie Integriteit (in het Engels: *honesty-humility*; positief gerelateerd aan integriteit) werden gebruikt om geschikte responsopties te schrijven. Daarnaast werden vier categorieën cognitieve verstoringen, dat wil zeggen, onjuiste gedachtepatronen die kunnen leiden tot antisociaal gedrag (negatief gerelateerd aan integriteit), gebruikt om ongeschikte responsopties te schrijven. De resulterende Integriteit-SJT bestond uit 57 scenario's, elk gevolgd door vier responsopties die beoordeeld moesten worden op een zes-punt beoordelingsschaal van zeer ongepast tot zeer gepast. In Hoofdstuk 3 werd de invloed van de SJT-kenmerken, ontwikkelmethode en geschiktheid van de responsopties, op de constructvaliditeit onderzocht. De Integriteit-SJT had significante en middelgrote verbanden met vier externe integriteitgerelateerde vragenlijsten, duidend op convergente validiteit. Daarnaast werd geen verband gevonden met een vragenlijst over zelfeffectiviteit, wat duidt op discriminante validiteit. Opvallend was de sterkere convergente validiteit voor een SJT-score gebaseerd op ongeschikte responsopties (op basis van de cognitieve verstoringen) dan voor een SJT-score gebaseerd op geschikte responsopties (op basis van *honesty-humility*). Een grotere consensus over welke reacties als ongepast, in plaats van gepast, worden beschouwd in lastige situaties kan dit verschil in constructvaliditeit mogelijk verklaren. De resultaten van Hoofdstuk 3 geven aan dat bestaande theoretische modellen een nuttige leidraad kunnen bieden bij de ontwikkeling van een constructgebaseerde SJT en dat het gunstig kan zijn om bij de ontwikkeling van een SJT te focussen op de vaardigheid om ongeschikte responsopties correct te identificeren als ongepast (dat wil zeggen, wat *niet* te doen).

De studie in **Hoofdstuk 4** behandelt een kwaliteitscriterium van de SJT dat vaak wordt bekritiseerd in *high-stakes* selectiesituaties, namelijk de vatbaarheid voor *faken*. *Faken* wordt gedefinieerd als de opzettelijke verdraaiing van antwoorden om zo een betere indruk te maken en de kans op toelating te vergroten. *Faken* werd onderzocht door tweemaal dezelfde tien SJT-scenario's af te nemen bij dezelfde groep kandidaten. De eerste afname (T1) was in

een *low-stakes* situatie, namelijk tijdens een vrijwillige coachingsdag waarop kandidaten informatie ontvingen over de selectieprocedure en waar geen selectie plaatsvond. De tweede afname (T2) was in een *high-stakes* selectiesituatie, namelijk een toetsdag waarop kandidaten drie cognitieve toelatingstoetsen moesten maken waarop ze werden geselecteerd. Vanwege het verschil in *stakes* werd verwacht dat kandidaten meer gemotiveerd waren om te *faken* tijdens de tweede afname. *Faken* werd geoperationaliseerd als een toename in het gebruik van extreme beoordelingsschaalpunten en als een toename in de SJT-score tussen T1 en T2. De SJT-kenmerken die in Hoofdstuk 4 werden onderzocht zijn de scoringsmethode en de geschiktheid van de responsopties. Er werden drie scoringsmethoden gebruikt die verschilden in hun manier van controleren voor systematische fouten veroorzaakt door individuele verschillen in het gebruik van beoordelingsschalen (bijvoorbeeld het gebruik van alleen de uitersten of alleen het midden van beoordelingsschalen). De resultaten van deze studie toonden aan dat kandidaten meer extreme beoordelingsschaalpunten gebruikten op T2 dan op T1, wat impliceert dat kandidaten anders reageren op een SJT wanneer er meer op het spel staat. Of een T1-T2 toename in het gebruik van extreme beoordelingsschaalpunten is geassocieerd met een T1-T2 toename in de SJT-score bleek af te hangen van de methode die wordt gebruikt voor het scoren van de SJT. Een SJT-score die controleerde voor systematische fouten steeg van T1 naar T2 met een toename in het gebruik van extreme beoordelingsschaalpunten. Een SJT-score die *niet* controleerde voor individuele verschillen in het gebruik van beoordelingsschalen daalde daarentegen van T1 naar T2 met een toename in het gebruik van extreme beoordelingsschaalpunten. Deze bevindingen suggereren dat een *faking* effect gemaskeerd zou kunnen worden als een SJT gescoord wordt met een methode die niet controleert voor systematische fouten veroorzaakt door individuele verschillen in het gebruik van beoordelingsschalen. Daarnaast werd een sterker *faking* effect gevonden voor een SJT-score gebaseerd op geschikte responsopties dan voor een SJT-score gebaseerd op ongeschikte responsopties. Dit resultaat duidt erop dat de vaardigheid om te herkennen wat men niet moet doen in lastige situaties mogelijk moeilijker te *faken* is dan de vaardigheid om herkennen wat men wel moet doen in lastige situaties.

De studie beschreven in **Hoofdstuk 5** verschuift van het perspectief van de toelatingscommissie naar het perspectief van de kandidaat, door het kwaliteitscriterium kandidaatpercepties te bestuderen. Een vrijwillige, online enquête werd afgenomen na de selectie-toetsdag maar voordat de kandidaten de uitslag van de selectie ontvingen. De enquête vroeg kandidaten naar hun opinie over elf toelatingsmethoden, inclusief een SJT. Kandidaatpercepties werden geëvalueerd op basis van een overall waarderingsscore en vijf items met betrekking tot factoren die de kandidaatpercepties van toelatingsmethoden kunnen beïnvloeden. De online enquête bevatte twee voorbeeld SJT-items, die varieerden in zowel hun responsinstructies – kandidaten moesten beoordelen wat ze zouden moeten doen (*should do*) of wat ze werkelijk zouden doen in de gepresenteerde scenario's (*would do*) – als in hun responsformat – kandidaten moesten of één van de gegeven responsopties kiezen of ze moesten elke responsoptie apart beoordelen door middel van een beoordelingsschaal. Het manipuleren van deze twee SJT-kenmerken, responsinstructies en responsformat, resulteerde in vier verschillende SJT-versies, die willekeurig werden toegewezen aan de respondenten. De resultaten lieten zien dat beoordelingsformats positievere kandidaatpercepties ontvingen dan meerkeuzeformats en dat *would do* instructies als makkelijk te *faken* werden beschouwd dan sh*ould do* instructies. Daarnaast werden subgroepverschillen in kandidaatpercepties van een SJT onderzocht. Verschillen tussen subgroepen op basis van geslacht, etnische en socio-economische achtergrond bleken niet significant. Echter, er werden significante interactie-effecten aangetroffen tussen deze demografische subgroepvariabelen en de SJT-kenmerken, wat impliceert dat subgroepen mogelijk verschillen in hun voorkeur voor bepaalde SJT-

kenmerken. Kandidaten met een lage socio-economische achtergrond, maar niet kandidaten met een hoge socio-economische achtergrond, hadden bijvoorbeeld positievere percepties van beoordelingsformats dan van meerkeuzeformats.

De studie in **Hoofdstuk 6** is gericht op het vermogen van de Integriteit-SJT om toekomstig onprofessioneel gedrag van eerstejaars geneeskundestudenten te voorspellen. Een methodologisch probleem bij het voorspellen van onprofessioneel gedrag is de lage prevalentie van onprofessioneel gedrag bij geneeskundestudenten (in ons geval: 8.8%), ook wel het *base rate* probleem genoemd. Hoewel onprofessioneel gedrag zeldzaam is, kunnen moeilijkheden gerelateerd aan de professionaliteit van geneeskundestudenten en artsen ernstige gevolgen hebben voor individuen die afhankelijk zijn van hun zorg en samenwerking. Om de predictieve validiteit van de Integriteit-SJT met betrekking tot deze zeldzame uitkomstmaat te onderzoeken is in Hoofdstuk 6 het base rate probleem benaderd met behulp van nieuwe technieken uit het domein van *machine learning*. In machine learning wordt een set training-data aangeboden aan een computer, waaruit het automatisch de onderliggende patronen "leert". Deze patronen kunnen door verschillende algoritmen worden weergegeven. Machine learning wordt vaak toegepast op ongebalanceerde datasets (bijvoorbeeld bij het voorspellen van zeldzame ziektes) en biedt daarom verschillende methoden om het base rate probleem aan te pakken. Zes algoritmen en drie methoden gericht op het base rate probleem werden toegepast op de classificatie van onprofessioneel gedrag bij eerstejaars geneeskundestudenten op basis van verschillende cognitieve en niet-cognitieve selectievariabelen. Deze variabelen omvatten vwo-cijfers, extracurriculaire activiteiten, cognitieve toets-scores, persoonlijkheidstestscores, SJT-score en de vrijwillige deelname aan een coachingsdag. Het base rate probleem werd in onze dataset weerspiegeld door een lage classificatienauwkeurigheid voor de groep niet-professionele studenten (de zogeheten *cases*) en een hoge classificatienauwkeurigheid voor de groep professionele studenten (de zogeheten *controls*). De meest effectieve methode om de classificatienauwkeurigheid voor de *cases* te verhogen, was de *synthetic minority oversampling technique* (SMOTE). Deze techniek vergroot het aantal *cases* op basis van de kenmerken van bestaande *cases*. Deze bevinding impliceert dat de gecontroleerde *oversampling* van het aantal *cases* een effectievere oplossing voor het *base rate* probleem lijkt te zijn dan de verwijdering van een groot aantal *controls*. De evaluatie van de bijdrage van elke selectievariabele aan de classificatie van onprofessioneel gedrag toonde aan dat de vrijwillige deelname aan de coachingsdag de meest waardevolle variabele was in de classificatie van de uitkomstmaat, wat suggereert dat gedragsindicatoren mogelijk de beste voorspellers zijn van toekomstig (onprofessioneel) gedrag. Met betrekking tot de Integriteit-SJT werden de SJT-kenmerken, scoringsmethode en geschiktheid van de responsopties, onderzocht. Er werden geen substantiële verschillen gevonden tussen de scoringsmethoden in hun effect op de classificatienauwkeurigheid. Een SJT-score gebaseerd op geschikte responsopties bleek waardevoller in de classificatie van toekomstig onprofessioneel gedrag dan een SJT-score gebaseerd op ongeschikte responsopties. Dit wordt mogelijk verklaard door de focus van de uitkomstmaat op professioneel gedrag in plaats van onprofessioneel gedrag.

De Algemene Discussie (**Hoofdstuk 7**) gaat dieper in op drie hoofdthema's van dit proefschrift. Allereerst wordt besproken hoe *profile similarity metrics* en polynome regressieanalyse SJT-scoringsmethoden mogelijk verder kunnen doorgronden. Ten tweede worden verschillende categorie-indelingen voor de verschillende soorten SJT-responsopties overwogen. Ten derde volgt een bespreking over methodologische en ethische kwesties die het gebruik van een integriteit-gebaseerde SJT voor de selectie van geneeskundestudenten mogelijk bemoeilijken. De algemene discussie wordt afgesloten met een overzicht van de beperkingen van dit proefschrift, inclusief suggesties voor toekomstig onderzoek.

Dankwoord

De afgelopen jaren waren rijk aan nieuwe ervaringen, kennis en vriendschappen. Ik heb een hoop mensen ontmoet zonder wie de voltooiing van mijn proefschrift nooit was gelukt. Anderen waren aanwezig lang voor de start van mijn promotieonderzoek en zijn met hun steun nooit van mijn zijde geweken. Mijn waardering voor jullie is groot, waarvoor hier een woord van dank.

Allereerst natuurlijk mijn promotieteam. Beste Axel, hartelijk dank voor je begeleiding, die ik als zeer prettig heb ervaren. Jouw kritische feedback maar ook praktische instelling hebben mij geholpen om zoveel mogelijk uit mijn promotietijd te halen en om verschillende papers te publiceren. Daarnaast vond ik het erg fijn dat ik af en toe met jou mee terug kon rijden naar Dordt. Ik ben blij dat je mij bleef begeleiden tijdens jouw emeritaat, waar je nu na mijn promotie eindelijk vol van kan genieten samen met Els.
Beste Marise, ik vond het heel fijn dat jij vanuit de psychologie was betrokken bij mijn promotietraject. Als een soort levende zoekmachine wist jij op elk vraagstuk wel een relevante studie of theoretisch model te bedenken. Bovendien konden jouw goed getimede complimenten mij uit ieder promotiedipje halen. Dankzij jouw steun en kennis ging ik na onze afspraken altijd weer met nieuwe ideeën en goede moed verder aan mijn onderzoek.
Beste Karen, bedankt voor je fijne dagelijkse begeleiding. Ik heb ontzettend veel van je geleerd. Jij wist altijd de juiste vragen te stellen en op een slimme en creatieve manier problemen op te lossen waar ik tijdens mijn promotietraject tegenaan liep. Jouw passie en enthousiasme voor onderzoek en goede selectie van geneeskundestudenten werken zeer aanstekelijk. Ik heb genoten van onze congresbezoeken waar we elkaar goed hebben leren kennen. Ik ben blij dat jij mijn copromotor was en dat we onze samenwerking kunnen voorzetten in het CLI project.
Met zijn vieren vormden we een goed team, waar dit proefschrift een mooi resultaat van is.

Vervolgens wil ik prof.dr. Maarten Frens, prof.dr. Dimitri van der Linden en dr. Janneke Oostrom bedanken voor hun bereidheid om plaats te nemen in de beoordelingscommissie en voor de tijd en moeite die zij hebben gestoken in het lezen en beoordelen van het manuscript. Daarnaast wil ik prof.dr. Rob Meijer, prof.dr. Walther van Mook en dr. Kitty Cleutjens bedanken dat zij willen opponeren tijdens mijn verdediging. Jullie onderzoeken en papers waren een grote inspiratiebron tijdens mijn promotie. Ik ben daarom blij dat jullie zes deel uitmaken van de verdediging van dit proefschrift.

Dear dr. Adrian Husbands, thank you for so kindly providing the SJT that you developed in Scotland, which really helped me to quickly start up my PhD research. Dear prof.dr. Jon Dowell, thank you for your co-authorship and feedback on my first paper.
Dear prof.dr. Geoff Norman, thank for your feedback on my research during your visits at the Erasmus MC. Your valuable remarks inspired us to write the paper on scoring methods.

Aansluitend wil ik alle personen bedanken die betrokken zijn geweest bij de ontwikkeling van de SJT. Ada, Benno, Gert, Hannie, Jenny, Leen, Liesbeth, Noor en Peter, bedankt voor jullie deelname aan de interviews over onprofessioneel gedrag binnen de geneeskundeopleiding. Jullie waardevolle bijdrage was onmisbaar in de ontwikkeling van de SJT en de totstandkoming van dit proefschrift. Emely, ik wil je bedanken voor de keren dat ik mijn promotieonderzoek mocht presenteren tijdens de werkgroep professionele

ontwikkeling. De feedback vanuit de werkgroep heeft mij geholpen om altijd kritisch te blijven kijken naar mijn onderzoek. Daarnaast ben ik de onderzoekers, docenten, stafleden en studenten uit de honours class die hebben geholpen bij het bedenken van de responsopties zeer dankbaar. Verder wil ik alle decentrale selectie kandidaten die hebben deelgenomen aan het onderzoek bedanken. De passie en ambitie die jullie tijdens de decentrale selectie lieten zien heeft mij altijd geïntrigeerd en benadrukt voor mij het belang van een goede selectieprocedure. Tot slot mijn grote dank voor Denise, Marian en Felicia. De dataverzameling voor mijn onderzoek liep dankzij jullie nauwkeurigheid en harde werken iedere keer weer gesmeerd.

Anouk, Marieke, Nienke en Sanne, het is altijd gezellig en leerzaam om met jullie op congressen en tijdens bijeenkomsten van gedachten te kunnen wisselen over onze onderzoeken naar de selectie. Susan, ik ben blij dat wij via Marise 'gekoppeld' zijn. Ik heb veel geleerd van jouw onderzoek en ik ben dankbaar dat ik via jou bij de COTAN terecht ben gekomen.

De onderzoekers van de Dutch-Flemish Network for Selection Research en de promovendi van de pubgroep op de Erasmus Universiteit wil ik bedanken voor hun waardevolle input op mijn onderzoek.

De collega's van de afdeling hoger onderwijs van de Inspectie van het Onderwijs, en in het bijzonder van de projectgroep Selectie en Toegankelijkheid, Susanne, Perry, Boy, Gerard, Willem en JW, wil ik bedanken voor een fantastische detacheringsperiode, waar ik regelmatig met veel plezier op terugkijk.

Mijn collega's van het NIP wil ik bedanken voor hun collegialiteit en alle lekkere traktaties. Ik heb in mijn leven nog nooit zoveel taart en koekjes gegeten als in het afgelopen jaar. Jennifer, ik hoop dat we samen nog vele pauzerondjes mogen lopen en dat het ons ooit lukt om Arlette te verslaan met tafelvoetbal. Marion, Marjolein, Nathaly, Nicole, Renate, Rosalinde en Wilma, dankzij jullie voelde ik mij in no-time thuis op het NIP. Laten we snel weer een keer gaan bowlen!

Iris en Karin, ik ben ongelofelijk blij dat jullie mijn COTAN-collega's zijn. Door jullie behulpzaamheid en vriendelijkheid voelde ik mij gelijk welkom. Karin, jouw vrolijkheid maakt werken bij de COTAN extra leuk. Ik vind het altijd erg gezellig om met jou terug te reizen na het werk. Iris, het afgelopen jaar heb ik veel geleerd van jouw ervaring met testbeoordelingen. Het was ook fijn dat ik met jou zaken rondom de afronding van mijn promotie kon bespreken, waar je altijd een goede raad op wist.
Petra en de andere COTAN-leden wil ik bedanken voor hun interessante en gepassioneerde discussies, waar ik iedere keer weer ontzettend veel van opsteek.

Mijn (oud)collega's op het Erasmus MC wil ik bedanken voor hun samenwerking en hun tolerantie voor mijn rare, niet altijd geluidloze, fratsen. Door jullie ging ik elke dag met plezier naar mijn werk. Ik ben daarom blij dat ik mijn onderzoek parttime kan voortzetten. Mijn dank voor ieder van jullie is groot, maar om de lengte van dit dankwoord binnen de perken te houden, lukt het helaas niet om iedereen persoonlijk te bedanken. Rianne, het is fantastisch om een collega te hebben die net zo enthousiast – zo niet enthousiaster – is over escape rooms en Wie is de mol. Marja, bedankt dat ik af en toe je hoelahoep mocht lenen. Miranda, bedankt dat ik naast jouw bureau mocht hoelahoepen en jongleren, terwijl jij

stoïcijns bleef doorwerken. Rita, ik hoop dat je snel herstelt, want het is stil zonder jou. Priscilla, bedankt voor je onuitputtelijke enthousiasme in het schrijven van de nieuwsbrief en het organiseren van activiteiten op de afdeling. Jolanda, bedankt voor al onze gezellige gesprekken tijdens de wekelijkse vrijdagmiddag sensation. Eric, merci de pratiquer mon français avec vous. Mathijs, bedankt voor je hulp met het ontcijferen van Scorion en EPASS. Tot slot wil ik graag de saladeclub bedanken voor alle lekkere en gezonde middagmalen.

Mijn collega's van iMERR, Laura, Sílvia, Mary en Walter, wil ik bedanken voor hun wijze raad tijdens mijn onderzoek. Het was fijn om van jullie ervaring te leren. De feestkamer- en ballententbewoners, dankzij jullie was promoveren een feestje. Josepha, met jou kan ik altijd praten over mijn promotie-struggles, maar ook urenlang over andere dingen onder het genot van een drankje. Tjitske, bedankt voor je adviezen en tips, die grote problemen minder groot maken, en voor je oneindige gastvrijheid. Chantal, bedankt voor onze gezellige gesprekken over onderzoek en andere, meer geurige, onderwerpen. Justine, wij hebben maar even tegenover elkaar gezeten, maar je humor heeft een blijvende indruk achter gelaten. Lokke, één van de allerleukste dingen aan promoveren was dat ik het tegelijkertijd met jou kon doen. Samen hebben we hinkelbanen aangelegd, gastoeters onder bureaustoelen geïnstalleerd en heel veel kamers versierd. Je stond altijd klaar om te helpen met het verzamelen van data en een bemoedigende blik tijdens presentaties. Duizendmaal dank! Inge, Suzanne en Vera, ik ben blij dat iMERR met jullie drie nieuwe, leuke en intelligente promovendi er bij heeft. Ik wens jullie allemaal heel veel succes maar vooral ook veel plezier tijdens de rest van jullie promotietrajecten.

Susanne en Frederique, jullie waren er vanaf het begin van mijn promotie bij. Dankzij jullie had ik een goede start en heb ik geleerd dat werk en plezier prima samengaan. Het is altijd fijn om met jullie bij te praten.

Anne, het is al weer even geleden, maar bedankt voor je prettige en leuke begeleiding tijdens mijn masteronderzoek, die mijn enthousiasme voor onderzoek heeft aangewakkerd.

Kim en Vivian, dankzij jullie werd mijn studietijd in Tilburg een onvergetelijke tijd. Ik ben blij dat wij elkaar nog steeds regelmatig zien en ik ben benieuwd waar onze volgende reis naar toe zal gaan.

Sally en Ariëlle, ik ben dankbaar dat jullie mijn paranimfen zijn. Sally, bedankt voor je levenslange vriendschap, die mij veel waard is. Samen hebben we een hoop gezamenlijke herinneringen. Laten we er nog veel meer bij maken! Ariëlle, jouw trotsheid bij elke mijlpaal tijdens mijn promotie deed mij altijd goed. Je bent een goede vriendin en bovendien, samen met René, maker van twee fantastische kinderen. Lieve Nura en Ilaria, onze playdates waren een fijne afwisseling van het werk en ik hoop dat er nog vele zullen volgen.

Rick van Driel, bedankt dat je de omslag voor mijn proefschrift wilde ontwerpen.

Ed, Irma, Floris, Fabian, Patricia, Rowan, Eltica, kleine Nikkie, Ruud, Eliane, Ilse, Stefan, Indy, Mees, Jannie, Arjan, Daniëlle, Sascha, Marvin, Kees, Sjak, Kevin, Jeff, grote Nikkie, Henny, Adriaan en Tim, bedankt voor jullie warmte en vriendschap. Jullie maken het leven mooier.

Opa, het is elke vrijdag feest als ik jou weer zie. Bedankt dat ik een week bij jou mocht bivakkeren om aan mijn proefschrift te werken. Ondanks dat je af en toe zit te klieren, ben je een ontzettend lieve opa.

Lieve Mamsel en Papa, bedankt voor jullie onvoorwaardelijke liefde en steun. Hoe zwaar een week soms ook was, als ik hem vrijdag bij jullie kon afsluiten met een wijntje, was alles weer goed. Johnno en Steven, jullie zijn twee hele fijne grote, kleine broertjes. Bedankt voor jullie gezellige gekkigheid. Jullie staan altijd voor mij klaar. Ik hou van jullie!

Jolly, bedankt voor je fluffy aanwezigheid, waarmee je in de letterlijke zin van het woord niet van mijn zijde bent geweken.

Lieve Ruud, er wordt wel eens gezegd dat het niet makkelijk is om een relatie te hebben met een promovendus en ik ben bang dat ik daar geen uitzondering op was. Bedankt voor al je knuffels en steun de afgelopen jaren. Ik kijk er naar uit om de weekenden meer met jou en minder met mijn onderzoek door te brengen.

Publications and presentations

Publications

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2019). Faking on a situational judgment test in a medical school selection setting: Effect of different scoring methods? *International Journal of Selection and Assessment, 27*, 235-248.

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2019). MUM effect in medical education: taking into account the recipient and training setting. *Medical Education, 53,* 106-108.

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2018). Influence of response instructions and response format on applicant perceptions of a situational judgement test for medical school selection. *BMC Medical Education, 18*, 282.

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2018). Integrity situational judgement test for medical school selection: Judging 'what to do' versus 'what not to do'. *Medical Education, 52*, 427-437.

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., Frens, M.A., & Themmen, A.P.N. (2017). Participation in a scientific pre-university programme and medical students' interest in an academic career. *BMC Medical Education, 17,* 150.

De Leng, W.E., Stegers-Jager, K.M., Husbands, A., Dowell, J.S., Born, M.Ph., & Themmen, A.P.N. (2017). Scoring method of a Situational Judgment Test: Influence on internal consistency reliability, adverse impact and correlation with personality? *Advances for Health Sciences Education, 22*, 243-265.

Oral presentations

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2017, November 17). *Faking op een Situational Judgement Test voor de selectie van geneeskunde studenten.* Paper presented at NVMO conference, Egmond aan zee, the Netherlands.

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N.(2017, October 20) *Faking on a Situational Judgement Test for medical school selection*. Paper presented at the annual meeting of the Dutch-Flemish Network for Selection Research, Ghent, Belgium.

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2017, August 31). *Faking on a Situational Judgement Test*. Paper presented at Rogano meeting, Helsinki, Finland.

De Leng, W.E., Stegers-Jager, K.M., Husbands, A., Dowell, J.S., Born, M.Ph., & Themmen, A.P.N. (2017, July 24). *Twenty-eight SJT scoring methods: influence on internal consistency reliability, adverse impact and correlation with personality*. Paper presented at DREAMS workshop, Sydney, Australia.

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2016, November 25). *Situational Judgment Test for selecting medical students on integrity: the "bright" and "dark" side.* Paper presented at WAOP conference, Rotterdam, the Netherlands.

De Leng, W.E., Stegers-Jager, K.M., Husbands, A., Born, M.Ph., & Themmen, A.P.N. (2015, November 12). *Het effect van scoringsmethode voor een Situational Judgement Test op de betrouwbaarheid en etnische subgroep verschillen.* Paper presented at NVMO conference, Rotterdam, the Netherlands.

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2015, October 16).

Situational Judgment Tests for selection into medical school: Scoring methods and hybrid development. Paper presented at the annual meeting of the Dutch-Flemish Network for Selection Research, Rotterdam, the Netherlands.

De Leng, W.E., Stegers-Jager, K.M., Husbands, A., Born, M.Ph., & Themmen, A.P.N. (2015, September 7*). The effect of the scoring method for a Situational Judgement Test on adverse impact and reliability.* Paper presented at AMEE conference, Glasgow, Scotland.

Poster presentations

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2016, November 17). *Hybride ontwikkeling en validering van een integriteit Situational Judgement Test voor de selectie van geneeskundestudenten*. Poster presented at NVMO conference, Egmond aan Zee, the Netherlands.

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2017, August 28). *Development and construct validity of an Integrity SJT for medical school selection: judging 'what to do' versus 'what not to do'.* Paper presented at AMEE conference, Helsinki, Finland.

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. (2016, March 21). *Hybrid development of an integrity-based Situational Judgment Test for selection into medical school.* Paper presented at Ottawa conference, Perth, Australia.

Symposia

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. *Influence of response instructions and response format on applicant perceptions of a Situational Judgement Test for medical school selection.* In: Schmid, E., Moldzio, T., De Leng, W.E., Kim, L.E., & Jörg, F. (2019, May 31). *Casting light on Situational Judgement Tests from multiple perspectives: Possibilities of differential uses across disciplines and countries.* EAWOP conference, Turin, Italy.

De Leng, W.E., Stegers-Jager, K.M., Born, M.Ph., & Themmen, A.P.N. *Development of a Situational Judgement Test to measure integrity for medical school selection.* In: De Leng, W.E., Stegers-Jager, K.M., Niessen, A.S.M., & Meijer, R.R. (2017, April 30). *Current developments in selective admission to higher education in Europe.* NCME conference, San Antonio, United States.

De Leng, W.E., Stegers-Jager, K.M., Lucieer, S.M., Frens, M.A., & Themmen, A.P.N. *Junior Med School: Evaluatie en opbrengsten.* In: Van der Valk, T., Tromp, S., De Leng, W.E., & Walsari Wolff, S. (2015, June 17). Opbrengsten van pre-universitaire verrijkingsprogramma's. ORD conference, Leiden, the Netherlands.

Curriculum Vitae

Wendy de Leng was born on 21 June 1989 in Dordrecht, the Netherlands. In 2007, she completed her secondary education at Insula College in Dordrecht. She obtained her Bachelor's degree in Psychology at Tilburg University in 2012 and obtained her Master's degree in Methodology and Statistics in Psychology at University Leiden in 2013. After graduating, she worked as a teaching assistant in statistics at VU Amsterdam and as a research assistant at University Leiden. In April 2014, she started her PhD research on the use of a situational judgement test for medical school selection at the institute of Medical Education Research Rotterdam (iMERR), Erasmus Medical Centre. This research resulted in various national and international presentations and publications. During her PhD research, she continued teaching in statistics at the Erasmus University and, in 2017, she worked for six months as a part-time researcher focused on selection and admissions at the Inspectorate of Education. As of October 2018, she works as an editor at the Dutch Committee on Tests and Testing (COTAN) and as a scientific researcher at iMERR.

PhD Portfolio

Summary of PhD training and teaching

Name PhD student: Wendy de Leng Erasmus MC Department: institute of Medical Education Research Rotterdam (iMERR)	PhD period: April 2014 – July 2018 Promotors: Prof. dr. ir. A.P.N. Themmen Prof. dr. M.Ph. Born Supervisor: Dr. K.M. Stegers-Jager

1. PhD training		
	Year	**Workload (Hours/ECTS)**
General courses		
- Biomedical English Writing and Communication	2015-2016	3 ECTS
- Research Integrity	2015	0.3 ECTS
- How to Survive your PhD	2014	2.5 ECTS
- Presenting and Networking	2014	2.5 ECTS
- Systematic Literature Research	2014	1 ECTS
Specific courses (e.g. Research school, Medical Training)		
- Introduction to Bayesian Methods in Clinical Research	2016	1.4 ECTS
Seminars and workshops		
- KNAW Hendrik Muller Summer Seminar	2016	2 ECTS
Presentations		
- Oral (8x)	2015-2017	8 ECTS
- Poster (2x)	2016	1 ECTS
- Symposium (3x)	2015-2019	3 ECTS
(Inter)national conferences		
- 7 international conferences	2014-2019	7 ECTS
- 12 national conferences	2014-2017	6 ECTS
2. Teaching		
	Year	**Workload (Hours/ECTS)**
Supervising practicals and excursions, Tutoring		
- Tutor Statistic II: Explaining and Predicting (2x)	2014-2015	6 ECTS
- Tutor Applied Multivariate Data Analysis	2017	3 ECTS
Other		
- Supervising Bachelor's thesis	2016	1 ECTS