Adv in Health Sci Educ (2017) 22:243–265
DOI 10.1007/s10459-016-9720-7

Scoring method of a Situational Judgment Test: influence on internal consistency reliability, adverse impact and correlation with personality?

W. E. De Leng[1] · K. M. Stegers-Jager[1] · A. Husbands[2] · J. S. Dowell[3] · M. Ph. Born[4] · A. P. N. Themmen[1,5]

Received: 12 February 2016 / Accepted: 6 October 2016 / Published online: 18 October 2016

Abstract Situational Judgment 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.

✉ W. E. De Leng
w.deleng@erasmusmc.nl

[1] Institute of Medical Education Research Rotterdam (iMERR), Erasmus MC, Room AE-239, PO Box 2040, 3000 CA Rotterdam, The Netherlands

[2] Medical School, University of Buckingham, Buckingham, UK

[3] School of Medicine, University of Dundee, Dundee, UK

[4] Department of Psychology, Erasmus University Rotterdam, Rotterdam, The Netherlands

[5] Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands

Keywords Situational Judgment Test · Scoring method · Medical school selection · Internal consistency reliability · Adverse impact · Integrity · Big Five

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 et al. 2002; Salvatori 2001). Adding non-cognitive-based measures to cognitive-based measures may improve the predictive quality of a selection procedure (Kulatunga-Moruzi and Norman 2002; Lucieer et al. 2015; Powis 2015). Non-cognitive-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 non-cognitive-based measure for selection into medical school is the Situational Judgment 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 and Ployhart 2013). Administering SJTs in work-related selection procedures has several beneficial characteristics: (1) good predictive validity with regard to job performance (McDaniel et al. 2001), (2) incremental validity over and above cognitive ability and personality (Clevenger et al. 2001), (3) less adverse impact than cognitive measures (McDaniel and Nguyen 2001), (4) higher favorability ratings by candidates than in cognitive tests (Lievens 2013) and (5) more efficient administration to large groups of applicants than other non-cognitive-based instruments (e.g., assessment centers) (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. 2005; Lievens and Sackett 2012; Patterson et al. 2009, 2011, 2015).

Despite the good qualities mentioned above, some aspects of SJTs require more research. One of these aspects is the scoring method (Whetzel and 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 judgments 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 judgments 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 et al. 2013).

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, standardized, 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. *Standardized* consensus calculates the distance after conducting a within-person z standardization 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 standardizing or dichotomizing 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 and O'Malley 1984; Hui and Triandis 1989). By standardizing or dichotomizing 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 standardized 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 standardized 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 et al. (2010) found a higher concurrent validity for a standardized scale than a raw scale.

Next to using raw, standardized 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 and 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 judgments 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 behavior is effective and ineffective in their field (Motowidlo and 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 and 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 and 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 et al. (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 panelists (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 judgments of the reference group are summarized (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 and 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 and 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 analyzing 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., standardized 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 and MacCrae 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 et al. 2007). Moreover, the validity and reliability of the scores on the Big Five measure used in this study [i.e., NEO-PI-R (Costa and MacCrae 1992)] has repeatedly been demonstrated (Costa and McCrae 2008), including in samples of adolescents (De Fruyt et al. 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 hypothesize 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 et al. 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 Judgment 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 and Ehrhart 2003)]. An example of an SJT item is presented in Appendix 1.

A rational scoring key for this SJT was developed based on the judgments 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 socio-economic 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 and MacCrae 1992; Hoekstra et al. 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 (Fig. 1). These scoring methods followed the categorization into raw, standardized and dichotomous consensus scoring methods as proposed by McDaniel et al. (2011).

Within each of the raw and standardized 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 standardized 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.

The 24 scoring methods based on either raw, standardized 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

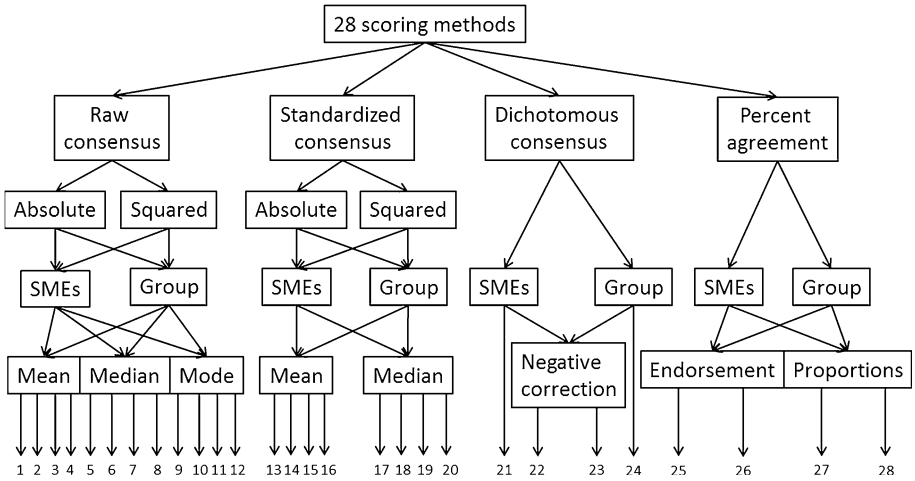

Fig. 1 Schematic representation of the 28 scoring methods. *SMEs* Subject Matter Experts

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 2. 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 judgments on the SJT to a score, using the different scoring methods. The raw and standardized consensus scoring methods that used the group of respondents itself as a reference were conducted using a leave-one-out method (Hastie et al. 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 et al. 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 [$X^2(1) = 6.02$, $p = .014$, $\varphi = .08$]. 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.

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 standardized 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,

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

SD standard deviation, *SES* socio-economic status

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

Scoring method	M (SD)	Min.–Max.	Alpha
Raw consensus			
1. Absolute distance—SME mean	34.32 (6.02)	20.01–64.99	.67
2. Absolute distance—SME median	33.11 (6.61)	13.50–66.50	.56
3. Absolute distance—SME mode	32.95 (6.52)	14.50–65.50	.55
4. Squared distance—SME mean	36.25 (12.50)	11.48–107.72	.67
5. Squared distance—SME median	42.44 (13.27)	12.75–122.75	.61
6. Squared distance—SME mode	41.81 (13.18)	13.25–121.25	.60
7. Absolute distance—Group mean	31.26 (6.31)	16.32–63.09	.73
8. Absolute distance—Group median	28.93 (7.00)	11–63	.61
9. Absolute distance—Group mode	29.07 (6.99)	11–63	.59
10. Squared distance—Group mean	30.35 (11.56)	8.47–100.35	.73
11. Squared distance—Group median	35.67 (12.85)	11–113	.65
12. Squared distance—Group mode	36.28 (13.01)	11–115	.63
Standardized consensus			
13. Absolute distance—SME mean	32.86 (4.63)	21.24–51.67	.44
14. Absolute distance—SME median	33.52 (4.68)	19.09–51.54	.41
15. Squared distance—SME mean	34.46 (9.57)	14.31–34.46	.49
16. Squared distance—SME median	36.29 (9.61)	13.47–79.99	.45
17. Absolute distance—Group mean	30.42 (3.91)	20.99–50.67	.33
18. Absolute distance—Group median	29.91 (4.57)	18.27–51.00	.43
19. Squared distance—Group mean	29.11 (7.77)	13.58–74.24	.45
20. Squared distance—Group median	30.08 (8.89)	12.63–79.44	.51
Dichotomous consensus			
21. SME as reference	34.34 (3.55)	21–44	.34
22. SME as reference—negative correction	18.78 (7.04)	−8–38	.34
23. Group as reference	37.56 (3.59)	22–47	.34
24. Group as reference—negative correction	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

M mean, *SD* standard deviation, *SME* Subject Matter Expert, *Min.* minimum, *Max.* maximum

standardized 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

Table 3 Results of the independent *t* tests for Dutch versus non-Western differences in SJT scores generated by the 28 different scoring methods

Scoringmethod	Dutch (N = 490)	Non-Western (N = 269)	*d*
Raw consensus			
1. Absolute distance—SME mean	32.88 (5.38)	36.51 (6.50)	0.61
2. Absolute distance—SME median	31.47 (5.92)	35.58 (7.05)	0.63
3. Absolute distance—SME mode	31.34 (5.82)	35.37 (6.95)	0.63
4. Squared distance—SME mean	33.28 (10.86)	40.82 (13.98)	0.60
5. Squared distance—SME median	39.24 (11.50)	47.39 (14.84)	0.61
6. Squared distance—SME mode	38.70 (11.37)	46.67 (14.74)	0.61
7. Absolute distance—Group mean	29.95 (5.61)	33.16 (7.03)	0.50
8. Absolute distance—Group median	27.29 (6.27)	31.31 (7.49)	0.58
9. Absolute distance—Group mode	27.37 (6.21)	31.51 (7.48)	0.60
10. Squared distance—Group mean	27.94 (9.88)	33.96 (13.37)	0.51
11. Squared distance—Group median	32.66 (11.06)	40.02 (14.35)	0.57
12. Squared distance—Group mode	33.13 (11.10)	40.86 (14.59)	0.60
Standardized consensus			
13. Absolute distance—SME mean	31.69 (4.23)	34.52 (4.43)	0.65
14. Absolute distance—SME median	32.30 (4.25)	35.22 (4.60)	0.66
15. Squared distance—SME mean	32.07 (8.52)	37.80 (9.36)	0.64
16. Squared distance—SME median	33.88 (8.51)	39.69 (9.56)	0.64
17. Absolute distance—Group mean	29.53 (3.63)	31.55 (3.72)	0.55
18. Absolute distance—Group median	28.83 (4.25)	31.30 (4.34)	0.58
19. Squared distance—Group mean	27.47 (7.14)	31.13 (7.40)	0.50
20. Squared distance—Group median	28.10 (8.11)	32.52 (8.58)	0.53
Dichotomous consensus			
21. SME as reference	35.07 (3.32)	33.43 (3.46)	0.48
22. SME as reference—negative correction	20.22 (6.59)	16.98 (6.86)	0.48
23. Group as reference	38.31 (3.37)	36.69 (3.44)	0.48
24. Group as reference—negative correction	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

All differences were significant ($p < .001$)

SME Subject Matter Expert, *d* Cohen's *d* (effect size)

that calculated the absolute distance from the SME median using standardized 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 standardized 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.

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 standardized 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

Table 4 Pearson correlation coefficients between the SJT score and the three Big Five personality dimensions for which we expect a correlation with the integrity-based SJT assessed by the NEO-PI-R in cohort 2014 only (N = 382)

Scoringmethod	N	A	C
Raw consensus			
1. Absolute distance—SME mean	−.03	**−.11**	−.04
2. Absolute distance—SME median	.01	**−.11**	−.07
3. Absolute distance—SME mode	0	**−.11**	−.06
4. Squared distance—SME mean	−.03	**−.12**	−.04
5. Squared distance—SME median	−.01	**−.12**	−.06
6. Squared distance—SME mode	−.01	**−.12**	−.05
7. Absolute distance—Group mean	−.06	−.07	.02
8. Absolute distance—Group median	−.06	−.08	0
9. Absolute distance—Group mode	−.03	−.08	0
10. Squared distance—Group mean	−.06	−.09	0
11. Squared distance—Group median	−.06	**−.11**	0
12. Squared distance—Group mode	−.05	**−.11**	−.01
Standardized consensus			
13. Absolute distance—SME mean	0	**−.15**	**−.12**
14. Absolute distance—SME median	−.01	**−.12**	**−.12**
15. Squared distance—SME mean	0	**−.15**	**−.10**
16. Squared distance—SME median	0	**−.13**	**−.11**
17. Absolute distance—Group mean	.01	−.10	−.07
18. Absolute distance—Group median	0	**−.11**	−.06
19. Squared distance—Group mean	.02	−.10	−.06
20. Squared distance—Group median	.01	**−.11**	−.06
Dichotomous consensus			
21. SME as reference	−.07	.07	.10
22. SME as reference—negative correction	−.07	.07	.10
23. Group as reference	.02	**.14**	.05
24. Group as reference—negative correction	.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

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

N neuroticism, *A* agreeableness, *C* conscientiousness, *SME* Subject Matter Expert

were found for scoring methods calculating the absolute distance from the SME mean and median both using standardized consensus. Due to the low effect sizes and the small range of significant correlation coefficients, no general linear models were tested.

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 standardized 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). Standardized 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 artifact 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 modeling (SEM) reliability estimate (Yang and 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 standardized 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., standardized 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 standardized 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 et al. 2013). Dichotomizing 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 and 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 behaviors 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 et al. 2012; Jha et al. 2015) and since integrity is an important aspect of medical professionalism, it too might depend on cultural context (Arnold and 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 standardized 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 et al. 2010). This non-association might be explained by the fact that personality tests assess non-cognitive traits, whereas MMIs and SJTs assess non-cognitive behaviors. Non-cognitive behaviors are more dependent on situational factors than personality traits (Eva 2005). This is in line with a previous study which demonstrated that a contextualized personality measure had higher criterion validity for academic performance and counterproductive academic behavior than a generic personality measure (Holtrop et al. 2014). The lack of contextualization 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 summarized using the mean. Since the mean is strongly influenced by extreme scores (Field 2013), asymmetrically distributed data are better summarized 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 judgments 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 generalizability 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 and 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.

Appendix 1: 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 2

See Table 5.

Table 5 Correlation between the SJT scores resulting from the 28 different scoring methods

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.	.67	.72	.74	.66	.69	.70	.50	.60	.63	.54	.63	.64		
14.	.70	.78	.79	.70	.73	.74	.50	.59	.62	.54	.61	.64	.98	
15.	.65	.70	.72	.68	.70	.72	.48	.58	.61	.55	.64	.66	.96	.94
16.	.65	.72	.74	.68	.72	.74	.47	.57	.60	.54	.62	.64	.95	.95
17.	.61	.64	.68	.61	.63	.66	.61	.70	.72	.63	.71	.72	.88	.84
18.	.60	.64	.67	.60	.62	.65	.60	.75	.76	.62	.74	.74	.87	.82
19.	.57	.62	.65	.60	.62	.65	.53	.64	.65	.61	.70	.70	.87	.84
20.	.58	.62	.66	.60	.62	.65	.53	.67	.69	.60	.73	.73	.87	.83
21.	−.58	−.62	−.61	−.57	−.58	−.58	−.31	−.37	−.39	−.34	−.39	−.41	−.74	−.82
22.	−.59	−.63	−.61	−.57	−.58	−.58	−.31	−.37	−.39	−.34	−.39	−.41	−.75	−.82
23.	−.51	−.52	−.56	−.51	−.51	−.54	−.49	−.63	−.63	−.49	−.61	−.61	−.75	−.71
24.	−.52	−.53	−.56	−.51	−.51	−.54	−.50	−.63	−.63	−.50	−.61	−.61	−.75	−.71
25.	−.88	−.91	−.93	−.84	−.84	−.85	−.75	−.77	−.80	−.74	−.76	−.77	−.75	−.78
26.	−.73	−.74	−.77	−.72	−.72	−.75	−.73	−.80	−.80	−.75	−.79	−.79	−.75	−.74
27.	−.90	−.91	−.92	−.86	−.86	−.87	−.78	−.80	−.82	−.77	−.79	−.80	−.77	−.80
28.	−.82	−.81	−.83	−.82	−.81	−.83	−.87	−.91	−.91	−.87	−.90	−.89	−.71	−.70

Table 5 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.	−.74	−.77	−.58	−.57	−.60	−.60							
22.	−.75	−.77	−.58	−.57	−.60	−.60	1						
23.	−.73	−.70	−.81	−.85	−.76	−.80	.61	.60					
24.	−.74	−.71	−.81	−.86	−.76	−.80	.60	.60	1				
25.	−.71	−.71	−.67	−.67	−.64	−.65	.67	.67	.59	.58			

Table 5 continued

Method	15.	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.
26.	−.76	−.71	−.85	−.85	−.82	−.81	.56	.55	.69	.68	.77		
27.	−.74	−.74	−.71	−.71	−.67	−.68	.68	.68	.64	.63	.97	.81	
28.	−.71	−.70	−.83	−.83	−.78	−.79	.49	.47	.69	.68	.82	.94	.85

All correlations are significant. The numbers in the table correspond to the scoring methods in Tables 2, 3 and 4

References

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

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.

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.

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.

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

Campion, M. C., Ployhart, R. E., & MacKenzie, W. I., Jr. (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., & 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.

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

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.

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 behavioural sciences*. Hillsdale: Earlbaum Associates.

Costa, P. T., & MacCrae, 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). *The SAGE Handbook of Personality Theory and Assessment, 2*, 179–198.

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

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 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: A study in a multi-ethnic police setting. *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.

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

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

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

Flaskerud, J. H. (1988). Is the Likert scale format culturally biased? *Nursing Research, 37*, 185–186.

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.

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

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

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.

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

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.

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.

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.

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.

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.

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 Science 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.

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: U.S. Army Research Institute for the Behavioral and Social Sciences 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: U.S. Army Research Institute for the Behavioral and Social Sciences Technical Report No. 1272.

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: Hogrefe & Huber Publishers.

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

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

Lievens, F., Corstjens, J., Ángel Sorrel, M., 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., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37*, 426–441.

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.

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.

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), 1–34.

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. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*, 63–91.

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

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

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

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., Zibarras, L., & Ashworth, V. (2015). *Situational judgement tests in medical education and training: Research, theory and practice* (pp. 1–15). AMEE Guide No. 100. Medical Teacher.

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.

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.

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

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.

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

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.

Weekley, J. A., & Ployhart, R. E. (2013). *Situational judgment tests: Theory, measurement, and application*. New York: Psychology Press.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2013). On the development of situational judgement 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). New Jersey: Lawrence Erlbaum Associates Inc.

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.

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