Received: 5 August 2018 | Revised: 8 May 2019 | Accepted: 9 May 2019

DOI: 10.1111/ijsa.12251

FEATURE ARTICLE

International Journal of Selection and Assessment | WILEY

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

W. E. de Leng[1] | K. M. Stegers-Jager[1] | M. Ph. Born[2,3] | A. P. N. Themmen[1]

[1]Institute of Medical Education Research Rotterdam, Erasmus MC, Rotterdam, the Netherlands

[2]Institute of Psychology, Erasmus University Rotterdam, Rotterdam, the Netherlands

[3]Optentia and Faculty of Economic and Management Sciences, North-West University, Potchefstroom, South Africa

Correspondence
W. E. de Leng, IMERR, Erasmus MC, Room AE-227, PO Box 2040 3000 CA, Rotterdam, the Netherlands.
Email: w.deleng@erasmusmc.nl

Abstract

We examined the occurrence of faking on a rating situational judgment 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.

KEYWORDS

extreme responding, faking, high-stakes selection, scoring methods, situational judgment test

1 | INTRODUCTION

The predictive validity evidence on situational judgment 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 organization or the educational program for which an individual applies. Generally, SJTs are used to measure noncognitive attributes. SJTs demonstrate sufficient criterion-related validity in personnel selection (McDaniel, Hartman, Whetzel, & Grubb, 2007) and educational admissions (Lievens, Buyse, & Sackett, 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, Patterson, Corstjens, Martin, & Nicholson, 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 favorable 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.

1.1 | 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, Dwight, & Schneider, 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, Krings, & Binggeli, 2016).

1.2 | 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, Viswesvaran, & Reiss, 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, Kleinmann, König, and Melchers (2015) demonstrated a positive relation between faking and job performance and thus proposed that faking should be viewed as socially adequate behavior. 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.

1.3 | 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 & Oswald, 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 & Oswald, 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 summarize, research on the effectiveness of various approaches to deal with faking on personality measures has mixed results.

1.4 | 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, Biderman, & McDaniel, 2005; Oostrom, Köbis, Ronay, & Cremers, 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 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 (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, Weekley, Holtz, & Kemp, 2003) and another study reporting better SJT performance for nonapplicants (Weekley, Ployhart, & Harold, 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., behavioral tendency instructions) (Nguyen et al., 2005).

TABLE 1 Overview of published papers on faking on situational judgment tests (SJTs)

Study	Study type	Study design	Study setting	Response format	Response instruction	Results
Lievens and Peeters (2008)	Lab	BS	Edu	Pick one	WD	Smaller faking effect when students elaborated on their judgment, 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 behavioral 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 and Lievens (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 BS = Between-subjects; d = Cohen's d (effect size); Edu = Educational Pers = Personnel; Field = Real-life study; IM = Impression management; Lab = Faking-induced study; OWD = Others would do; SD = Should do; WD = Would do; WS = Within-subjects.

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 behavioral 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, Sackett, & Buyse, 2009). Finally, due to the higher susceptibility to faking, behavioral 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.

1.5 | 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, Meijer, & Tendeiro, 2017), 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 familiarization with the test format (Hooper, Cullen, & Sackett, 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-emphasized (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, Stemler, Sternberg, Grigorenko, & Hoffman, 2011; Stemler, Aggarwal, & Nithyanand, 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 behaviors during job applications revealed that the proportion of respondents indicating to de-emphasize negative traits was larger than the proportion of respondents indicating to exaggerate positive characteristics (Donovan, Dwight, & Hurtz, 2003). Accordingly, we hypothesized 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 judgment to determine the scoring key (McDaniel, Psotka, Legree, Yost, & Weekley, 2011) and calculate the distance on the rating scale between an individual's judgment and the consensus judgment. 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, Yang, Lievens, & McDaniel, 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.

2 | METHODS

2.1 | 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.

2.2 | 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 ($t(360) = 0.56$, $p > 0.05$) and gender ($\chi^2(1) = 1.22$, $p > 0.05$). On *T2*, the sample that provided data on *T1* and *T2* was significantly younger ($t(589) = 2.43$, $p = 0.015$, $d = 0.20$) and consisted of significantly more females ($\chi^2(1) = 7.77$, $p = 0.005$, $\varphi = 0.11$) than the sample that only provided data on *T2* (*N* = 274). The results of this study were based on the overlapping group (*N* = 317).

2.3 | Situational judgment 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, Ashton, & de Vries, 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, Brugman, & Koops, 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, minimizing, and assuming the worst). See De Leng, Stegers-Jager, Born and Themmen (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 behavior, 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 Supporting Information 1. Supporting Information 2 shows the intercorrelations between SJT scores and the other variables collected during the selection procedure.

2.4 | 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 judgment and the average judgment 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 standardized consensus scoring method calculated the absolute distance between the applicant's judgment and the average SMEs' judgment, but first performed a within-person *z* standardization 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. Standardized 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 judgment was located on the same side of the rating scale as the average judgment 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 judgment is located at the extremes or near the midpoint of the rating scale. The distribution of the SMEs' judgments across the rating scales of the SJT items is presented as Supporting Information 3. 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.

TABLE 2 Average raw, standardized, 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	0.74	**121.94 (7.14)**	0.81	**120.22 (7.32)**
Desirable response options	0.52	61.74 (3.88)	0.70	61.39 (3.73)
Undesirable response options	0.65	**60.20 (4.19)**	0.73	**58.63 (4.75)**
Standardized consensus				
All response options	0.81	**84.31 (4.42)**	0.83	**85.33 (4.11)**
Desirable response options	0.66	**41.71 (2.55)**	0.67	**42.38 (2.13)**
Undesirable response options	0.75	**42.60 (2.46)**	0.79	**42.95 (2.53)**
Dichotomous consensus				
All response options	0.72	**36.38 (2.98)**	0.83	**37.21 (2.93)**
Desirable response options	0.42	**17.62 (1.60)**	0.66	**18.37 (1.35)**
Undesirable response options	0.72	18.76 (1.88)	0.81	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.

2.5 | 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 (α = 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 ($X^2(1) = 0.36$, $p = 0.551$) and age ($t(315) = -0.51$, $p = 0.610$). Responders had a significantly higher pu-GPA than nonresponders ($t(239) = -2.08$, $p = 0.039$, $d = 0.27$). Additionally, responders obtained a significantly higher standardized consensus score on *T2* than nonresponders for all response options ($t(315) = -3.21$, $p = 0.002$, $d = 0.37$), for desirable response options ($t(315) = -2.79$, $p = 0.006$, $d = 0.31$) and for undesirable response

options ($t(315) = -2.87$, $p = 0.004$, $d = 0.32$). Responders also obtained a significantly higher dichotomous consensus score on *T2* than nonresponders for all response options ($t(315) = -2.49$, $p = 0.016$, $d = 0.28$) and undesirable response options only ($t(315) = -2.64$, $p = 0.010$, $d = 0.29$), indicating that 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 standardized and dichotomous SJT scores on *T1* and the dichotomous SJT score based on the desirable response options on *T2*.

3 | RESULTS

3.1 | Mean differences

The mean raw consensus SJT score was significantly lower (worse) on *T2* than *T1* (Table 2), $t(316) = 3.82$, $p < 0.001$, $d_{RM} = 0.23$ (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 ($t(316) = 1.41$, $p = 0.161$, $d_{RM} = 0.11$). For undesirable response options, respondents obtained a significantly lower score on *T2* than *T1*, $t(316) = 4.88$, $p < 0.001$, $d_{RM} = 0.28$ (small effect). On the contrary, the mean standardized consensus SJT score was significantly higher (better) on *T2* than *T1* for all response options ($t(316) = -4.45$, $p < 0.001$, $d_{RM} = -0.25$, small effect), for desirable response options ($t(316) = -4.72$, $p < 0.001$, $d_{RM} = -0.27$, small effect), and for undesirable response options ($t(316) = -2.59$, $p = 0.010$, $d_{RM} = -0.15$). The *T1–T2* difference in the dichotomous consensus SJT score was also significant, $t(316) = -5.01$, $p < 0.001$, $d_{RM} = -0.28$ (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 ($t(316) = -8.25$, $p < 0.001$, $d_{RM} = -0.46$, medium effect), but not for undesirable response options ($t(316) = -0.73$, $p = 0.469$, $d_{RM} = -0.04$). Thus, a faking effect (i.e., higher score in a high-stakes than in a low-stakes setting) was detected for the standardized and dichotomous consensus scoring method, but not for the raw consensus scoring method. In contrast to Hypothesis 1, the faking effect on the standardized and dichotomous SJT scores was larger for desirable response options than undesirable response options.

3.2 | 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, $t(316) = -7.36$, $p < 0.001$, $d_{RM} = -0.46$ (medium effect; Table 2). A significantly higher percentage of extreme ratings was found for both desirable ($t(316) = -8.10$, $p < 0.001$, $d_{RM} = -0.50$, medium effect) and undesirable response options ($t(316) = -5.24$, $p < 0.001$, $d_{RM} = -0.32$, 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 ($t(316) = -0.25$, $p = 0.804$, $d_{RM} = 0.02$). So, the difference between desirable and undesirable response options in extreme responding was not explained by a difference in the opportunity to fake.

3.3 | 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 standardized 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

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	**−0.47**
Desirable response options	**−0.37**
Undesirable response options	**−0.43**
Standardized consensus	
All response options	**0.45**
Desirable response options	**0.37**
Undesirable response options	**0.37**
Dichotomous consensus	
All response options	**0.38**
Desirable response options	**0.32**
Undesirable response options	**0.30**

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 ($p < 0.001$, two-tailed).

consensus scoring method, $t(316) = -2.50$, $p = 0.013$. 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 ($t(316) = 1.72$, $p = 0.087$) or on desirable response options ($t(316) = 0.89$, $p = 0.370$). Additionally, the correlation between the ERS difference and the standardized consensus score difference was significantly stronger than the correlation between the ERS difference and the dichotomous consensus score difference for all response options ($t(316) = 2.42$, $p = 0.016$) and undesirable response options ($t(316) = 2.03$, $p = 0.043$), but not for desirable response options ($t(316) = 1.42$, $p = 0.160$). 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 standardized consensus score difference for all response options ($t(316) = 0.37$, $p = 0.710$), desirable response options ($t(316) = -0.15$, $p = 0.880$), and undesirable response options ($t(316) = 1.10$, $p = 0.270$). 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 standardized 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.

3.4 | Construct validity

As expected, the correlation of the raw consensus SJT scores with honesty-humility was smaller than the correlation of the standardized and dichotomous SJT scores with honesty-humility (Table 4). However, the standardized consensus score based on all response options had a significant and positive correlation to honesty-humility on *T1* ($r = 0.17$, $p = 0.029$) and *T2* ($r = 0.24$, $p = 0.001$). For the dichotomous consensus scoring method, the overall SJT score was also significantly and positively correlated to honesty-humility, but only on *T2* ($r = 0.25$, $p = 0.001$). The SJT score based on undesirable response options correlated significantly and positively to honesty-humility on both *T1* and *T2* for the standardized consensus scoring method ($r_{T1} = 0.22$, $p_{T1} = 0.004$ and $r_{T2} = 0.30$, $p_{T2} < 0.001$) and for the dichotomous consensus scoring method ($r_{T1} = 0.17$, $p_{T1} = 0.023$ and $r_{T2} = 0.33$, $p_{T2} < 0.001$). For the standardized 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).

3.5 | 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 standardized consensus scoring method, a significant difference was found for the overall SJT score ($t(589) = -3.28$, $p = 0.001$, $d = 0.27$, small effect), desirable response options ($t(589) = -2.98$, $p = 0.003$, $d = 0.24$, small effect), and undesirable response options ($t(589) = -2.93$, $p = 0.004$, $d = 0.24$, small effect), all in favor of repeat test takers. In addition, a significant difference favoring 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 ($t(589) = -2.87$, $p = 0.004$, $d = 0.24$, small effect), and undesirable response options ($t(589) = -2.73$, $p = 0.007$, $d = 0.23$, small effect). Finally, repeat test takers used significantly more extreme rating scale points on *T2* than novel test takers, $t(589) = -2.44$, $p = 0.015$, $d = -0.20$ (small effect). Prior exposure to an SJT resulted in

TABLE 4 Correlation to honesty-humility for three scoring methods and for all, desirable and undesirable response options on *T1* and *T2*

		Scoring method		
		Raw consensus	Standardized consensus	Dichotomous consensus
T1	All response options	−0.01	**0.17**	0.13
	Desirable response options	0.01	0.07	0.04
	Undesirable response options	−0.03	**0.22**	**0.17**
T2	All response options	−0.12	**0.24**	**0.25**
	Desirable response options	−0.14	0.08	0.01
	Undesirable response options	−0.08	**0.30**	**0.33**

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, $p < 0.05$.

no retest effects when using the raw consensus score and in small retest effects when using the standardized and dichotomous consensus score. Thus, retest effects—faking or practice—were only detected for scoring methods that controlled for response tendencies.

The SPSS syntax for the above analyses can be found in Supporting Information 4.

4 | 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., standardized 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.

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)
Standardized 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)**
Dichotomous 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). Standard deviations between brackets. Bold numbers indicate a significant difference.

4.1 | Faking

Because the standardized 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 standardized (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 behavioral 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 behavioral tendency instructions. McDaniel et al. (2007) describe SJTs with knowledge instructions as maximal performance tests and SJTs with behavioral tendency instructions as typical performance tests and argue that self-reports of typical behavior 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.

4.2 | Desirable and undesirable response options

The *T1*–*T2* increase in the SJT score based on desirable response options was significant for the standardized (d = −0.27) and dichotomous (d = −0.46) scoring method. For undesirable response options, only the *T1*–*T2* increase in the standardized SJT score was significant (d = −0.15), albeit considerably smaller than the *T1*–*T2* increase

for desirable response options. Additionally, the *T*1–*T*2 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 behavior. Measures of typical behavior 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 behaviors. 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-emphasizing 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 characterized 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.

4.3 | 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 judgment, 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 standardized 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 judgment 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 judgment 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 standardized and dichotomous SJT score—the construct validity was stronger on *T*2 than *T*1, possibly because applicants are familiarized with the SJT format on *T*2 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 *T*2 than *T*1 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) emphasize the importance of standardizing 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).

4.4 | 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 standardized 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 familiarization 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 standardized 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.

4.5 | 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.

4.6 | 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 familiarization (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 standardized 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 generalizability of our findings to real selection settings. However, Niessen et al. (2017) 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.

CONFLICT OF INTEREST

The authors report no conflict of interest.

ORCID

W. E. de Leng https://orcid.org/0000-0001-8296-7239

REFERENCES

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. https://doi.org/10.1007/s10459-017-9796-8

Ashton, M. C., & Lee, K. (2005). Honesty-humility, the Big Five, and the five-factor model. *Journal of Personality*, *73*, 1321–1354. https://doi.org/10.1111/j.1467-6494.2005.00351.x

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: Official Journal of the International Society for Research on Aggression*, *22*, 333–343. https://doi.org/10.1002/(SICI)1098-2337(1996)22:5<333:AID-AB2>3.0.CO;2-K

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. https://doi.org/10.1111/j.1468-2389.2006.00354.x

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

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. https://doi.org/10.1111/j.1468-2389.2006.00340.x

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. https://doi.org/10.1111/medu.13498

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. https://doi.org/10.1007/s10459-016-9720-7

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

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. https://doi.org/10.1207/S15327043HUP1601_4

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 Psychology*, *29*, 479–493. https://doi.org/10.1007/s10869-013-9318-5

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. https://doi.org/10.1111/j.1468-2389.2011.00549.x

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance*, *16*, 1–23. https://doi.org/10.1207/S15327043HUP1601_1

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods*, *4*, 265–287. https://doi.org/10.1177/109442810143005

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. https://doi.org/10.1037/0021-9010.92.2.386

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. https://doi.org/10.1080/01411920903420016

Goffin, R. D., & Boyd, A. C. (2009). Faking and personality assessment in personnel selection: Advancing models of faking. *Canadian Psychology/Psychologie Canadienne*, *50*, 151–160. https://doi.org/10.1037/a0015946

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. https://doi.org/10.1111/j.0965-075X.2003.00256.x

Griffin, B., & Wilson, I. G. (2012). Faking good: Self-enhancement in medical school applicants. *Medical Education*, *46*, 485–490. https://doi.org/10.1111/j.1365-2923.2011.04208.x

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. https://doi.org/10.1108/00483480710731310

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. https://doi.org/10.1111/j.1754-9434.2008.00053.x

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. https://doi.org/10.1177/0022022114534773

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. https://doi.org/10.1037/0021-9010.91.1.9

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. https://doi.org/10.1037/0022-3514.66.2.276

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. https://doi.org/10.1037/0022-3514.72.3.515

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology*, *92*, 1270–1285. https://doi.org/10.1037/0021-9010.92.5.1270

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 Publishers.

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. https://doi.org/10.1080/1359432X.2014.915215

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. https://doi.org/10.1207/S15327043HUP1304_3

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. https://doi.org/10.1111/ijsa.12007

Kulas, J. T. (2013). Personality-based profile matching in personnel selection: Estimates of method prevalence and criterion-related validity. *Applied Psychology*, *62*, 519–542. https://doi.org/10.1111/j.1464-0597.2012.00491.x

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. https://doi.org/10.1037/a0020375

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*, 329–358. https://doi.org/10.1207/s15327906mbr3902_8

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. https://doi.org/10.1207/s15327043hup1802_4

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. https://doi.org/10.1111/ijsa.12084

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. https://doi.org/10.1037/0021-9010.90.3.442

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. https://doi.org/10.1111/j.1744-6570.2005.00713.x

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. https://doi.org/10.1111/medu.13060

Lievens, F., & Peeters, H. (2008). Impact of elaboration on responding to situational judgment test items. *International Journal of Selection and Assessment*, *16*, 345–355. https://doi.org/10.1111/j.1468-2389.2008.00440.x

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. https://doi.org/10.1037/a0014628

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. https://doi.org/10.1111/j.1744-6570.2007.00065.x

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. https://doi.org/10.1037/a0021983

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, *85*, 812–821. https://doi.org/10.1037110021-9010.85.5.812

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. https://doi.org/10.1111/j.1744-6570.2007.00089.x

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. https://doi.org/10.1037//1082-989X.7.1.105

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.

Mueller-Hanson, R., Heggestad, E. D., & Thornton Iii, 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. https://doi.org/10.1037/0021-9010.88.2.348

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. https://doi.org/10.1027/1015-5759.24.3.181

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. https://doi.org/10.1111/j.1468-2389.2005.00322.x

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017). 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. https://doi.org/10.1016/j.paid.2016.11.014

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. https://doi.org/10.1016/j.paid.2016.03.075

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. https://doi.org/10.1080/08959285.1998.9668033

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. https://doi.org/10.1037/0021-9010.81.6.660

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. https://doi.org/10.1016/j.jrp.2017.09.001

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. https://doi.org/10.1111/1467-6494.00041

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. https://doi.org/10.1177/0013164404268672

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. https://doi.org/10.1080/08959285.2017.1407320

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. https://doi.org/10.1111/j.1744-6570.2003.tb00757.x

Robson, S. M., Jones, A., & Abraham, J. (2007). Personality, faking, and convergent validity: A warning concerning warning statements. *Human Performance*, *21*, 89–106. https://doi.org/10.1080/08959280701522155

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. https://doi.org/10.1037/0021-9010.83.4.634

Roulin, N., Krings, F., & Binggeli, S. (2016). A dynamic model of applicant faking. *Organizational Psychology Review*, *6*, 145–170. https://doi.org/10.1177/2041386615580875

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. https://doi.org/10.1037/a0016810

Schmitt, N., & Oswald, F. L. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology*, *91*, 613–621. https://doi.org/10.1037/0021-9010.91.3.613

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251. https://doi.org/10.1037/0033-2909.87.2.245

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. https://doi.org/10.1111/ijsa.12143

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. https://doi.org/10.1037/a0025711

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. https://doi.org/10.1146/annurev-orgpsych-032516-113349

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. https://doi.org/10.1037/0021-9010.85.1.50

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*, 197–210. https://doi.org/10.1177/00131649921969802

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

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. https://doi.org/10.1207/s15327043hup1704_5

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. https://doi.org/10.1016/j.jvb.2017.11.005

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, *19*, 188–202. https://doi.org/10.1016/j.hrmr.2009.03.007

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. https://doi.org/10.1027/1015-5759/a000292

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: de Leng WE, Stegers-Jager KM, Born MP, Themmen APN. Faking on a situational judgment test in a medical school selection setting: Effect of different scoring methods? *Int J Select Assess.* 2019;27:235–248. https://doi.org/10.1111/ijsa.12251