A closer look at response options: Is judgment in situational judgment tests a function of the desirability of response options?

Katarina KAMINSKI

Jorg FELFE

Philipp SCHAEPERS
Singapore Management University

Stefan KRUMM

Citation

A closer look at response options: Is judgment in situational judgment tests a function of the desirability of response options?

Katarina Kaminski[1]

Jörg Felfe[1]

Philipp Schäpers[2]

Stefan Krumm[3]

[1]Work & Organizational Psychology, HelmutSchmidt-Universität Hamburg, Hamburg, Germany

[2]Lee Kong Chian School of Business, Singapore Management University, Singapore, Singapore

[3]Institute of Psychology, Freie Universität Berlin, Berlin, Germany

Correspondence Katarina Kaminski, Work & Organizational Psychology, Helmut-Schmidt-Universität Hamburg, Holstenhofweg 85, 22043 Hamburg, Germany. Email: katarina.kaminski@hsu-hh.de

Abstract

The current study builds on the current scholarly debate about SJTs potentially being less situational than previously assumed. Specifically, we respond to recent calls to examine general (situation unspecific) information included in response options as a guide to SJT responses. Across three consecutive studies and three different forms of SJT administration (standard, without situation descriptions, under fake-good instructions), the relevance of social desirability of response options on SJT responses was examined. Results suggest that social desirability of response options is significantly related to test takers' response. This finding generalized across different forms of SJT administration. Across studies and together with the plausibility of response options, desirability explained about one-third of reliable variance in test takers' response to an SJT. Implications for SJT theory and development are discussed.

1 INTRODUCTION

Situational judgment tests (SJTs) are widely used instruments in personnel selection. Typically, SJTs consist of brief descriptions of job-related critical situations along with multiple-choice response alternatives. Using a closed response format, SJTs ask participants how they should or would react in such situations (Motowidlo, Dunnette, & Carter, 1990). Until recently, the unanimous understanding among researchers was that SJTs function as (low-fidelity) simulations, that is, participants envision the described situations and respond according to their specific construal of the situation (Lievens & Motowidlo, 2016). In other words, it was assumed that SJTs capture *situational* judgment.

This notion, however, has been questioned by a recent study (Krumm et al., 2015). Specifically, these authors administered SJT items either with or without situation descriptions in the item stem. Their results were intriguing: Although a central component of SJTs, that is, the situation descriptions, was not available for test takers, their performance in many items was comparable to a control group, which were given the

situation descriptions. These authors concluded that judgment in SJTs may not be as situational as initially assumed.

Understandably, this finding has led to a controversial debate among scholars (Borneman, 2016; Brown, Jones, Serfass, & Sherman, 2016; Chen, Fan, Zeng, & Hack, 2016; Crook, 2016; Fan, Stuhlmann, Chen, & Weng, 2016; Harris, Siedor, Fan, Listyg, & Carter, 2016; Harvey, 2016; Lievens & Motowidlo, 2016; McDaniel, List, & Kepes, 2016; Melchers & Kleinmann, 2016; Naemi, Martin-Raugh, & Kell, 2016; Torres & Beier, 2016; Whetzel & Reeder, 2016). Among other points, it was discussed whether response options are a valid source of information in SJTs and might enable test takers to deduce correct solutions (Harris et al., 2016; Melchers & Kleinmann, 2016). In fact, recent research suggests that response options choice in SJTs is determined by more than test takers' perusal of the situation (Leeds, 2012, 2018). For instance, Schmitt, Ryan, Bradburn, and Nye (2018) revealed that the extent to which response options reflect cultural content (e.g., individualism) differentially affects responses of people with different cultural backgrounds. So, a closer look at response options is not only a timely topic, it also became clear that the debate about the role of situations in SJTs cannot be settled without simultaneously examining if and how response options function as sources of information in SJTs. This research question is relevant for several reasons: First, we provide knowledge about the role of response options (specifically: their desirability) for SJT performance, which is an issue that has received little research attention so far. Second, by including SJT items without situation descriptions and a faking instruction (see below), we examine important boundary conditions of the relevance of response options. Third and more generally, our research adds to current theorizing about the functioning of SJTs and responds to recent calls to open the black box of SJTs (Ployhart, 2006; Schmitt & Chan, 2006; Whetzel & McDaniel, 2009). From a practical perspective, we raise test developers' awareness for the importance of response option choice and make initial recommendations about how to critically examine the response options of each SJT item.

To this end, we examined the effect of information included in response options on judgment in an SJT across three consecutive studies and three different forms of SJT administration (standard, without situation descriptions, under fake-good instructions). Specifically, we targeted the social desirability of response options as a source of information (a) because it is applicable to a broad range of SJT responses (since most SJTs target social skills; Christian, Edwards, & Bradley, 2010), and (b) because previous research has shown that individuals search for and detect signs of desirability in SJTs (Krumm et al., 2015; Leeds, 2012, 2018). Therefore, in the current research, we first assessed the social desirability of each response option included in an SJT (Study 1). Subsequently, we examined whether social desirability of response options— understood as a feature of the SJT and not as an individual difference variable—predicts test takers' responses to an SJT, which was administered in its standard form and with a fake-good instruction (Study 2) or in its standard form and without situation descriptions (Study 3).

1.1 Situational judgment in SJTs

Since their reintroduction in the 1990s, SJTs have received ample attention, both in research and practice (Christian et al., 2010; Motowidlo et al., 1990). Combining advantages of simulations with the cost-efficient administration and scoring of tests, SJTs typically receive favorable applicant reactions (e.g., Chan & Schmitt, 1997). Importantly, SJTs can predict job-related criteria reasonably well (Christian et al., 2010; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001), also incrementally above and beyond personality and cognitive ability (McDaniel, Hartman, Whetzel, & Grubb, 2007). Furthermore, the available evidence suggests that they may be less easy to fake than personality tests (Kanning & Kuhne, 2006). Finally, with advancing technological opportunities, video-based and interactive forms of test administration are available to provide more realistic representations of real job-related situations (MacCann, Lievens, Libbrecht, & Roberts, 2016).

Until recently, there was no reason to doubt that SJTs function as (low-fidelity) simulations. In fact, there was a unanimous agreement among scholars that SJTs capture test takers construal of the job-related situations presented and their context-dependent knowledge about appropriate reactions to these situations (McDaniel et al., 2007). In line with this notion, situations are usually viewed as the core component of SJTs. This is also why many guidelines for test developers emphasize the need (i) to identify job-related

critical situations by interviewing subject matter experts (SMEs), (ii) to screen and select situations for appropriateness, and (iii) to write comprehensive item stems enabling test takers to envision the critical situations and react to them (Corstjens, Lievens, & Krumm, 2017; Weekley & Ployhart, 2013). In short, a costly procedure is seen as vital to equip SJTs with their purportedly most important feature, that is, situation descriptions.

Interestingly, this notion has rarely been put to an empirical test. Only recently, researchers began to explicitly examine the role of situation construal in SJTs (Krumm et al., 2015; Rockstuhl, Ang, Ng, Lievens, & Van Dyne, 2015). For instance, Rockstuhl et al. asked test takers to not only respond to an SJT, but also to report their situation construal. In line with the widely agreed notion of SJTs as simulations, these authors found that the appropriate construal of situations was significantly correlated with SJT performance. Krumm et al., however, presented results that point to the opposite conclusion. In several studies, they administered SJTs either with or without situations included in the item stems. That is, half of the test takers were not able to make any situation construal based on situation descriptions. Despite this immense handicap, such test takers were able to perform at least as well (as the ones who received situation descriptions) in between 46% and 71% of the administered SJT items (depending on the correction for alpha inflation). This was true for several samples (students and nonstudents), across several SJTs (from diverse construct domains), and across response instructions (would- vs. should-do). Thus, Krumm et al. concluded that SJTs may be less context-dependent than previously assumed.

1.2 Response options as sources of information

The results reported by Krumm et al. (2015) have sparked a controversial debate about the relevance of situation descriptions in SJTs. Besides other arguments for and against a dismissal of situations in SJTs (e.g., Crook, 2016; Chen et al., 2016; Fan et al., 2016; Harvey, 2016; Lievens & Motowidlo, 2016; McDaniel et al., 2016), Melchers and Kleinmann (2016) as well as Harris et al. (2016) argued that response options may provide valid and oftentimes sufficient information. Specifically, Melchers and Kleinmann stated that, in the absence of situation descriptions, "test takers try to understand the situation to which the response options are related, with the only difference that an understanding of this situation is (at least somewhat) more difficult in comparison with when an item stem with a description of the situation is included" (p. 32). Likewise, Harris et al. argued that "the content of SJT response options retain situational descriptors" (p. 24). So, these authors provide an alternative interpretation of the Krumm et al. results and emphasize the role of response options in SJTs.

Albeit the specific claim made by several researchers (Fan et al., 2016; Harris et al., 2016; Melchers & Kleinmann, 2016), that is, that test takers gather situational information from response options, has not been tested explicitly, evidence that response options serve as sources of information in SJTs is indeed available. Leeds (2012) distinguished two sequential steps of processing when taking an SJT. In the first step, he posited that participants evaluate each response option per se (e.g., with regard to its reasonableness) and in relation to other response options (e.g., is this a better response than other options). Specifically, Leeds referred to this first process as cognitive acuity, which he defined as "the capacity to detect correctness and to distinguish between differences in correctness among simultaneously presented situation-specific response options" (p. 166). Using methods from signal detection theory, he confirmed that individual and relative signs of correctness in response options (obtained through ratings of SMEs) were indeed related to response choice in SJTs. In the second step of processing, it is assumed that test takers inspect the situation descriptions to validate their initial response choice. In short, theorizing and evidence provided by Leeds suggest that response options play a pivotal role in responding to SJTs.

While the insights provided by Leeds (2012) are intriguing, the extent to which the context (provided in the situation descriptions) influenced participants' correctness perceptions remained unclear. In other words, participants' ability to distinguish between response options (i.e., their cognitive acuity) may be, in part, contingent on the item-specific context and participants' level of context-specific knowledge. In the current research, we advance previous research by focusing on features of response options that may suffice, without any further context, to guide participants' decision-making. Next, we introduce social desirability as a feature of response options that may provide vital information for test takers.

1.3 Social desirability in SJTs

Socially desirable responding is frequently understood as a characteristic of individuals (i.e., an individual difference variable) and, in this tradition, defined as "the tendency […] to select answers on tests that will result in others viewing them in the most favorable way" (McFarland & Ryan, 2000, p. 813), which may be the result of an unconscious or an intentional act (Paulhus, 1986). The impact of intentionally presenting oneself in a favorable way on results of personnel selection instruments has been subject to vivid and long-lasting debates among scholars. Despite disagreements whether intentional socially desirable responding contributes to or distorts criterion-related validity of personnel selection instruments (e.g., Ellingson, Sackett, & Connelly, 2007; Goffin & Boyd, 2009; Griffith, Chmielowski, & Yoshita, 2007), ample evidence attests that applicants can improve their scores if instruments make use of rating scales (e.g., Hooper & Sackett, 2008; Kanning & Kuhne, 2006; Viswesvaran & Ones, 1999).

Social desirability may also be understood as a characteristic of a test. That is, high correlations between individual differences in socially desirable responding (see above) and outcomes of a particular test (e.g., an SJT) are interpreted as the extent to which this test is prone to social desirability (e.g., Furnham, 1986). In the same vein, mean differences in a particular test between groups of honest responders and those instructed to present themselves in a favorable way are used to describe the test with regard to its proneness to socially desirable responding (e.g., from the realm of SJTs, see Kanning & Kuhne, 2006).

While it is widely acknowledged that personality tests are prone to socially desirable responding, much less is known about socially desirable responding in SJTs (Hooper, Cullen, & Sackett, 2006). While some researchers found no or only small effects of intentional socially desirable responding on SJTs (Kanning & Kuhne, 2006; Nguyen, Biderman, & McDaniel, 2005), others reported large effects of up to $d = 0.89$ (Peeters & Lievens, 2005). Against the backdrop of this, Hooper et al. conceded that too few studies and too heterogeneous results are available to judge how prone SJTs are to test takers' socially desirable responding. However, they speculated that design features of SJTs (such as a "would-do" vs. "should-do" response instruction) may affect social desirable responding. While consented that "would-do" response instruction rather capture personality constructs and "should-do" response instruction rather capture cognitive constructs (McDaniel & Nguyen, 2001; Ployhart & Ehrhart, 2003), it is additionally assumed that performance of SJTs with "would-do" response instructions (i.e., Peeters & Lievens, 2005) may be faked more easily than performance of SJTs with "could-do response" instructions (Nguyen et al., 2005).We will next elaborate on social desirability of response options as a feature of SJTs in more detail.

1.3.1 Social desirability as a design feature of SJTs

In light of the notion that response options in SJTs may contain enough information to even render situation descriptions superfluous (Harris et al., 2016; Krumm et al., 2015; Leeds, 2012; Melchers & Kleinmann, 2016), the social desirability of response options may be an important design feature influencing test takers' responses to SJT items. In fact, results provided by Krumm et al. (2015) support this view. These authors conducted verbal protocol analyses with participants taking an SJT without situation descriptions. That is, participants were asked to inspect sets of response options, to identify the most appropriate response, and to verbalize their strategies in identifying appropriate responses. Strategies described by these participants were classified in five broad categories. In line with Leeds, a frequently reported strategy was to compare response options. Notably, additional strategies to evaluate response options included scrutinizing their feasibility or plausibility, as well as their fairness and motivating effect on others. Considering that Krumm et al. administered a teamwork SJT, judging responses on their fairness and motivating effect describes—to some extent—socially desirable responding. So, a preliminary answer to the question of what constitutes the correctness valence of SJT responses (Leeds, 2012), besides their general effectiveness, may be their plausibility as well as their social desirability.[1]

Albeit knowing how the social desirability of response options influences responding to SJTs is important both from a theoretical (i.e., understanding response processes in SJTs) and a practical (i.e., avoid differences in desirability of response options in the construction) perspective, no study has so far explicitly

examined this. In the current research, we conducted three consecutive studies to test whether the social desirability of response options affects SJTs responses under various conditions; that is, for a standard form of SJT administration, under a faking instruction, and when an SJT was administered without situation descriptions. Building on previous research suggestive of response options being valid sources of information, we posit:

Hypothesis 1 *The social desirability of response options is significantly and positively related to test takers' responses in an SJT.*

Furthermore, we argue that the link between the social desirability of response options and test takers' responses in an SJT will be amplified when (a) test takers are instructed to present themselves as favorably as possible. We posit that in this case, test takers will base their response less on their construal of the situation and their knowledge about effective behavior and more on information about social desirability as included in response options.

Hypothesis 1a *The social desirability of response options is more strongly related to test takers' responses in an SJT taken under a fake-good instruction than in an SJT taken under an honest instruction.*

Finally, we posit that when no additional information in form of situation descriptions in the item stem is presented, test takers will base their judgment on the plausibility and general effectiveness of response options (Krumm et al., 2015), but also on their social desirability. The main reason for this is that the typically assumed processes in SJTs, situational construal and reliance on procedural knowledge about effective behavior, are less helpful in the absence of situation descriptions. Thus, test takers have to rely on other strategies, which exploit the information inherent in response options. We thus hypothesize:

Hypothesis 1b *The social desirability of response options is more strongly related to test takers' responses in an SJT without situation descriptions than in an SJT with situation descriptions.*

Below, three consecutive studies are presented. Study 1 was conducted to gauge the social desirability of response options in an SJT. In Studies 2 and 3, we administered an SJT in its standard form, hence providing a test of H1. In Study 2, we additionally administered the same SJT under a fake-good instruction as a test of H1a. Finally, Study 3 also included a version of the SJT in which situation descriptions were omitted, thus testing H1b.

2 STUDY 1

2.1 Methods

2.1.1 Participants and procedure

The sample included $N = 132$ subject matter experts (SMEs). Participants (11% female) were on average 29.2 years old ($SD = 6.2$, range from 21 to 52). The majority of participants were experienced soldiers (5% enlisted ranks, 78% noncommissioned officers, 17% officers) from different military bases across Germany. On average, they had 7.7 years of professional military experience ($SD = 5.59$, range from 1 to 33). The soldiers participated voluntarily. Anonymity of their results was guaranteed.

2.1.2 Study design and materials

Data about the desirability of response options were collected in a paper–pencil format and in a proctored setting. All participants received the same study materials. The study lasted 1 hr in total.

2.1.3 Situational Judgment Test

The SJT used in all the herein reported studies examines knowledge of leadership and decision-making skills in the military context (Felfe, Wunderlich, & Kaminski, 2016), similar to the Leadership Judgment Indicator (LJI; Lock, Wheeler, Burnard, & Cooper, 2005). This SJT consists of 17 leadership situations presented in a written format. Situations were obtained through critical incident interviews (Flanagan, 1954) and screened by experts for appropriateness, comprehensiveness, and readability. A sample situation is:

You are in the position of a platoon leader. You have been asked to assign one of your soldiers to an international exchange program to Denmark. The assignment will start in two month and will last for six weeks. To your opinion, it is important to select a soldier who is highly motivated. You have just started your commandment one month ago and you are still getting to know each other. From your point of view the decision should be well-considered. (Felfe et al., 2016, p. 10)

After each military leadership situation, participants are asked to rate four different responses to that particular situation. Specifically, participants are asked to rate how effective they think each response is in a given situation. Each response option reflects a distinct leadership style, that is, directive, consultative, consensual, and delegating, as specified by established leadership theories (Lock et al., 2005; Vroom & Yetton, 1973). In the typical application of the SJT, scores are obtained by assessing the convergence of a participant's ratings with the ideal solution as delineated from the theory of situational leadership (Vecchio, 1987; Vroom & Yetton, 1973) and confirmed by a panel of military leadership experts (Felfe et al., 2016). In the herein reported studies, we used the "pure" rating of participants (and not its convergence with the ideal solution), since we were interested in the extent to which actual responding to an SJT (regardless of its effectiveness) was driven by desirability of response options.

This SJT enables a conservative test of our hypotheses: First, each item of this SJT contained response options representing four distinct leadership styles. Hence, appropriateness of a response was solely determined on the fit between a leadership style and the situation described in the item stem and not, like in several other SJTs, on the general appropriateness or desirability of responses.[2] Second, this SJT came with a rating response format (i.e., participants rated each response option). Effects of social desirability on SJTs have—to our knowledge—so far only been found for pick-the-best/worst response formats (e.g., Peeters & Lievens, 2005). Third, this SJT used a knowledge instruction. Albeit evidence is sparse, Nguyen et al. (2005) present more consistent evidence for desirability effects on SJTs with behavioral tendency than for SJTs with knowledge instructions. In fact, we would argue that when given a knowledge response instruction, test takers should have less inclination to score socially desirable but rather to receive high test scores. This would be consistent with McDaniel and Nguyen (2001), who reported a higher cognitive saturation of an SJT with a knowledge instruction. Fourth, social desirability of response options was determined by presenting all response options in a random sequence to participants (see below). So, participants were not able to compare the four response options belonging to a specific item with regard to their desirability. In other words, we deliberately made the process of determining social desirability independent from the process of responding to the actual SJT. In sum, the current research will likely represent the lower bound of social desirability effects on SJTs.

Notably, Study 1 did not include the SJT in its standard form. Since we aimed at gauging the desirability of response options regardless of their effectiveness in a particular situation, all 68 response options (the standard form of the SJT included 17 situations with 4 response options each) were presented in a random sequence and without situations. Participants were asked to rate each response with regard to their social desirability ("How socially desirable is this response alternative?"), their professional desirability ("How desirable is this response alternative in terms of military leadership?"), and their general plausibility ("How plausible do you think is this response alternative?") on a 5-point Likert scale ranging from 1 (*undesirable* or *implausible*) to 5 (*desirable* or *plausible*). We included professional desirability to also account for the (military) context in which desirability ratings were obtained. We also considered plausibility of response options as another viable information included in response options (Leeds, 2012, 2018).

2.2 Results

On average, response options yielded a social desirability score of 3.45 ($SD = 0.48$, range from 1.95 to 4.30). An example of a response option rated as highly socially desirable is "A solution for the assignments of duties is jointly developed in a short meeting with your soldiers." An example of a response option that received low social desirability ratings is "As you have got all necessary information you will make the decision for the get together on your own." Similar average scores were obtained for professional desirability, $M = 3.53$ ($SD = 0.41$, range from 1.98 to 4.22), and for plausibility of the response options, $M = 3.45$ ($SD = 0.39$, range from 2.19 to 4.02).

Importantly for subsequent analyses, the average social desirability score obtained for each response option was only lowly correlated with the ideal effectiveness rating as reported by the test authors ($r = 0.14$), indicating that the social desirability scores obtained in Study 1 are distinct from the effectiveness of the response options. Professional desirability and response plausibility, however, showed slightly higher correlations with effectiveness ratings (rs = 0.26 and 0.24, respectively).

The above-reported average scores—obtained from a sample of SMEs and through ratings on response options which were presented without any further information about the situation they refer to—were used as a reference in further analyses (Studies 2 and 3). That is, we subsequently treated the scores obtained per response option as an estimate of their relevant features, that is, their social desirability, professional desirability, and plausibility.

3 STUDY 2

Study 2 was conducted to test H1 and H1a. It therefore included an SJT in its standard form as well as the same SJT administered with a fake-good instruction.

3.1 Methods

3.1.1 Participants and procedure

The sample used in Study 2 consisted of $N = 135$ soldiers (9.6% female) from different military bases across Germany. The majority of soldiers were either enlisted ranks (44%) or served as noncommissioned officers (50%; another 6% were officers). Their mean age was 27.1 years ($SD = 4.9$, range from 17 to 53). On average, they had 6.4 years of professional military experience ($SD = 4.7$, range from 1 to 34).

Study 2 employed a between-subjects design. Participants were randomly assigned to the SJT administered in its standard form ($n = 68$) or received a fake-good instruction ($n = 67$).[3] In the standard condition, participants were instructed to respond as honestly as possible. To ensure honest responding, test takers were told to participate as independent experienced experts in the field of military leadership and that their responding will be used for further development of the test. Being in this supervisory role, there was no need to be dishonest or to meet social standards. In the fake-good condition, participants received the following instruction: "Please answer the questions in the most favorable way. Imagine you were applying for the post you always wanted to be offered. Your aim is to score as high as you can in this test." Response instructions were adapted from Peeters and Lievens (2005) and Nguyen et al. (2005). We included this condition to examine if test takers' rely more strongly on the desirability of response options when instructed to fake good. In both conditions, participants took the SJT in a proctored setting. Test duration was 1 hr. Participation was voluntary and anonymous.

3.1.2 Situational Judgment Test

The SJT described along with Study 1 was administered in its full version, that is, with situation descriptions and four response alternatives per each situation. Scores (indicating the convergence of a participant with the ideal solution as determined by military leadership experts) were sufficiently reliable,

Cronbach's alpha = 0.62 and 0.61, respectively, in the standard and fake-good condition and exceeded the typically reported reliability estimates for SJTs (see Campion, Ployhart, & MacKenzie, 2014; Catano, Brochu, & Lamerson, 2012; Kasten & Freund, 2016).

3.2 Results

The link between social desirability as a feature of response options and the actual responses of the current sample was determined by correlating each participant's ratings ($17 \times 4 = 68$ ratings) with the social desirability scores of each response option as obtained from Study 1. Hence, in the standard condition, as many correlations as there were participants were calculated and averaged (using Fisher-z-transformation, Fisher, 1921). The same procedure was conducted for the fake-good condition as well as for professional desirability and response plausibility.

The results are presented in Table 1. In line with H1, the social desirability of response options was moderately and significantly related to test takers' responses (mean $r = 0.29$). A similar result was obtained for the professional desirability of response options (mean $r = 0.30$). Interestingly, the plausibility of response options showed a significantly closer link to test takers' responses (mean $r = 0.38$) than the two forms of desirability (as evidenced by the nonoverlapping confidence intervals).

Table 1. Bivariate correlations between characteristics of response alternatives and test takers' responses (study 2)

	Mean r[a]	Fisher's z	90% BC[a] confidence interval[b]	
			Lower bound	Upper bound
Social desirability of responses				
SJT in its standard form	0.29	0.303	0.259	0.348
SJT with fake-good instruction	0.32	0.327	0.294	0.361
Professional desirability of responses				
SJT in its standard form	0.30	0.311	0.280	0.343
SJT with fake-good instruction	0.29	0.302	0.277	0.328
Plausibility of responses				
SJT in its standard form	0.38	0.401	0.372	0.429
SJT with fake-good instruction	0.39	0.411	0.380	0.444

[a] Mean rs are computed by back-transforming average Fisher's z scores.
[b] Bias-corrected and accelerated confidence intervals based on 1,000 bootstrap samples.

When administered under a fake-good instruction, social desirability was also moderately related to test takers' responses (mean $r = 0.32$). The average correlation did not differ from the one observed in the standard condition (as indicated by the nonoverlapping confidence intervals), thus failing to support H1a. The same is true for professional desirability and response plausibility (see Table 1).

4 STUDY 3

Study 3 was conducted to test H1 and H1b. It therefore included an SJT in its standard form as well as the same SJT administered without situation descriptions.

4.1 Methods

4.1.1 Participants and procedure

The sample of Study 3 included $N = 95$ soldiers (26.3% female), which participated in a military training course. All soldiers served as officers. Their mean age was 26.3 years ($SD = 2.3$, range from 23 to 36). On average, they had 6.7 years of professional military experience ($SD = 2.1$, range from 2 to 16).

Paralleling Study 2, we again employed a between-subjects design. Participants were randomly assigned to the SJT administered in its standard form ($n = 51$) or received the same SJT without situation descriptions ($n = 44$).[4] For examples of SJTs that can be completed even when no information about the situations is provided, see Krumm et al. (2015). We included this condition in the current research to examine if test takers' rely more strongly on the desirability of response options when no further information (i.e., about the situation in question) is available. In both conditions, participants took the SJT in a proctored setting. Test duration was 1 hr. Participation was voluntary and anonymous.

4.1.2 Situational Judgment Test

The SJT described along with Study 1 was administered. Scores (indicating the convergence of a participant with the ideal solution as determined by military leadership experts) were sufficiently reliable, Cronbach's alpha = 0.50 and 0.61, respectively, in the standard and in the no-situation condition.

4.2 Results

The main results are presented in Table 2. Similar to results reported in Study 2, the social desirability of response options was significantly related to test takers' responses (mean $r = 0.17$). Albeit the magnitude of the correlation was lower than in Study 2, this finding is in line with H1 (since the confidence interval excluded 0). The professional desirability of response options as well as their plausibility were significantly related to test takers' responses (mean rs = 0.31). The magnitude of these correlations was similar to those observed in Study 2.

When comparing conditions with and without situation descriptions, no significantly different results were obtained. Though there was a small increase in the link between the social desirability of response options and test takers' responses (from $r = 0.17$ to 0.24), this difference was not significant (as evidenced by overlapping confidence intervals, see Table 2), thus failing to confirm H1b. The same findings were observed for professional desirability and response plausibility.

Finally, we combined data obtained from the SJTs in their standard form as included in Studies 2 and 3. We predicted each response of each participant by the three response option features simultaneously (i.e., by their social desirability, professional desirability, and plausibility). This was conducted to determine the total amount of variance in SJT responses that can be explained by social and professional desirability as well by plausibility of response options. Hence, we ran $n = 119$ (which is the number of participants included in Studies 2 and 3 who took the standard SJT form) independent regression analyses and calculated the average multiple R across these analyses. The average multiple correlation was $R = 0.431$. In other words, the three response option features together explained 19% of variance in SJT responses. We followed up on this results with relative weights analyses. This was done to quantify the relative importance of each of the three response option features. Results revealed that they were about equally important: social desirability explained 5%, professional desirability 6%, and plausibility 7% of variance in test takers' responses.

Table 2. Bivariate correlations between characteristics of response alternatives and test takers' responses (study 3)

	Mean r[a]	Fisher's z	90% BC[a] confidence interval[b]	
			Lower bound	Upper bound
Social desirability of responses				
SJT in its standard form	0.17	0.172	0.121	0.222
SJT without situation descriptions	0.24	0.247	0.184	0.308
Professional desirability of responses				
SJT in its standard form	0.31	0.325	0.292	0.358
SJT without situation descriptions	0.31	0.321	0.292	0.350
Plausibility of responses				
SJT in its standard form	0.31	0.320	0.289	0.350
SJT without situation descriptions	0.36	0.371	0.334	0.406

[a] Mean rs are computed by back-transforming average Fisher's z scores.
[b] Bias-corrected and accelerated confidence intervals based on 1,000 bootstrap samples.

5 GENERAL DISCUSSION

The current research examined the extent to which information provided by response options of an SJT-affected test takers' response behavior. Across several forms of SJT administration (standard, under fake-good instruction, without situation descriptions), we found a moderate and significant correlation between social desirability of response options and test takers' responses. Importantly, this correlation was not moderated by SJT administration. That is, SJT responses obtained from (i) the standard SJT form or (ii) under a fake-good instruction or (iii) without presenting situation descriptions all yielded similar correlations with the social desirability of response options. The herein considered response option features (social desirability, professional desirability, and plausibility) explained 19% of SJT response variance.

These results have several theoretical implications. First, our results add to recent theorizing that SJTs may be less situational than originally assumed (Krumm et al., 2015; Lievens & Motowidlo, 2016; Schäpers, Lievens, & Krumm, 2017). Recall that social and professional desirability as well as plausibility of response options were determined by SMEs who did not receive any information about the job-related situations to which the response options belonged. Yet, 19% of variance in test takers' responses to the situations presented by the SJT was explained by (non-situational) response option features. Considering the reliability estimates of approximately 0.60 obtained for the herein used SJT, this means that about one-third of reliable variance can be attributed to response option features. In fact, this proportion is likely higher in other SJTs, in which response options are not derived to reflect four aspects of an overarching theory (in our case leadership styles) and, as a result, in which the social desirability of response options may vary more strongly—thus potentially rendering other SJTs even less situational.

Second, our results also extend previous theorizing by offering insights that complement the prevailing focus on personal attributes. That is, the ongoing debate on the context-(in)dependency of SJTs has mainly

addressed test takers' general domain knowledge (e.g., Lievens & Motowidlo, 2016) and, relatedly, their implicit trait policies (Motowidlo, Hooper, & Jackson, 2006) as well as broad personality dimensions and cognitive ability (Ziegler & Horstmann, 2017). These approaches may be complemented by more fine-grained knowledge about aspects of SJTs that activate these personal attributes. Differences among response options in their desirability and plausibility have the potential to capture candidates' general knowledge (about what is generally desirable and plausible) as well as their cognitive ability (to correctly discern subtle differences in desirability and plausibility). So, we suggest that identifying personal attributes as explanatory variables of SJT performance is one way to answer how situational SJTs in fact are, whereas identifying structural aspects of the test is another way to look at it.

Third, we also complement research by Leeds (2012), who posited that participants make use of response options in two different ways. That is, he suggested that test takers evaluate each response option per se and in relation to other response options. A notable difference between his approach and ours is that we determined the information provided by response options regardless of the context they were presented in (i.e., the situation description they belonged to). That is, their desirability and plausibility were determined by judges who were unaware of the corresponding situation descriptions. Hence, this study provides insights about unique features of response options, while Leeds provided insights about response processes that occur together with test takers' perusal of the context.

Fourth, we posited that the focus of test takers would shift under certain forms of SJT administration. That is, when asked to present themselves as favorably as possible (fake-good instruction) or when left uninformed about the situation in question, we assumed that desirability and plausibility of responses would be a stronger guide in deciding about the effectiveness of responses. The current results, however, suggest that this is not the case. Instead, we found that desirability and plausibility of response options are heavily used in the standard form of SJT administration already (as they explained one-third of reliable variance in test takers' responses). One possible implication for research on faking in SJTs is that SJTs may be less prone to faking (Kanning & Kuhne, 2006; Nguyen et al., 2005) because the relevant information on how to best present oneself is already exploited in SJTs administered under an honest instruction (especially if administered with a knowledge or "should-do" instruction).

Similarly, and related to the first and second implication, our results may imply that SJTs can also be solved without situation descriptions (Krumm et al., 2015; Schäpers et al., 2017) if the desirable response options are also the correct options. In this case, test takers may make use of information inherent in response options (see also Leeds, 2012) regardless whether the situation descriptions are available or not.

To summarize our main contributions, the herein reported findings support the notion that response options are an important source of information in SJTs. While this may not seem surprising to many scholars, it is nevertheless insightful in light of the ongoing debate on the context-dependency of SJTs (e.g., Lievens & Motowidlo, 2016). Specifically, our study suggests that response options may convey context-*in*dependent information about their correctness in terms of their desirability and plausibility. As another key contribution, our findings show that the information inherent in response options is used regardless of the way an SJT is administered (i.e., under fake-good instructions, without further context provided by situation descriptions). From a practical perspective, we hope to draw test developers attention more closely to response options in SJTs. The current results stress the importance of presenting response options for each scenario that are balanced in desirability and plausibility. Another way to make use of the current insights when developing an SJT would be to ensure that desirability and plausibility of response options are unrelated to their effectiveness as determined by experts (which was the case in the current study). Otherwise, knowledge about the effectiveness of a given response cannot be discerned from socially desirable responding. Building on experimental test validation strategies (Bornstein, 2011; Krumm, Hüffmeier, & Lievens, 2017), we recommend that test developers conduct experimental approaches to examine whether test scores are affected by social desirability and plausibility of responses. While knowledge about desirable behavior may add to criterion-related validity (Ones, Viswesvaran, & Reiss, 1996), it is in contrast to recent trends toward a more construct driven development of SJTs (e.g., Christian et al., 2010; Lievens, 2017). Another practical takeaway would be to design situation descriptions in a way that the influence of the desirability of response options is minimized. This may be achieved by presenting

situational information, in which a (correct) response based on context-dependent knowledge is in contrast to desirable responding. Finally, we suggest that test developers could also come up with SJTs in which socially desirable responding is contingent on specifics of a situation. Note that our research examined the desirability and plausibility of responses as judged in absence of situational information. However, if the desirability of a response can only be judged considering specifics of a given situation, this may make SJTs more immune to the herein reported effects and, instead, capture relevant knowledge about how to act under such circumstances.

5.1 Limitations and future research

Several limitations must be noted. First, we deliberately used independent samples to rate the desirability and plausibility of response options and to respond to the actual SJT. While this approach avoids common source bias and potential reciprocal influences (i.e., "I find this solution effective, ergo it is also desirable"), we do not know whether test takers' own judgment of responses' desirability is related to their rating of response options. Instead, our focus was on desirability and plausibility understood as features of each response options, as identified by independent SMEs. Second, SMEs rated the desirability of each response option without being able to compare them to other response options (as included in the actual SJT item). However, socially desirable responding may also be understood as a comparative process where test takers weight alternative pieces of information against each other. We suspect that the relationships between social desirability ratings and test responses would have been even stronger if desirability had been determined through a comparison of the response alternatives included in each SJT item. As mentioned above, however, we sought to provide a conservative test of our hypothesis and therefore chose to gather social desirability ratings for each response option independently.

Another limitation may be seen in the relatively small variance in desirability and plausibility of response options. While it was our aspiration to test our hypotheses on a psychometrically sound SJT, which therefore did not include many response options that were obviously not plausible or very undesirable, this may have led to a rather conservative test of our hypotheses. In addition, we averaged correlations between response option features and effectiveness ratings across all candidates. Naturally, such an approach ignores potential differences among test takers. Not all participants may have equally used information inherent in response options to guide their effectiveness ratings. Future research might focus on candidates' attributes, such as their cognitive ability or personality, as potential moderators of the extent to which information in response options is used. This includes test takers cultural background: Cultural differences (e.g., individualism-collectivism, power distance) are known to influence response choice in SJTs (e.g., Schmitt et al., 2018) and may also determine the desirability of leadership behavior (e.g., directive vs. consensual leadership behavior).

Finally, we acknowledge that our findings were obtained from only one SJT tapping into leadership skills in the military context. Thus, it is an open question whether our findings transfer to other SJTs and contexts. However, as mentioned in the Methods section, the herein used SJT enabled a conservative test of our hypotheses. That is, we applied an SJT with design features that—based on previous research (Nguyen et al., 2005; Peeters & Lievens, 2005)—would already minimize effects of social desirability (despite other more favorable properties of SJTs with behavior tendency response instructions; Ployhart & Ehrhart, 2003). These design features were: a theory-driven and fixed set of response options, a rating response format, and a knowledge response instruction. Although further research is needed, the herein reported results may be viewed as a lower bound. We also emphasize that our SJT was developed for the military context, but otherwise comparable (e.g., in length of situation descriptions, scoring) to other SJTs (outside the military context) with the same design features (e.g., Lock et al., 2005). In terms of effects of omitting situation descriptions in SJTs, Krumm et al. (2015) and Schäpers et al., (2017) examined several design features of SJTs as potential moderators (e.g., response instruction, sample, construct domain of SJTs, presentation format). However, no significant moderators were found. We nevertheless encourage further research examining the relevance of desirability and plausibility of response options on structurally different SJTs from several construct domains that are presented with and without situation descriptions.

In addition to the above suggestions, we encourage future research that examines design features on a more fine-grained level. Specifically, the experimental manipulation of smaller parts of SJT response options could be a useful approach to shed light on the question when and why social desirability affect test takers' responses. For instance, the implementation of social desirable versus nonsocial desirable context cues could be a promising approach to investigate the magnitude of social desirability on test takers' response behavior (see Baumgarten, Süß, & Weis, 2015). Finally, future research might concentrate on other design features (e.g., trait-activating cues) that may moderate SJT response behavior. Such research might not only examine how such design features contribute to SJT performance, but also how they interact in predicting SJT performance. So, we encourage process tracing methods (e.g., eye tracking) to further scrutinize the relative importance of SJT design features on test takers' responses.

6 CONCLUSION

The current research revealed that one-third of reliable variance in test takers' effectiveness ratings in an SJT can be attributed to the desirability and plausibility of response options. Since desirability and plausibility of responses were determined independently from the SJT's situations, these findings add to the notion that response behavior may be less situational and—unlike expected from simulations—more prone to general judgments based on information inherent in response options.

ENDNOTES

1 In addition, culturally relevant content on response options may determine response choice (see Schmitt et al., 2018, for an approach similar to the one presented here); however, culture was not considered in the current study.

2 In fact, the social desirability rating, as obtained in Study 1, correlated lowly ($r = 0.14$) with the effectiveness of responses as determined by experts (see below).

3 Both groups did not differ in age, $t(114.1) = 1.187$, $p = 0.238$, but in gender ratio, $\chi^2 (1, N = 135) = 4.29$, $p = 0.04$).

4 Both groups did not differ in age, $t(93) = 1.178$, $p = 0.242$ or gender ratio, $\chi^2 (1, N = 95) = 0.073$, $p = 0.79$).

REFERENCES

Baumgarten, M., Süß, H. M., & Weis, S. (2015). The cue is the key. European Journal of Psychological Assessment, 31, 38–44. https://doi. org/10.1027/1015-5759/a000204

Borneman, M. J. (2016). Further considerations in SJT development. Industrial and Organizational Psychology, 9, 55–59. https://doi. org/10.1017/iop.2015.117

Bornstein, R. F. (2011). Toward a process-focused model of test score validity: Improving psychological assessment in science and practice. Psychological Assessment, 23, 532–544. https://doi.org/10.1037/a0022402

Brown, N. A., Jones, A. B., Serfass, D. G., & Sherman, R. A. (2016). Reinvigorating the concept of a situation in situational judgment tests. Industrial and Organizational Psychology, 9, 38–42. https://doi. org/10.1017/iop.2015.113

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. https://doi.org/10.1080/08959285.2014.929693

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. https://doi.org/10.1111/j.1468-2389.2012.00604.x

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. https://doi. org/10.1037/0021-9010.82.1.143

Chen, L., Fan, J., Zheng, L., & Hack, E. (2016). Clearly defined constructs and specific situations are the currency of SJTs. Industrial and Organizational Psychology, 9, 34–38. https://doi.org/10.1017/iop.2015.112

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. https://doi. org/10.1111/j.1744-6570.2009.01163.x

Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational judgment tests for selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore, & C. Semedo (Eds.), The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention (pp. 226–246). Oxford, UK: Wiley. https://doi.org/10.1002/9781118972472.ch11

Crook, A. E. (2016). Unintended consequences: Narrowing SJT usage and losing credibility with applicants. Industrial and Organizational Psychology, 9, 59–63. https://doi.org/10.1017/iop.2015.118

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

Fan, J., Stuhlman, M., Chen, L., & Weng, Q. (2016). Both general domain knowledge and situation assessment are needed to better understand how SJTs work. Industrial and Organizational Psychology, 9, 43–47. https://doi.org/10.1017/iop.2015.114

Felfe, J., Wunderlich, I., & Kaminski, K. (2016). TEFEM: Test zur Erfassung von Führungs- und Entscheidungsverhalten im Militär [TEFEM: Testing leadership decision making in the military]. Unpublished manual. Hamburg, Germany: Helmut-Schmidt-Universität.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. Metron, 1, 3–32.

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327–358. https://doi.org/10.1037/h0061470

Furnham, A. (1986). Response bias, social desirability and dissimulation. Personality and Individual Differences, 7, 385–400. https://doi. org/10.1016/0191-8869(86)90014-0

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

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

Harris, A. M., Siedor, L. E., Fan, Y., Listyg, B., & Carter, N. T. (2016). In defense of the situation: An interactionist explanation for performance on situational judgment tests. Industrial and Organizational Psychology, 9, 23–28. https://doi.org/10.1017/iop.2015.110

Harvey, R. J. (2016). Scoring SJTs for traits and situational effectiveness. Industrial and Organizational Psychology, 9, 63–71. https://doi. org/10.1017/iop.2015.119

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: Erlbaum.

Hooper, A. C., & Sackett, P. R. (2008, April). Self-presentation on personality measures: A meta-analysis. Paper presented at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology. San Francisco, CA.

Kanning, U. P., & Kuhne, S. (2006). Social desirability in a multimodal personnel selection test battery. European Journal of Work and Organizational Psychology, 15, 241–261. https://doi. org/10.1080/13594320600625872

Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of situational judgment tests (SJTs). European Journal of Psychological Assessment, 32, 230–240. https://doi. org/10.1027/1015-5759/a000250

Krumm, S., Hüffmeier, J., & Lievens, F. (2017). Experimental test validation: Examining the path from test elements to test performance. European Journal of Psychological Assessment, Advance Online Publication. https://doi.org/10.1027/1015-5759/a000393

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? Journal of Applied Psychology, 100, 399–416. https://doi. org/10.1037/a0037674

Leeds, J. P. (2012). The theory of cognitive acuity: Extending psychophysics to the measurement of situational judgment. Journal of Neuroscience, Psychology, and Economics, 5, 166–181. https://doi. org/10.1037/a0027294

Leeds, J. P. (2018). Applying cognitive acuity theory to the development and scoring of situational judgment tests. Behavior Research Methods, 50, 2215–2225. https://doi.org/10.3758/s13428-017-0988-1

Lievens, F. (2017). Construct-driven SJTs: Toward an agenda for future research. International Journal of Testing, 17, 269–276. https://doi.org /10.1080/15305058.2017.1309857

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. Industrial and Organizational Psychology, 9, 3–22. https:// doi.org/10.1017/iop.2015.71

Lock, M., Wheeler, R., Burnard, N., & Cooper, C. (2005). Leadership judgement indicator manual. Oxford, UK: Hogrefe.

MacCann, C., Lievens, F., Libbrecht, N., & Roberts, R. D. (2016). Differences between multimedia and text-based assessments of emotion management: An exploration with the multimedia emotion management assessment (MEMA). Cognition and Emotion, 30, 1317–1331. https://doi.org/10.1080/02699931.2015.1061.482

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., List, S. K., & Kepes, S. (2016). The "hot mess" of situational judgment test construct validity and other issues. Industrial and Organizational Psychology, 9, 47–51. https://doi.org/10.1017/iop.2015.115

McDaniel, M. A., Morgeson, F. P., Finnegan Bruhn, E., 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. https://doi. org/10.1037//0021-9010.86.4.730

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. International Journal of Selection and Assessment, 9(1/2), 103–113. https://doi. org/10.1111/1468-2389.00167

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. Journal of Applied Psychology, 85, 812–821. https:// doi.org/10.1037/0021-9010.85.5.812

Melchers, K. G., & Kleinmann, M. (2016). Why situational judgment is a missing component in the theory of SJTs. Industrial and Organizational Psychology, 9, 29–34. https://doi.org/10.1017/iop.2015.111

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. https://doi. org/10.1037/0021-9010.75.6.640

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. Journal of Applied Psychology, 91, 749–761. https://doi. org/10.1037/0021-9010.91.4.749

aemi, B., Martin-Raugh, M., & Kell, H. (2016). SJTs as measures of general domain knowledge for multimedia formats: Do actions speak louder than words? Industrial and Organizational Psychology, 9, 77–83. https://doi.org/10.1017/iop.2015.121

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

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

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), Personality assessment via questionnaire (pp. 143–165). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-3-642-70751-3_8

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

Ployhart, R. E. (2006). The predictor response process model. In J. A. Weekley & R. E. Ployhart (Eds.), Situational judgment tests. Theory, measurement, and application (pp. 83–106). Mahwah, NJ: Erlbaum.

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. https://doi. org/10.1111/1468-2389.00222

Rockstuhl, T., Ang, S., Ng, K. Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgment tests: Evidence from intercultural multimedia SJTs. Journal of Applied Psychology, 100, 464– 480. https://doi.org/10.1037/a0038098

Schäpers, P., Lievens, F., & Krumm, S. (2017, April). How situational are video-based situational judgment tests? Paper presented at the 32th annual convention of the Society for Industrial and Organizational Psychology, Orlando, FL.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley & R. E. Ployhart (Eds.), Situational judgment tests. Theory, measurement, and application (pp. 135–155). Mahwah, NJ: Erlbaum.

Schmitt, N., Prasad, J. J., Ryan, A. M., Bradburn, J. C., & Nye, C. D. (2018). Culture as a determinant of option choice in a situational judgement test: A new look. Journal of Occupational and Organizational Psychology, Advance Online Publication. https://doi.org/10.1111/joop.12251

Torres, W. J., & Beier, M. E. (2016). It's time to examine the nomological net of job knowledge. Industrial and Organizational Psychology, 9, 51–55. https://doi.org/10.1017/iop.2015.116

Vecchio, R. P. (1987). Situational leadership theory: An examination of a prescriptive theory. Journal of Applied Psychology, 72, 444–451. https://doi.org/10.1037/0021-9010.72.3.444

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

Vroom, V. H., & Yetton, P. W. (1973). Leadership and decision-making. Pittsburgh, PA: University of Pittsburgh Press.

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

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

Whetzel, D. L., & Reeder, M. C. (2016). Why some situational judgment tests fail to predict job performance (and others succeed). Industrial and Organizational Psychology, 9, 71–77. https://doi.org/10.1017/ iop.2015.120

Ziegler, M., & Horstmann, K. T. (2017). Importance of testing validity. European Journal of Personality, 31, 486–487. https://doi. org/10.1002/per.2119