Effects of Response Instructions on Faking a Situational Judgment Test

Nhung T. Nguyen*
Towson University

Michael D. Biderman
University of Tennessee

Michael A. McDaniel
Virginia Commonwealth University

A situational judgment test (SJT) and a Big 5 personality test were administered to 203 participants under instructions to respond honestly and to fake good using a within-subjects design. Participants indicated both the best and worst response (i.e., Knowledge) and the most likely and least likely response (i.e., Behavioral Tendency) to each situation. Faking effect size for the SJT Behavioral Tendency response format was ($d = .34$) when participants responded first under honest instructions and ($d = .15$) when they responded first under faking instructions. Those for the Big 5 dimensions ranged from $d = .26$ to $d = 1.0$. For the Knowledge response format results were inconsistent. Honest condition Knowledge SJT scores were more highly correlated with cognitive ability ($r = .56$) than were Behavioral Tendency SJT scores ($r = .38$). Implications for researchers and practitioners are discussed.

Introduction

Situational judgment tests (SJTs), designed to assess an applicant's response to a series of work-related scenarios, are becoming popular selection tools (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Although research on SJTs indicates that these tests have incremental predictive validity of job performance over and above cognitive ability and personality tests (e.g., Chan & Schmitt, 2002; Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001), whether these tests are fakable remains unclear. SJTs were found to be fakable in one study (Vasilopoulos, Reilly, & Leaman, 2000) using a behavioral tendency response format and not fakable in another study (Juraska & Drasgow, 2001) using the same response format. Further, no research has examined the role of response formats on faking in SJTs. One study (Ployhart & Ehrhart, 2003) examined the influence of response formats on the validity and reliability of SJTs. They found that different response formats were associated with differences in the validity and reliability of their SJT. Although it seems obvious that the fakability of SJTs would depend on how test takers are instructed to respond, no studies have addressed that issue.

The importance of studies on applicant faking depends on the impact of faking on test scores. In an employment setting, it seems that most applicants would want to achieve the highest possible score. If that desire and the ability to fake remain constant across applicants, faking merely adds a constant to everyone's score and its importance as a factor in the selection process is minimal. However, McFarland and Ryan (2000) showed that there are individual differences in faking. They also found that individuals scoring highly on the Big Five personality dimensions of conscientiousness and emotional stability faked less than those scoring low on the same dimensions. Thus, faking has the potential to affect both the reliability and validity of selection measures. However, the extent of individual differences in faking ability and faking motivation in SJTs as well as the impact of that variance on reliability and validity still remain to be assessed.

SJTs have two main types of response formats. For tests using a Best/Worst or Knowledge format, test-takers are presented a scenario and asked to choose the best and/or the worst response to that scenario or otherwise evaluate the effectiveness of the response from a list of possible responses. In the second, the Most Likely/Least Likely or Behavioral Tendency format, test-takers are asked to

*Address for correspondence: Nhung T. Nguyen, Towson University, Department of Management, 8000 York Road, Towson, MD 21252. E-mail: nnguyen@towson.edu

choose the response they would be most likely or least likely to make. For the first response format, the test taker's responses are determined by the ability to identify "correct" responses, in a fashion similar to traditional cognitive ability or job knowledge tests. For the second, most likely/least likely format, the test taker's responses are determined by his or her behavioral tendencies. In the first response format, the SJT is a knowledge test and can be expected to correlate with general cognitive ability as most job knowledge tests do. The second response format elicits self-reports of behavioral tendencies, not unlike personality tests, and should correlate with personality tests. In fact, McDaniel and Nguyen (2001) found moderate correlations of SJT scores with three of the Big Five personality dimensions (i.e., emotional stability, conscientiousness, and agreeableness) in a meta-analysis of SJT studies using a mixture of response formats. McDaniel, Hartman, and Grubb (2003) analyzed the SJT data separately by response format. SJTs using a Knowledge response format were more highly correlated with cognitive ability and less correlated with personality test scores. In contrast, SJTs using the Behavioral-tendency response format were more highly correlated with personality test scores and less highly correlated with cognitive ability.

The purpose of this study was to determine whether the ability to fake the SJT is related to response formats and to compare the generalizability of the findings with those of a personality test. A within-subjects design was used in which participants took both a SJT and a personality measure under instructions to respond honestly and under instructions to "fake good." Our research addressed three specific questions. First, we investigated the extent to which faking was related to the test response format. McDaniel and Nguyen (2001) speculated that the "Pick the best and worst answer," i.e., Knowledge response format, would be more faking resistant than the "Pick the most likely and least likely answer," i.e., Behavioral Tendency format. They suggested that the former response format assesses knowledge of respondents regarding a particular procedure, fact, or concept, which is more difficult to fake. The latter response format measures respondents' behavioral tendency.

Although personality is commonly measured by trait adjectives (e.g., Goldberg, 1990, 1992) behavioral tendencies have often been used as indicators of personality traits. Behavioral tendencies have been used to measure personality traits including leadership (e.g., Kognor & Nordvik, 2004) and workaholism (e.g., Mudrack & Naughton, 2001), to name a few. Further, the trait theory of personality defines personality as "dimensions of individual differences in tendencies to show consistent patterns of thoughts, feelings, and *actions* ... [italic added by authors]" (McCrae & Costa, 1995, p. 235). Thus, the Behavioral Tendency SJT responses should be analogous to personality test items that describe behavioral tendencies. In the present study, both response formats, a Knowledge format and a Behavioral Tendency format, were used with the same test items. Based on the hypothesis of McDaniel and Nguyen (2001), we expected greater faking for the Behavioral Tendency response format than for the Knowledge response format.

Hypothesis 1: The behavioral tendency response format in a SJT will be more fakable (i.e., yield higher test scores) than the knowledge response format.

The second question concerned the extent to which a SJT can be faked relative to a personality test. Meta-analytic findings on applicant faking have shown that respondents can raise their scores on a personality inventory by one-half standard deviation using between-subjects study designs and by about one standard deviation for within-subjects study designs (Viswesvaran & Ones, 1999; McFarland & Ryan, 2000). However, that same research on SJT fakability is lacking. Only one study has compared faking in SJT with faking in personality tests. That study found a SJT using the "most likely" response format to be less fakable than a personality inventory (Vasilopoulos *et al*., 2000). Evidence in applicant faking literature suggests that the heterogeneity of the test, a typical characteristic of SJTs (Chan & Schmitt, 1997), makes it less transparent to the test takers and thus more difficult to fake (e.g., Lautenschlager, 1994). In the present study we included an inventory of the Big Five personality dimensions along with the SJT described by Smith and McDaniel (1998) in a within-subjects design. Based on the little research comparing faking in the two types of tests, we expected to find less faking in a SJT than in personality tests.

Hypothesis 2: The SJT will be less fakable than the Big Five personality dimensions.

The third question addressed whether fakability of a SJT is a function of its cognitive loading. We explored cognitive ability as measured by the Wonderlic Personnel Test (Form A) (Wonderlic Inc., 1999). Whereas the Knowledge response format makes the SJT analogous to a job knowledge test, SJTs using that format should correlate more highly with a standard measure of cognitive ability, such as the Wonderlic, than scores from measures using the Behavioral Tendency response format. Therefore, we hypothesize:

Hypothesis 3: SJT scores based on a Knowledge response format will correlate more highly with cognitive ability than those based on a behavioral tendency response format.

Method

Setting and Participants

Two hundred and sixty-one undergraduate and graduate students from two southeastern public universities

participated in the study in exchange for partial course credit. Of these participants, fifteen left the testing room before the experiment was completed, resulting in a sample of 246 participants. The exclusion of 43 others is described later. Of the final 203 participants, the average age was 25.33 years (SD = 6.24). 86 were male and 117 were female. By race, 113 (55.7%) participants were White, 49 (24.1%) were Black, 30 (14.8%) were Asian, two (1.0%) were Hispanic, and 9 (4.4%) indicated "other."

Procedure

The battery of instruments was administered to groups of participants ranging in size from 3 to 25. Participants signed a consent form and were assigned an identification number at the beginning of the experiment. After completing the Wonderlic Personnel Test (Form A), all participants completed the SJT measures and the Big Five measures twice; once under the "faking good" instruction and once under the "honest" instructions. The order of instructions was counterbalanced with half of the participants completing the measures under the "honest" instructions first and the other half under the "faking good" instructions first. The following instructions were used.

> *Honest instructions:* You are asked to complete a two-part selection measure. The first part presents you with situations that you might experience on the job. Each situation is followed by several possible actions. Identify which action would be the *BEST ACTION* and which action would be the *WORST ACTION* to take in the situation. Then identify which action you would *MOST LIKELY* take and which action you would *LEAST LIKELY* take in the situation. The second part of the selection measure presents you with statements describing yourself. You are asked to decide how representative the statements are to your own characteristics. It is very important that you answer as honestly as possible even if you think the action and/or description is negative or unflattering. Remember that your responses will be used for research purposes only and no one will have access to your responses.

> *Faking instructions:* You are asked to play the role of a job applicant in completing this two-part selection measure. The first part presents you with situations that you might experience on the job. Each situation is followed by several possible actions. Identify which action would be the *BEST ACTION* and which action would be the *WORST ACTION* to take in the situation. Then identify which action you would *MOST LIKELY* take and which action you would *LEAST LIKELY* take in the situation. The second part of the selection measure presents you with statements describing yourself. You are asked to decide how representative the statements are to your own characteristics. Respond to this measure as if you were applying for the job of customer service representative. Some examples of customer service jobs include bank teller and call center representative. Please respond in a way that would best guarantee that you would get the customer service representative job. Remember that your responses will be used for research purposes only and no one will have access to your responses.

After completing the test battery, participants completed a background survey including typical demographic questions. At the end of the experiment, participants were debriefed and thanked for their participation. The total testing time was approximately 1.5 hours.

Measures

Demographic Questionnaire. The demographic questionnaire included questions of age, sex, race, and education level. Respondents reported their previous experience in customer service jobs on a six-category scale labeled from *No experience* to *3 or More Years experience* in customer service-related jobs.

Cognitive Ability. Cognitive ability was measured by the Wonderlic Personnel Test (Form A). The test has been used in previous research to measure cognitive ability of adults with test-retest reliabilities above .90 (Wonderlic Inc., 1999).

Situational Judgment Test. The situation judgment test used here was the Work Judgment Survey described by Smith and McDaniel (1998). The test developers sought to make the test potentially useful to a wide range of jobs. This was done in three ways. First, no item scenarios required the applicant to assume the role of a supervisor. Thus, the test could apply to non-supervisory positions. It could also be used with supervisory positions although it would not tap content unique to supervisors. Second, no scenarios mentioned equipment specific to a particular job. Third, no scenarios mentioned a specific job (e.g., salesperson). The resulting test is thus potentially applicable to a wide range of jobs. The content of the items included difficulties with one's supervisor (e.g., lack of supervisor's support, unreasonable supervisor), problems with the work tasks (e.g., insufficient resources, difficulty of work), and issues with co-workers (e.g., personality conflicts, lazy co-worker). The 31 item test was empirically-keyed and also keyed based on expert judgment. The keys correlated in the .90 s. In operational use, the empirical key has typically been used. The empirical key was used here.

For each situation, respondents were asked to select from five given courses of actions. In the knowledge instruction condition, the respondents were asked to identify the best and worst action. In the behavioral tendency condition, the respondents were asked to indicate which response they would most likely and least likely perform. Thus participants knew that they would respond twice to each SJT scenario – once using the Knowledge response format and once using the behavioral tendency format. Response sheets for both formats were in plain view as participants read the SJT scenarios. Since participants knew that both types of responses would be required,

it was decided that there was no need to counterbalance the order in which the two responses were obtained, and participants were instructed to provide the Knowledge response first, followed by the Behavioral Tendency response. We computed reliability estimates based on split-half correlations for the SJT measures using the Spearman-Brown prophesy formula (e.g., DeVellis, 2003). Combining the two orders of instructions, for the Knowledge instruction, odd item-even item split-half estimates of reliability were .77 and .76 for the Honest and Fake Good instructional conditions respectively. For the Behavioral Tendency response format, split-half estimates respectively for the Honest and Fake Good instructional condition were .78 and .82. As noted by Chan and Schmitt (1997), the multidimensionality of SJTs may cause alpha to underestimate true reliability. As a lower bound estimate of reliability, the alphas were .67 and .68 for the Knowledge response format and .74 and .78 for the Behavioral Tendency response format.

Goldberg's Big 5 Personality Measure. The Big 5 personality inventory developed by Goldberg (http://ipip.ori.org/ipip) was used in this study. The inventory we used consists of 50 items taken from a large pool of personality items available for public use on the International Personality Item Pool web site. The inventory measures personality dimensions of the Big 5 model. Each dimension was measured by 10 Likert-type items. Respondents were asked to indicate the accuracy of each item as a descriptor of them with a number ranging from 1 (*very inaccurate*) to 5 (*very accurate*). The scales have been validated against other established scales (e.g., NEO-PI) and shown to have good reliabilities (Goldberg, 1999, in press). Internal consistency estimates of reliability ranged from .74 to .90, consistent with previous meta-analytic estimates of reliability coefficients for Big 5 personality measures (Viswesvaran & Ones, 2000), and are reported in Table 1.

Faking Ability. For tests of Hypothesis 1 involving only the SJT performance, raw SJT scale scores were used, with differences in means under the Faking and Honest conditions defining faking ability. Although criticism against the use of difference scores as a measure of change has been widely discussed (e.g., Edwards, 1995; Edwards, 2002), we think that its use is justified when individual differences in true change (i.e., faking) exists (McFarland & Ryan, 2000; Rogosa & Willett, 1983). For tests of Hypothesis 2 involving comparisons across SJT and Big 5 dimensions, faking ability was measured by creating a standardized difference score for each participant on each variable using the following formula:

$$d = (X_F - X_H)/\mathrm{SD}.$$

In the formula, d is the standardized difference score for a participant, and X_F and X_H are the participant's scores on a measure under Faking and Honest instructions respectively. The SD term is a doubly pooled standard deviation obtained by first computing the pooled variance of X_F across order conditions and X_H across order conditions, then averaging the X_F and X_H pooled variances and finally taking the square root to create SD. The computation was based on the discussion of Dunlap, Cortina, Vaslow, and Burke (1996) of effect sizes for correlated designs. Since these variables represent standardized effect sizes, they allowed comparisons across the SJT and personality measures, just as those in comparisons based on mean effect sizes. These variables are analogous to the effect size variables whose means were reported by McFarland and Ryan (2000; Table 1) except that we first pooled variance across order conditions prior to averaging across H and F conditions.

Manipulation Check. Two items were created to determine if participants attended to the instructions. The questions asked participants to indicate whether they responded honestly under the honest condition and tried to present their best image under the faking good condition. Two hundred and twenty-one participants (90%) correctly responded to both manipulation-check items. Those who incorrectly responded to the manipulation check items ($N = 25$) were excluded from further analyses. Additionally, 18 cases were dropped because respondents failed to complete all the measures or they made errors in recording their identification numbers that prevented us from correctly matching their data from various measures. The final sample included 203 cases.

Results

Because several differences associated with order of conditions were found, Table 1 presents descriptive statistics for the various measures separately for each order along with statistics computed for the two orders combined. To test the first hypothesis – that of the relationship of faking to response format – a three-way analysis of variance ($2 \times 2 \times 2$) with two repeated measures factors – Faking instruction (Honest vs. Faked) and Response Format (Knowledge vs. Behavioral Tendency) and one between subjects factor – Order (Honest → Faking vs. Faking → Honest) was conducted. Because the three-way interaction in this analysis was significant ($F(1, 201) = 35.386$, $p < .001$, $\eta^2 = .150$) inspection of the pattern of means suggested that a more enlightening alternative approach would be to analyze the data of the two response formats separately.

Analyses of variance were conducted separately for the Knowledge and for the Behavioral Tendency response formats, with Honest vs. Faking instructions as a repeated measures factor and Order as a between-subjects factor in each. The results of these two analyses were clear. For the Knowledge instruction format, cell means formed a nearly perfect classic "crossover" interaction ($F(1, 201) = 36.541$,

Table 1. Descriptive statistics and intercorrelation among variables in the study

H → F order			$N = 99$														
F → H order			$N = 104$														
Variables	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Knowledge SJT_H	20.72	4.96	.70	.43	.15	.26	.29	.06	.26	.51	.38	.29	.35	.35	.25	.28	.46
	18.50	5.90	.80														
2. Behavioral SJT_H	14.38	6.45	.48	.78	.11	.13	.27	.19	.30	.35	.60	.18	.24	.28	.23	.28	.32
	13.03	6.68	.37	.77													
3. Extraversion_H	3.99	8.28	.02	.07	.90	.24	.13	.25	.30	.09	.11	.49	.11	.06	.14	.21	.00
	4.03	8.57	.26	.14	.90												
4. Agreeableness_H	14.20	4.77	.17	.09	.27	.79	.29	.06	.26	.16	.11	.18	.50	.19	.14	.16	.05
	13.81	5.29	.31	.15	.21	.79											
5. Conscientiousness_H	16.02	6.30	.30	.29	.04	.31	.82	.27	.44	.18	.16	.22	.29	.53	.24	.35	.15
	15.46	6.60	.27	.26	.22	.27	.85										
6. Emotional stability_H	−14.9	8.76	.06	.22	.21	−.00	.19	.89	.22	.06	.00	.15	.03	.12	.44	.11	.01
	−13.6	8.20	.09	.18	.28	.12	.37	.88									
7. Openness_H	20.24	5.37	.23	.30	.24	.22	.42	.23	.76	.20	.23	.31	.31	.37	.26	.52	.38
	20.17	5.27	.30	.30	.36	.29	.46	.21	.75								
8. Knowledge SJT_F	18.87	5.93	.62	.49	−.04	.08	.26	.02	.15	.80	.51	.30	.39	.32	.25	.29	.38
	20.96	4.96	.53	.27	.24	.26	.13	.07	.27	.69							
9. Behavioral SJT_F	16.67	6.45	.46	.55	.05	.11	.12	.06	.10	.76	.75	.29	.29	.34	.32	.32	.35
	14.06	7.39	.29	.63	.16	.10	.17	−.02	.35	.40	.86						
10. Extraversion_F	9.19	7.26	.29	.23	.27	.24	.27	.19	.29	.35	.32	.86	.48	.47	.56	.58	.29
	6.56	7.81	.25	.11	.70	.13	.17	.15	.33	.35	.23	.89					
11. Agreeableness_F	15.85	5.40	.37	.24	.05	.38	.34	.11	.32	.41	.36	.67	.84	.57	.40	.53	.34
	15.68	4.32	.35	.24	.17	.63	.23	−.08	.31	.39	.23	.30	.72				
12. Conscientiousness_F	20.36	6.67	.34	.33	−.04	.21	.53	.14	.35	.41	.33	.55	.69	.91	.61	.68	.33
	18.73	6.34	.33	.22	.16	.16	.53	.12	.39	.28	.32	.38	.43	.87			
13. Emotional stability_F	−6.76	7.98	.29	.29	.05	.11	.30	.45	.27	.35	.38	.56	.57	.69	.88	.60	.30
	−9.50	7.71	.17	.15	.23	.16	.19	.47	.26	.24	.22	.53	.20	.52	.87		
14. Openness_F	23.17	6.26	.28	.26	.13	.16	.44	.17	.42	.31	.30	.60	.60	.72	.65	.83	.43
	21.63	5.24	.26	.28	.30	.16	.26	.05	.65	.35	.33	.55	.43	.63	.54	.76	
15. Cognitive ability	25.61	6.39	.56	.38	−.14	−.00	.28	.12	.35	.41	.34	.30	.39	.40	.34	.27	–
	23.66	6.98	.37	.25	.12	.09	.04	−.07	.41	.45	.32	.26	.31	.25	.22	.58	–
Overall mean			19.58	13.69	4.01	14.00	15.73	−14.3	20.2	19.94	15.3	7.84	15.8	19.5	−8.2	22.4	24.61
Overall SD			5.56	6.59	8.41	5.04	6.45	8.48	5.31	5.54	7.05	7.64	4.86	6.54	7.94	5.80	6.75
Overall alpha			.77	.78	.90	.79	.84	.89	.75	.76	.82	.88	.79	.89	.88	.80	–

Note: Correlations between variables for order of instructions are shown in the lower triangle and for the whole sample in the upper triangle. Reliability estimates are shown underlined along the diagonal. SJT reliabilities are based on odd item – even item split half correlations.

$p < .001$, $\eta^2 = .154$). In each condition, the mean score for the first condition to which participants were exposed was greater than that for the second condition. In the Honest-Faked order, the mean score for the honest condition was larger than the mean for the Faking condition. But in the Faked → Honest order, the mean score for the Faking condition was larger than the mean for the Honest condition. Neither main effect was significant. This means that there was what would ordinarily be called a faking effect only for participants whose order of conditions was F → H. However, for those whose order was H → F, the difference between Honest and Faked instructions was negative, with higher scores in the honest condition.

For the Behavioral Tendency response format, there was no such crossover. The main effect of Honest vs. Faked Instruction was significant with means under the Faked instruction larger than in the Honest instruction ($F(1,201) = 14.913$, $p < .001$, $\eta^2 = .069$). The interaction of Instruction and Order was not significant ($F(1,201) = 2.138$, $p > .05$, $\eta^2 = .011$). There was also a main effect of Order – those participants who received the H → F order had higher mean scores under both instruction conditions than those receiving the F → H order ($F(1,201) = 5.479$, $p < .05$, $\eta^2 = .027$).

Taken together, these results indicate that participants are able to fake performance in a SJT in which a behavioral tendency response format is used. However, for a knowledge response format, the difference between performance under Honest and Faking instructions appears to depend on the order in which the instructions are received. Hypothesis 1 was supported for the Behavioral Tendency response format.

To test the second hypothesis, a standardized difference score was computed for each participant for each of the seven variables (i.e., five personality dimensions, Knowledge, and Behavioral Tendency SJT performance). Descriptive statistics on faking performance of the above seven variables as well as their reliabilities are presented in Table 2. The reliabilities of standardized difference faking scores were computed using the formula proposed by Rogosa and colleagues (Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1983). Inspection of means in the table shows that, as would be expected from the above analyses of the SJT measures, the mean effect size for the

Table 2. Correlation matrix of standardized faking effect scores for each order (lower triangle) and for whole sample (upper triangle)

H → F order						$N = 99$					
F → H order						$N = 104$					
Variables	Mean	SD	1	2	3	4	5	6	7	8	9
1. Knowledge SJT	−.34	.88	.38	.24	.07	.14	.07	.01	.07	−.08	.39
	.45	.97	.48								
2. Behavioral SJT	.34	.91	.38	.48	.13	.09	.21	.31	.13	.06	−.10
	.15	.90	.24	.52							
3. Extroversion	.65	1.18	.15	.10	.84	.42	.33	.45	.38	.27	−.17
	.32	.80	.16	.15	.65						
4. Agreeableness	.33	1.14	.18	.11	.50	.71	.40	.26	.33	.28	.02
	.38	.85	.11	.07	.29	.37					
5. Conscientiousness	.67	.97	.15	.17	.31	.44	.72	.50	.44	.19	−.09
	.50	.97	.09	.22	.36	.36	.70				
6. Emotional stability	1.00	1.08	.15	.26	.38	.30	.42	.79	.45	.26	−.23
	.51	1.00	.07	.33	.52	.24	.57	.76			
7. Openness	.53	1.14	.14	.21	.34	.34	.37	.39	.66	.09	−.14
	.26	.80	.14	−.01	.41	.34	.53	.51	.31		
8. Cognitive ability	25.61	6.39	−.07	−.04	.35	.37	.15	.19	−.03	–	−.14
	23.66	6.98	.00	.12	.15	.21	.21	.28	.20	–	
9. Test order	1	0									
	2	0									
Overall mean			.07	.24	.48	.36	.58	.75	.39	24.61	1.51
			(.11)	(.36)	(.54)	(.47)	(.69)	(.84)	(.53)		
Overall SD			1.01	.91	1.02	1.00	.97	1.07	.99	6.75	.50
Overall reliability			.44	.45	.79	.58	.71	.80	.54		

Note: Reliability estimates are shown underlined on the diagonals. Estimates within parentheses are effect sizes corrected for measurement error.

Knowledge response format for the H → F order was negative and nearly equal though opposite in sign of that from the F → H order. Because the Knowledge response format faking effects were so clearly dependent on the order in which the Honest/Faking conditions were received, we decided not to include the Knowledge response format data in subsequent comparisons. The means of the remaining standardized effect size variables were then compared using a repeated measures analysis of variance with the SJT Behavioral Tendency measure and the five personality dimensions as six levels of a repeated measures factor and Order as a between-subjects factor. Since the focus of this analysis was on a comparison of fakability of the SJT measure with the Big 5 measures, the repeated measures factor was decomposed into a set of dummy variable contrasts comparing each personality dimension to the SJT measure. The ANOVA revealed that there were differences in mean faking effect size across measures ($F(5, 197) = 9.968$, $p < .001$, $\eta^2 = .040$). Tests of the individual contrasts revealed that the size of the faking effect for the SJT measure was smaller than that for Extroversion ($p < .01$, $\eta^2 = .035$), Conscientiousness ($p < .001$, $\eta^2 = .077$) and Emotional Stability ($p < .001$, $\eta^2 = .162$) and perhaps smaller than that of Openness ($p < .10$, $\eta^2 = .014$). Although the mean faking effect scores for all measures were positive, faking was greater among those participants who received the H → F order ($F(1, 201) = 6.594$, $p < .05$, $\eta^2 = .032$). A significant Measure by Order interaction indicated that the size of this order effect varied across measures ($F(5, 197) = 2.308$, $p < .05$, $\eta^2 = .011$). To elaborate on the Measure × Order interaction, *t*-tests of simple effects revealed that the advantage of the H → F order over the F → H order was the greatest for Emotional Stability ($p < .01$, $\eta^2 = .054$), and Extroversion ($p < .05$, $\eta^2 = .027$). These results are presented in Figure 1. Hypothesis 2 thus received mixed support: mean faking on the Behavioral Tendency SJT was less than that of three of the Big 5 personality dimensions.

To test the third hypothesis concerning moderation of the correlation of SJT scores with cognitive ability by response format we computed correlations of cognitive ability with Knowledge and Behavioral tendency SJT scores and then compared those correlations using a test of related correlations (Cohen & Cohen, 1983, p. 57). Because the crossover interaction found for the Knowledge response format suggested that participant performance in that response format decreased from the first to the second administration regardless of order, we restricted this analysis to the first SJT response format encountered by participants in each order. From the H → F order, we compared the cognitive ability – honest Knowledge SJT correlation with the cognitive ability – honest Behavioral Tendency SJT correlation. Then from the F → H order we compared the cognitive ability – faked Knowledge SJT correlation with the cognitive ability – faked Behavioral

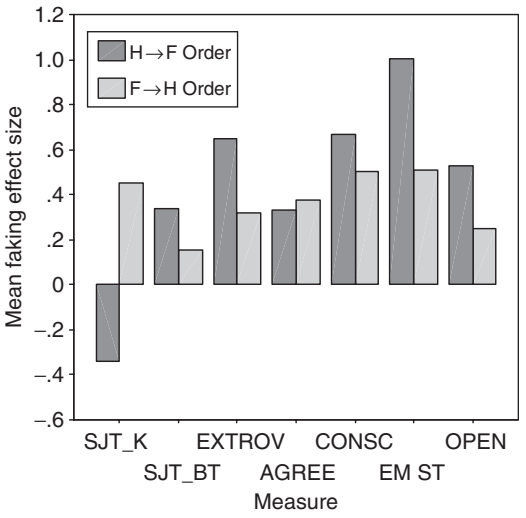

Figure 1. Mean faking effect sizes for the SJT knowledge, SJT behavioral tendency and Big 5 measures for each order.

Tendency SJT correlation. The second SJT score in each order was not used in either analysis. Results are presented in Table 3. All correlations with cognitive ability were positive and significantly different from zero. The comparison indicated that the cognitive ability – honest Knowledge SJT correlation of .555 was significantly larger than the cognitive ability – honest Behavioral Tendency SJT correlation of .375 ($t = 2.08$, df = 96, $p < .05$). However, the cognitive ability – faked Knowledge SJT correlation of .445 was not significantly different from the cognitive ability – faked Behavioral Tendency SJT correlation of .321 ($t = 1.27$, df = 101, $p > .05$). These results make a tentative case for the stronger relationship of cognitive ability to SJT performance for the Knowledge response format than the Behavioral Tendency format. Hypothesis 3 was thus supported.

We further explored the correlates of SJT performance under two response formats and the Big 5 personality dimensions using *t*-tests based on correlated *r*s like those above. The results are also presented in Table 3. Inspection of Table 3 reveals that SJT scores in both the Knowledge and Behavioral Tendency response formats were significantly correlated with openness in both the honest and faked conditions and with conscientiousness in the honest condition. None of the differences between correlations with Knowledge SJTs and those with Behavioral Tendency SJTs reached statistical significance (note that the statistical power for these comparisons was low). These results did not support any expectation that personality characteristics might be better reflected under the Behavioral tendency response format. Although the SJT used here is related to conscientiousness and openness to experience regardless of response format, we found no significant differences in correlations with the Big 5 measures as a function of the response formats.

Table 3. Results of cognitive and big 5 measure saturation in SJT performance

Cognitive ability correlates of SJT performance by response format

	Knowledge *r*	Behavioral tendency *r*	*t*(df)	*p*	80% CI	Power[1]
Condition						
Honest	.555	.375	2.077 (96)	.040	.068–.292	.53
Faked	.445	.321	1.271 (101)	.207	−.002 to .250	.19

Personality correlates of SJT performance by response format

	Knowledge *r*	Behavioral tendency *r*	*t*(df)	*p*	80% CI	Power
Honest condition						
Extraversion	.019	.074	−.531 (96)	.597	−.189 to.079	.08
Agreeableness	.172	.091	.791 (96)	.431	−.051 to .213	.12
Conscientiousness	.303*	.285*	.184 (96)	.855	−.109 to 145	.05
Emotional stability	.063	.221	−.1.560 (96)	.122	−.289 to .027	.34
Openness	.233*	.295*	−.626 (96)	.533	−.190 to .066	.09
Faked condition						
Extraversion	.237	.159	−.734 (101)	.465	−.059 to .215	.10
Agreeableness	.260	.103	1.483(101)	.141	.020 to .294	.28
Conscientiousness	.127	.171	−.409 (101)	.684	−.183 to .095	.06
Emotional stability	.070	−.021	.835 (101)	.406	−.050 to .232	.12
Openness	.274*	.353*	−.776 (101)	.440	−.210 to .052	.11

Notes: *Significant at $p < .05$.
[1]Power to detect a population difference equal in magnitude to the sample difference.

Discussion

Despite the fact that SJTs have been increasingly used in practice as a tool to screen job applicants, relatively little is known concerning the extent to which such tests are fakable. The purpose of this study was to explore the role of response instructions in SJT performance: how applicant faking of the exact same test items might be related to response instruction. The results presented here indicated that a SJT can be faked under the Behavioral tendency response format with faking effect size between .15 and .34. This was consistent with previous research (e.g., Vasilopoulos *et al*., 2000) utilizing a different SJT but with the same Behavioral tendency format.

With respect to the Knowledge response format, our results are more complicated. We found a positive F–H difference only for participants who received the "Faking good" instruction before the "Honest" instruction. Those who received the Honest → Faking order actually showed a negative F − H difference, almost equal in absolute value to the positive effect size of the participants who received the Faking → Honest order. One explanation for our results is that all participants did their best the first time they took the test. Those in the honest condition had no reason to falsify their responses, so they responded to the best of their ability. Similarly, those in the fake first condition had even greater reason to respond to the best of their ability. Then, when faced with the task again and the instruction to do something differently, to fake if the first condition was the honest condition, or to respond honestly if the first condition was the "fake good" condition, apparently the choice of participants was to change some of the responses to questions that to which they had already responded to the best of their ability. The result was that scores in the second condition were lower than scores in the first for both orders. We think the implication of this is that in practice, SJT scores obtained under the knowledge response format could be treated as being relatively immune from faking. Essentially the scores on the SJT knowledge response format will be as high as the applicant can make them regardless of the inclination to fake. However, without collecting data to test this possibility (e.g., probing participants during de-briefing to see how they altered their answers across response formats), our interpretation of the finding here should be considered tentative.

In line with the above discussion, it is our speculation that in case applicants are allowed to retest for the same job where they are given the same SJT, the order effect found here might translate into applicants' Knowledge SJT scores being unchanged on the retest regardless of faking, whereas scores on the behavioral tendency SJT might improve – if applicants responded honestly on the first administration and then decided to fake on the retest. Specifically, we speculate that if the applicant responded honestly the first time and faked the second time, his/her score on the behavioral tendency SJT might reflect the perception of what the best answers are, thus, the retest score might approximate the knowledge SJT score. In the situation where applicants apply for different jobs and take similar SJTs, we think that scores on retaking similar SJTs using the Knowledge response format for different jobs might not change regardless of faking. In the same situation, practitioners should certainly be cautious of using the Behavioral Tendency response format, because of its susceptibility to faking.

We did not collect criterion data in this study so we cannot provide evidence concerning the impact of faking on criterion-related validity of the SJT. Peeters and Lievens (2005) found that faking reduced the criterion-related validity of a SJT for undergraduate admissions from $r = .33$ to $r = .09$. But it is easy to envision situations in which faking might enhance such validity. For example, if faking ability is related to cognitive ability, as suggested in recent studies (e.g., Biderman & Nguyen, 2004; Wrensen & Biderman, 2005), then it is possible that SJT scores enhanced by faking would be more valid than SJT scores alone for criteria for which cognitive ability was salient. In the only study that investigated the effect of SJT response formats to criterion-related validities, Ployhart and Ehrhart (2003) found that the Behavioral Tendency response instruction consistently yielded higher criterion-related validity coefficients than did the Knowledge response instruction. This suggests that the impact of faking will depend on the relationship of characteristics of faking ability to characteristics of the criterion and on the nature of the response format used in the SJT.

We believe that having participants give the Knowledge response first followed by the Behavioral Tendency response had little impact on our essential findings. Specifically, it cannot be argued that the failure to find a consistent faking effect under the Knowledge response format was due to some carry-over effect, since the Knowledge response was provided first. Moreover, the finding of a consistent faking effect under the Behavioral Tendency format was in line with the one study that we are aware of using that format (e.g., Vasilopoulos *et al*., 2000) and was not related to the pattern of results obtained under the Knowledge response format.

Comparison of the Behavioral Tendency SJT faking effect sizes with those of the personality dimensions suggested that the SJT was certainly not more fakable than the Big 5 personality measures used in this study and might be less fakable than the extroversion, conscientiousness, and emotional stability measures. In spite of the faking effect sizes being not larger than those for the Big 5 scales, the fact that they were all positive suggests that practitioners wishing to have a measure completely immune to faking will wish to avoid the Behavioral Tendency response format. Further, we found that the magnitudes of the faking effect sizes for the Big 5 measures used here were consistent with previous meta-analysis of personality faking estimates. For example, Viswesvaran and Ones' (1999) meta-analysis showed the greatest amount of faking for emotional stability ($d = .93$, $N = 921$) and the least amount for agreeableness. Our results generally agreed with this pattern of findings.

Although the order effect was not significant in the analysis of the Behavioral Tendency SJT performance only, we did find a significant main effect of order across the SJT and Big 5 measures in the repeated measures ANOVA of standardized difference scores. In that analysis the main effect of order reflected that faking was generally larger in the Honest → Faking order than in the Faking → Honest order suggesting that having taken the test once made participants better able to fake. McFarland and Ryan (2000) conducted separate ANOVAs for each of seven dependent variables and found no significant differences. Since the sample effect size in our analysis was not large (Partial eta-squared = .032) it is possible that this effect is so small that it will exceed the threshold of significance in some studies and not in others. It is also possible that procedural differences between studies, including the time between administrations of the two measures and the number and kind of other tests given between the two administrations, may affect the size of the order effect.

Our expectation that the SJT scores under the Knowledge response format would have greater cognitive loading than those under the Behavioral Tendency response format was confirmed only when participants responded under the honest instructions first. The correlations with cognitive ability were small, suggesting that scores on the SJTs clearly were affected by factors other than cognitive ability. Other factors contributing to SJT performance include job knowledge and job experience (e.g., Vasilopoulos *et al*., 2000). In the Faking instruction condition, participants' attempts to fake also represent an extraneous variable that may have affected the differences in proportion of variance related to cognitive ability. Relationships to unmeasured variables such as these likely distort the comparisons made here. Although it would not have been possible to measure job knowledge for this study, we did have a self-reported measure of experience in customer service jobs. We repeated the comparisons of correlations reported above, this time using correlations partialling out job experience. The results of the re-analysis were essentially identical to those using zero-order correlations, indicating that this issue is as yet unresolved.

Finally we note that all the correlations between the six standardized difference scores (behavioral tendency

SJT and the Big 5 measures) were positive, with median correlation equal to .33. This is in agreement with the finding of McFarland and Ryan (2000). These two results support the notion that individual differences in faking ability as measured by difference scores generalize across measures. Although beyond the scope of the present paper (see Snell, Sydell, & Lueke, 1999), the possibility of the existence of a faking ability construct needs to be addressed in future research.

Limitations of the Current Study

Several limitations in this study should be noted. First, using participants who were college students limits the extent to which the findings can be generalized to an actual job applicant sample, since there may be differences in motivation to perform well between students and actual job applicants. Specifically, actual job applicants may have more motivation to "fake good" because their stakes are higher. Therefore, to the extent that actual applicants manipulate their responses through faking, the faking effect sizes reported here might be different from what may be found in a real-life applicant sample (e.g., Hough, 1998; Rosse, Stecher, Levin, & Miller, 1998). Second, the order effect we unexpectedly found restricted the interpretation of the findings to the honest – faked condition only. Last, we note that the faking effect sizes of SJT might be considered very small or small based on commonly accepted convention.

Conclusion

This study contributes to the applicant faking literature in general and SJT literature in particular. As we discussed earlier, the little previous research that existed was mixed concerning the fakability of SJTs. Also, little was known about the possibility of effects associated with the order of test sessions. We showed that both the response instructions in SJTs and order of test session might be responsible for those mixed findings. When respondents were asked to respond using a Knowledge response format, one hypothesis consistent with our results is that applicants respond to the best of their ability, regardless of the pressure to fake. Our results also are consistent with the hypothesis that the Knowledge response format is more cognitively loaded than the Behavioral Tendency response format in the absence of pressure to fake, although that cognitive load may be diluted when faking is present. All in all, this study showed that SJTs can be faked and the degree of faking varied as a function of the response format with the Knowledge response as more resistant to faking. By showing the conditions under which faking in SJTs might be expected, this study provides some guidance to practitioners considering use of these tests. Although how much applicants do fake and will fake in actual selection settings is not known, we hope this study will stimulate future research to address the effects of instruction differences on applicant faking and on the relationship of faking in SJTs vs. personality tests.

Acknowledgements

This research was based on the first author's doctoral dissertation completed at Virginia Commonwealth University. We'd like to thank the Associate Editor and two anonymous reviewers for their helpful comments on an earlier version of this manuscript.

Appendix A

Sample SJT items taken from Smith (1996).

Your work is shared with a co-worker. You work every afternoon and the co-worker works every morning. The co-worker is not doing a fair share of the work and as a result you have too much to do in the afternoon.

a. Ask the boss to handle it.
b. Talk with the co-worker and demand that the co-worker do more work.
c. Decrease the amount of work you do.
d. Try to have a friendly, non-threatening meeting with the co-worker to divide the tasks.
d. Ask the boss to assign a different co-worker to you.

References

Biderman, M.D. and Nguyen, N.T. (April, 2004) Structural equation models of faking ability in repeated measures designs. Poster presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

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

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

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

Cohen, J. and Cohen, P. (1983) *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

DeVellis, R.F. (2003) *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.

Dunlap, W.P., Cortina, J.M., Vaslow, J.B. and Burke, M.J. (1996) Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, **1**, 170–177.

Edwards, J.R. (1995) Alternatives to difference scores as dependent variables in the study of congruence in organizational research. *Organizational Behavior and Human Decision Processes*, **64**, 307–324.

Edwards, J.R. (2002) Alternatives to difference scores: Polynomial regression analysis and response surface methodology. In

F. Drasgow and N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis*. San Francisco, CA: Jossey-Bass.

Goldberg, L.R. (1990) An alternative "Description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, **59**, 1216–1229.

Goldberg, L.R. (1992) The development of markers for the Big-Five factor structure. *Psychological Assessment*, **4**, 26–42.

Goldberg, L.R. (1999) A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt and F. Ostendorf (Eds.), *Personality Psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L.R. (in press) The comparative validity of adult personality inventories: Applications of a consumer-testing framework. In S.R. Briggs, J.M. Cheek and E.M. Donahue (Eds.), *Handbook of adult personality inventories*. New York: Plenum.

Hough, L.M. (1998) Effects of intentional distortion in personality measurement evaluation of suggested palliatives. *Human Performance*, **11**, 209–244.

http://ipip.ori.org/ipip. International personality item pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences.

Juraska and Drasgow, F. (2001) Faking in a situational judgment test. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Kognor, H. and Nordvik, H. (2004) Personality traits in leadership behavior. *Scandinavian Journal of Psychology*, **45**, 49–54.

Lautenschlager, G.J. (1994) Accuracy and faking of background data. In G.S. Stokes, M.D. Mumford and W.A. Owens (Eds.), *The biodata handbook: Theory, research, and applications (*pp. 391–419). Palo Alto, CA: Consulting Psychological Press.

McCrae, R.R. and Costa, P.T. Jr. (1995) Trait explanations in personality psychology. *European Journal of Personality*, **9**, 231–252.

McDaniel, M.A., Hartman, N.S. and Grubb, W.L. III. (2003, April) Situational judgment tests, knowledge, behavioral tendency, and validity: A meta-analysis. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology. Orlando.

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

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

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

Mudrack, P.E. and Naughton, T.J. (2001) The assessment of workaholism as behavioral tendencies: Scale development and preliminary empirical testing. *International Journal of Stress Management*, **8**, 93–111.

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

Ployhart, R.E. and 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.

Rogosa, D.R. and Willett, J.B. (1983) Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, **20**, 335–343.

Rogosa, D.R., Brandt, D. and Zimowski, M. (1982) A growth curve approach to the measurement of change. *Psychological Bulletin*, **90**, 726–748.

Rosse, J.G., Stecher, M.D., Miller, J.L. and Levin, R.A. (1998) The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, **83**, 399–406.

Smith, K.C. (1996) *A situational Judgment Test: Criterion and construct validity evidence*. Paper presented at the annual meeting of the International Personnel Management Association Assessment Council, Boston, MA.

Smith, K.C. and McDaniel, M.A. (April, 1998) Criterion and construct validity evidence for a situational judgment measure. Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Snell, A.F., Sydell, E.J. and Lueke, S.B. (1999) Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review*, **9**, 219–242.

Vasilopoulos, N.L., Reilly, R.R. and 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.

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

Viswesvaran, C. and Ones, D.S. (2000) Measurement error in "Big Five Factors'" personality measurement. *Educational and Psychological Measurement*, **60**, 197–210.

Wonderlic Inc. (1999) *Wonderlic's personnel test manual and scoring guide*. Libertyville, IL: Wonderlic.

Wrensen, L.B. and Biderman, M.D. (April, 2005) Factors related to faking ability: A structural equation model application. Poster presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.