Predicting Performance on a Situational Judgement Test: The Role of Communication Skills, Listening Skills, and Expertise

Jeanne Strahan (jestrahan@bigpond.com)
Department of Psychology
University of Southern Queensland, Toowoomba QLD 4350 Australia

Gerard J. Fogarty (fogarty@usq.edu.au)
Department of Psychology
University of Southern Queensland, Toowoomba QLD 4350 Australia

M. Anthony Machin (machin@usq.edu.au)
Department of Psychology
University of Southern Queensland, Toowoomba QLD 4350 Australia

NB: The correct APA citation for this paper is: Strahan, J., Fogarty, G.J., Machin, A.M. (2005). Predicting performance on a situational judgement test: The role of communication skills, listening skills, and expertise. *Proceedings of the 40th Annual Conference of the Australian Psychological Society*, pp. 323-327, Sydney, Australia.

Abstract

Situational judgement tests (SJTs) present scenarios drawn from a work context and ask respondents to select the most appropriate response from among a range of options. They therefore attempt to assess aspects of social cognition that are not often measured in traditional selection batteries. The present study asked participants to complete a high-fidelity SJT constructed by staff at ETS to assess communication skills in a medical context and also to complete three self-report instruments assessing listening skills, communication skills, and personality. A total of 107 participants completed the computer-administered test battery. Results indicated that listening skills, communication skills, and the agreeableness personality dimension combined to predict 22% of the variance in performance on the SJT. An expert group of doctors and nurses performed better on the SJT than a group drawn from occupations outside the health area. However, problems were noted with internal consistency reliability estimates for the SJT, suggesting that the effects noted above are underestimates of the true relationships. We conclude that if these problems can be overcome, SJTs have the potential to contribute to the selection of health professionals.

Introduction

Tests of ability, personality, interests, and values have been used for almost a century to assist with selection decisions. However, for some time now researchers and practitioners have been aware that these tests do not assess the wider domains of interpersonal and communication skills. There is a need to develop and validate instruments that can be used to assess this aspect of cognitive functioning. Situational judgement tests (SJTs) are one proposed solution to this problem. SJTs present vignettes of simulated or actual interactions and ask the respondent to indicate from among a set of possible responses which one is the most appropriate in the situation shown. Unlike many selection tests, SJTs are not intended to be unidimensional. Job situations are complex and the behaviours that are sampled for inclusion in SJTs are therefore usually complex themselves. The question then arises as to whether there are identifiable skills that account for reliable variance on SJTs. The present study set out to examine the contribution of communication skills, personality, and expertise to an SJT developed to assist with the selection of medical students in the US.

SJTs are psychometric instruments designed to assess an individual's judgment concerning work-related situations (Chan & Schmitt, 2002; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Typically, the SJT is created from an analysis of a job and aspects of interest in the job, and is then used for employee selection and/or assessment of job performance. SJTs have been around since the 1920's at which time they were used to assess judgement in social situations (McDaniel et al., 2001). In World War II they were used to assess soldiers' ability to draw on common sense, experience, and general knowledge to respond to different scenarios (McDaniel et al., 2001). In the 1950s and 60s, their use was extended to predict, as well as assess, managerial success (McDaniel et al., 2001). Their use has increased greatly in the past two decades, driven by concerns about adverse impact and the growing interest in constructs such as emotional intelligence and practical intelligence. There is also an element of dissatisfaction in the general public with traditional psychometric testing, making face validity of any testing measure increasingly important. Related to both these reasons for increased interest in SJTs is the fact that current technological knowledge, in particular videos and computers, allows for a very realistic portrayal of real-life situations.

However, despite the long history and renewed interest in SJTs, there are problems associated with this method of testing. The most intractable problem relates to the scoring of answers. Attempts to address this issue include expert-novice differences, where an item is scored in the direction favoring the experts after the average ratings of experts and novices on each item are compared; expert judgment, where a team of experts decides the best answer to each question; target scoring, where the test author determines the correct answer; and consensual scoring, where a score is allocated to each option according to the percentage of people choosing that option (MacCann, Roberts, Matthews & Zeidner, 2004). From their discovery that expert scores correlate highly with consensus scoring (*r* = .89 to .99), Legree and colleagues (Legree, Psotka, Tremble, & Bourne, 2004) made a strong case for the use of consensual scoring in the knowledge domains where recognized experts do not always exist. One of the aims of the present study was to compare expert scoring and consensus scoring systems. Following Legree et al., it was expected that the two would yield similar outcomes.

As mentioned above, given the complexity of most SJTs, a further aim of this study was to identify sources of individual differences that contribute to performance on SJTs. To some extent, that will depend on the type of SJT used. In developing a SJT for civil servants, Chan and Schmitt (2002) identified two domains of performance that are found in any job; the technical domain (how the person performs tasks) and the contextual domain (how the person uses interpersonal skills). Some SJTs attempt to capture the performance of individuals across both these domains, whereas others emphasise only one. The present study asked participants to complete a high-fidelity SJT set in a medical context. The emphasis was on interpersonal skills rather than technical skills and we expected that self-report measures of communication skills would predict performance on this SJT.

Personality dimensions are also known to be related to performance on SJTs that assess interpersonal skills (Chan & Schmitt, 2002). Specifically, Chan and Schmitt found that neuroticism had a negative relationship with performance whilst extraversion, agreeableness, and openness were positively related to SJT scores. We expected the same relations to emerge in the present study.

The final aim of this study was to examine the impact of expertise on performance. As noted earlier, SJTs are characteristically domain specific. There are exceptions (e.g., Chan & Schmitt, 2002) but the general finding appears to be that participants with experience in the field under study typically score higher on the SJT than those with little experience (Cabrera & Nguyen, 2001; Hunter, 2003; Legree, 1995). With the SJT in the current study being very specific to the health profession, it was expected that health professionals would score higher on the SJT than non-health professionals.

Method

Participants

Participants were 107 adults (71 Women and 36 men) with a mean age of 38.7 years (range: 18-74 years), drawn from several sources. One group (*N = 69*) comprised private hospital employees, medical students, medical doctors, and members of the general public available to the first author in regional Queensland (Bundaberg). The remaining participants (*N = 38*) were University of Southern Queensland (USQ) psychology students of varying year levels.

Materials

The SJT used in this study was the American Medical Colleges Communication Skills Assessment Test (AMCCSAT), constructed by Educational Testing Service (ETS). AMCCSAT is a video based SJT designed to elicit responses to two scenarios involving trainee doctors, senior medical staff, patients, and their relatives. The first scenario initially shows a trainee doctor trying (unsuccessfully) to counsel a patient with panic disorder, and then various smaller vignettes are shown depicting possible communication techniques that are to be judged by the study participants in terms of appropriateness. The second scenario initially depicts a trainee doctor following his supervisor's directive to communicate with an Alzheimer patient's relative to ascertain her wishes regarding possible life support for her husband. A total of 12 questions in each scenario required judgment of communication techniques portrayed by the actors. Answers to these 24 questions were scored according to the consensus and expert judgment formats and summed to yield the dependent variables for our regression analyses.

A reduced version of the OCEANIC self-report personality inventory (Roberts, 2001) was used to measure the big-five personality dimensions. The components (and sample items) are as follows: Openness ("I am philosophical"), Conscientious ("I am thorough"), Extraversion ("I am talkative"), Agreeableness ("I am considerate of the feelings of others") and, Neuroticism ("I worry more than most people"). The reduced form consisted of two items per personality factor incorporated into the scale for a total of ten items. Respondents were required to rate on a six-point scale ranging from *never* (1) to *always* (6), how well each item statement described the way they think or feel.

The *Self-Perceived Communication Skills* (SPCS: Roberts, 2004) inventory consists of 20 items such as "I recognize other's emotions and feelings", and "I clarify unclear communication". Respondents rated their perceived frequency of the behaviour described in each item on a six-point scale from *Never* (1) to *Always* (6). There is currently no validation data available for this instrument but the data collected in this study were used to contribute to the process of validating the SPCS.

The *Self-Assessment of Listening Skills* (SALS: Roberts, 2004) inventory consists of 15 items such as “I understand the main idea of lectures and conversations” and, “I don’t have a problem understanding what people say”. The respondents rated the extent to which they agreed that the statement described their listening skills on a five-point scale from *completely agree* (1) to *completely disagree* (5). There is no published validation data for the SALS but data collected in the current study were used to contribute to the process of validating the SALS.

The equipment used in the study included an IBM T40 NotePad and desktop computers in the USQ computer laboratories. Headphones were available for audio control of the video segments of the test battery.

Procedure

The project received ethics clearance from the University of Southern Queensland’s Human Research Ethics Committtee. Participants were tested individually after providing a rationale and explaining that all data would remain confidential. Participants were drawn from two main populations as indicated above (USQ and Bundaberg). Subjects in the Bundaberg area were given a choice of testing venue: their home, their own office, the researcher’s residence, or an office provided by the researcher; whereas the USQ subjects were all tested in the psychology computer laboratories. A research assistant was available at all times during the testing sessions. All participants were given ID numbers with a prefix that identified their profession.

Results

The Statistical Program for Social Sciences (SPSS) version 11.5 was used for analysis of the data. Initially all data were screened for missing and/or duplicated data, out of range values, and incorrect entries. Missing data was near the end of the AMCCSAT in two data sets: three cells in one data set and one cell in another. These missing data were replaced with the modal response to these items from all other data sets (Tabachnick & Fidell, 2001).

Because of the lack of validation data for the SALS and SPCS, we began with an exploratory factor analyses (PCA with oblique promax rotation) of both scales to determine if the scales were unidimensional. Parallel analysis was used to determine the number of components to extract. There was strong evidence of unidimensionality for the SALS and the SPCS. Alphas for both scales were above .90. Intercorrelations between the two-item measures of personality were as follows: Openness (r = .43), Conscientiousness (r = .40), Extraversion (r = .56), Agreeableness (r = .51), and Neuroticism (r = .47).

In order to develop a consensual scoring key for the SJT, answer frequencies were analysed and the answers chosen by the highest percentage of participants considered correct (scoring 1) and all others wrong (scoring 0). In the event that the percentage of the two top answers for any one item had less than a 10% difference, then participants choosing either of those two top answers scored .5. Separate distributions were calculated for each item on the SJT for expert (N = 28) and non-expert (N = 79) groups. The experts consisted of participants who were either a qualified practising medical doctor or a qualified practising nurse (group 1). All other participants were designated as non-experts (group 2).

Chi Square goodness-of-fit tests were used to check for differences in the distributions across the response options for expert and non-expert groups. The distributions differed on one item only (item 2.1, from scenario one) where doctors and nurses were more likely to rate this response as appropriate whereas non-experts were likely to rate it as not appropriate ($\chi^2(1)$ = 5.09, p = .03). There were no other significant differences, indicating that if answers were marked as right or wrong on a consensus-scoring basis, expert and non-expert scoring systems would be exactly the same. This finding is supportive of the hypothesis that expert scoring systems yield the same result as consensus scoring systems.

The data from experts and non-experts were initially separated to enable the testing of how each group would respond to a SJT. In light of the finding that there were virtually no differences, the data were pooled so that the remaining hypotheses could be tested using the full data set. Scores on all SJT items were generated using consensus scoring based on all participants (N = 107). Reliability analysis of all 24 items on the SJT indicated 8 items with no variance (ceiling effects) and 3 items that did not contribute to the reliability of the scale. Together, these 11 items were deleted resulting in a 13-item scale. The 13-item scale was used in all further analyses involving the SJT. The reliability of this scale (α = .46) is not satisfactory for use in selection settings but is adequate for research (Anastasi & Urbina, 1997), where poor reliability will have the effect of suppressing correlations with other variables (Nunnally & Bernstein, 1994).

To investigate the relationships between the self-report measures and the SJT, these scales were analyzed using standard multiple regression with the SJT as the dependent variable and the self-report measures as predictors. Inspection of bivariate scatterplots of all possible combinations of the research variables determined that there was no evidence of heteroscedasticity or non-linearity. Table 1 displays the correlations between the variables, unstandardized (B) and standardized (β) regression coefficients, semi-partial correlations (sr^2) and overall R^2.

Table 1: Standard Multiple Regression of the SJT measure on the SALS, SPCS, and OCEANIC Subscales (*N* = 107).

Variables	SJT	SPCS	SALS	OceE	OceN	OceA	OceC	OceO	*B*	SEB	ß	*sr²*
SPCS	.32**								.02	.03	.08	
SALS	.44**	.57**							.12**	.04	.38	.09
OceE	.06	.46**	.27**						-.14	.11	-.14	
OceN	-.11	.00	-.10	.10					-.07	.09	-.08	
OceA	.20*	.64**	.30**	.52**	.26**				.15	.16	.13	
OceC	.07	.40**	.16	.27**	.27**	.52**			-.03	.11	-.03	
OceO	.15	.27**	.27**	.18	-.07	.13	.18		.04	.10	.04	
Means	8.73	89.56	51.94	8.48	5.58	9.83	8.67	7.87		*R* = .47		
SDs	1.92	11.11	6.07	1.92	2.01	1.58	1.88	1.88		*R²* = .22		

Note: **p* < .05, ***p* < .01.

There was only one correlation between the OCEANIC subscales (Agreeableness) and the SJT (*r = .20, p < .05*). Both the SALS and the SPCS, on the other hand, were correlated with the SJT. Altogether, the self-report measures (SPCS, SALS, OCEANIC) contributed 22% (R^2) to the variance in the SJT. Individually, only one regression coefficient was significant: SALS, *t* = 3.411, *p* < .001 with the SALS test explaining 9% of the total variance in the SJT score. A hierarchical multiple regression with the SJT as the dependent variable and with the SPCS and SAL entered as predictors at the first step and the OCEANIC scales entered as predictors at the second step showed that the two communications skills scales predicted 20% of the variance in the SJT. These findings support the hypothesis that scores on the measures of communication and listening skills will predict scores on the SJT. However, the scores on the personality measures were not significant predictors of the SJT measure when SPCS and SAL were included in the equation.

To determine if there was a difference between how health professionals (experts) and the general population (non-experts) scored on the interpersonal skills SJT, a t-test was run. The independent variable had two levels with the medical practitioners and nurses designated as group one and all other participants as group two (as identified by their ID numbers). The dependent variable was the total score on the 13-item SJT. There was a significant difference in the total score on the SJT (*t(105)* = 2.54, *p* = .01) between doctors and nurses (experts), and all others with the experts scoring higher (group 1, *M* = 9.0; group 2, *M* = 8.5). This is supportive of the hypothesis that experts do better on a context-specific STJ than non-experts.

Discussion

The SJT used in this study was the AMCCSAT which has been newly developed to assess interpersonal skills as part of the screening process for potential medical students. Past research has demonstrated that scoring SJTs is an area of study where sound statistical ground is hard to establish. However, some progress has been made. Firstly, scoring the SJT by consensus scoring was validated according to the initial hypothesis, which stated that an expert scoring system applied to the SJT would yield the same result as a consensus scoring system. The benefit of this knowledge is that in further projects designed to investigate better scoring methods for SJT, the difficulty of deciding who is an expert, and then engaging such experts will not always be necessary.

The findings of this study partially confirmed our expectation that scores on a measurement of communication skills and listening skills would be significant predictors of the SJT. The measure of listening skills uniquely predicted 9% of the variance in the SJT. The lack of a unique contribution from the communications skills measure may be partly explained by the strong correlation between the measures of communications skills and listening skills. The fact that the listening and communication skills scales predicted 20% of the SJT score suggests that the skills involved in making social judgments are linked with listening and communication skills but that there are many other contributors as well.

Our data also suggest that personality is not a strong determinant of performance on the SJT with only Agreeableness showing a significant correlation. This is typical of findings regarding the relationship between personality and performance-based measures of EI, of which the SJT is one type (MacCann et al. 2003).

Lievens, Butse and Sackett (in press) also examined the validity of a video-based SJT used as a predictor in a medical setting. They discovered that an SJT assessing the interpersonal skills required in doctor-patient interactions was a predictor of subsequent performance in medical college. The SJT was able to assess a student's knowledge of effective interpersonal behaviour.

The SJT in this study was developed to target a population who would be working in the health professional domain. The result which was especially noteworthy was that the health professionals scored higher on the SJT than non-health professionals. This provided the most significant source of encouragement from this research for the future use of SJT in selection processes. Despite the problems with reliability which made it difficult to ascertain the strength of the relations between the SJT and the other variables, there was evidence that people who have experience in the kinds of situations shown in the vignettes on the SJT perform better on these tasks than those who do not.

References

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River: Prentice-Hall Inc.

Cabrera, M. A., & Nguyen, N. T. (2001). Situational Judgment Tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*, 103-114.

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

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *International Journal of Aviation Psychology, 13*, 373-387.

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

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. (2004). *Using consensus based measurement to assess emotional intelligence*. Unpublished manuscript.

Lievens, F., Butse, T., & Sackett, P. R. (in press). The operational validity of a video-based Situational Judgment Test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology.*

MacCann, C., Matthews, G., Zeidner, M., & Roberts, R. D. (2003). Psychological Assessment of Emotional Intelligence: A Critical Review of Self-Report and Performance-Based Testing. *International Journal of Organizational Analysis, 11*(3), 247-275.

MacCann, C., Roberts, R. D., Matthews, G., & Zeidner, M. (2004). Consensus scoring and empirical option weighting of performance-based Emotional Intelligence (EI) tests. *Personality and Individual Differences, 36*, 645-662.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Roberts, R. D. (2001). *The Openness-Conscientiousness-Extraversion-Agreeableness-Neuroticism Index Condensed (OCEANIC): Technical Manual*. Entelligent Testing Products: Sydney, Australia.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics* (Fourth ed.). Needham Heights: Allyn & Bacon.

Acknowledgements

All test materials were supplied by ETS. We thank Dr Richard Roberts, in particular, for his advice and assistance throughout the project