

Knowing What NOT To Do Is a Critical Job Skill: Evidence from 10 different scoring methods

Steven E. Stemler*, Varun Aggarwal** and Siddharth Nithyanand**

*Wesleyan University, 207 High Street, Middletown, CT 06459, USA. steven.stemler@wesleyan.edu

**Aspiring Minds, Gurgaon, Haryana, India

Situational judgment tests (SJTs) have become an increasingly important tool for predicting employee performance; however, at least two key areas warrant further investigation. First, prior studies of SJTs have generally relied on samples from the western world, leaving open the question of the validity of using SJTs in the developing world where the majority of the world's workforce resides. Second, there is currently no standardized, theoretically-based method for the development and scoring of SJTs. Therefore, SJTs are highly domain-specific and must be developed anew for each new context. We report the results of three studies, conducted in India, that aim to: (1) test the cross-cultural validity of SJTs in a non-western context, and (2) examine the differential validity of 10 different approaches to scoring SJTs, some of which have the potential to resolve the problem of developing a theoretically-infused, standardized approach to scoring and future development.

1. Introduction

The success of an organization depends heavily upon the characteristics of its employees. At the highest ranks, successful organizations tend to be run by transformational leaders – charismatic individuals who are able to set out a strategic vision and who enjoy the loyal support of employees (Bass, 1990; Burns, 1978). At all levels of an organization, however, employees should be competent in their field and possess a variety of domain-general skills (e.g., communication ability) in order for an organization to be successful. Given the central importance of employee characteristics, the question of how best to identify and select the most outstanding individuals is of critical importance.

The traditional approach to personnel selection in most fields involves evaluating interview performance, letters of recommendation, and tests of cognitive abilities and personality characteristics (Kaplan & Saccuzzo, 2013). Of these methods, research has demonstrated that unstructured interviews tend to be a relatively poor predictor of job performance (Arvey & Campion, 1982;

Huffcutt & Arthur, 1994; Huffcutt, Culbertson & Weyhrauch 2013; Hunter & Hunter, 1984) whereas tests of cognitive abilities tend to be the best predictor, explaining approximately 30% of the variability in ratings of job performance (Hunter, 1986; Schmidt & Hunter, 1998). Yet, there remains a substantial amount of variance in job performance that is unexplained. Recently, researchers have begun to look toward situational judgment tests (SJTs) as a way of supplementing the predictive ability of other measures (McDaniel, Hartman, Whetzel, & Grubb, 2007; Sternberg & Horvath, 1999; Sternberg et al., 2000; Wagner, 1987; Weekley & Ployhart, 2005, 2006). Indeed, the results of such investigations have shown considerable promise, with SJTs adding as much as 10% of variance to the predictive equation (Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001; McDaniel, Morgerson, Finnegan, Campion, & Braverman, 2001). Yet, one of the open questions in the field is why SJTs tend to be predictive. Although many authors have speculated as to why this may be so (Lievens & Motowildo, 2016; see also Weekley & Ployhart, 2006 for a review), there is not yet a

clear answer to this question. The lack of a widely-accepted, theory-based approach to SJT development makes it elusive to determine how to improve the predictive validity of these tests. Furthermore, empirical methods of test development and scoring have become increasingly popular. The result of this trend has been that the correct answers, scoring methods and in some cases even the test items are developed and validated afresh for each organization/role. This leads to lack of test standardization, a lack of reuse, and has often prevented the use of the tests for formative purposes. In addition, most of the research on SJTs has taken place in the western world, so the question of cross-cultural applicability remains open to further investigation.

Therefore, the purpose of the current study is twofold. First, we investigate the extent to which two newly developed SJTs for management and sales are significant predictors of real world job performance in a non-western context. Specifically, we evaluate the predictive power of these tests for entry-level and mid-level roles involving interpersonal interactions and management within some of the top 100 companies in India. If the SJTs are shown to be predictive in a non-western context, this would provide some evidence to support the cross-cultural validity of the technique.

Second, we compare 10 different approaches to scoring SJTs. Some of these approaches have been reported previously while others we are proposing are new. We expect that each of these 10 scores will be differentially predictive of real-world outcomes. Among these, we hypothesize that scoring people for their ability to identify the worst response will exhibit the greatest predictive validity. Our rationale for this prediction is based on Elliot's (1999) performance-approach and performance-avoidance motivational distinction. We propose that if scoring people based on their ability to identify the worst response yields empirically superior validity results, this could potentially provide a framework to score SJTs consistently in a theoretically-infused, standardized way across job roles/organization.

2. Background

2.1. Cultural context of the current study

With over 1.2 billion residents, the country of India currently has the second largest population in the world (Central Intelligence Agency World Factbook, 2014). Recently, there has been an explosion of individuals trained in higher education with the practical result being too many applicants for too few positions. As a result, Aspiring Minds, India (<http://www.aspiringminds.com>), has developed a comprehensive battery of assessments designed to help identify, in a systematic, empirical, and theoretically informed way, those individual applicants who are likely to succeed and/or those who are likely to

fail in different industries. Given the rapid proliferation and empirical success of the technique of Situational Judgment Testing (SJTs) for predicting employee performance in the western world (Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001; McDaniel et al., 2001), Aspiring Minds has recently developed some SJTs for use in their larger assessment battery. In this paper, we report three studies in which the validity of different SJTs designed for different occupational groups was tested.

2.2. What are SJTs?

SJTs are written descriptions or video depictions of job-relevant situations that involve some sort of dilemma or conflict. The scenarios are typically derived from a job analysis study of critical skills required for a particular job (Schneider & Konz, 1989). SJTs have their roots in the critical incidents technique (Flanagan, 1954) in which participants are interviewed and asked to come up with job-relevant incidents, the responses to which distinguish high-performing employees from low-performing employees. Scenarios are then developed based on these incidents and a new set of participants are presented with the scenarios and asked to evaluate different ways of responding. The responses may be short answer, but more typically participants are given a finite number of potential responses and asked to rate the quality of the response option on a Likert scale. In the current investigation, participants were asked to select what they perceived to be the 'best' and 'worst' response options (see Figure 1).

2.3. Why do SJTs predict performance?

Whereas much prior research has supported the predictive validity of SJTs, there are a variety of subtle details that can impact their validity and reliability. The most important differences in SJT construction relate to three dimensions: (1) the framing of the prompt; (2) the method by which response options are developed; and (3) the method by which responses are scored.

With regard to the framing of the prompt, Motowidlo, Hooper, & Jackson (2006) have argued that questions that ask participants what they 'would do' tend to tap into the personality characteristics of the participants whereas scenarios that ask participants what they 'should do' tap into procedural knowledge, or what Sternberg and colleagues call tacit knowledge (Sternberg et al., 2000; Wagner, 1987). Still, Motowidlo et al. argue that despite the conceptual difference between these prompts there is unlikely to be a practical difference as it is reasonable to assume that those taking the tests are likely to interpret questions about what they would do as questions about what they should do, that job seekers are likely to try to put their best foot forward, and people are generally

Instructions: In this module, you will be provided with various situations one faces in the corporate environment. Based on the situation provided, we wish to understand what action you will take in the given situation. You have to choose two options: (a) The most desirable action amongst the options provided; and (b) The least desirable action amongst the options provided.

You are a salesperson with a company that deals in industrial cleaning equipment. You recently met a prospective client who is interested in your product as it meets his requirements. However, he tells you that his wife's uncle is in the same business; hence he would prefer buying the product from him. What would you do?

- (1) You feel this deal is tough sell since the client would prefer buying the product from a relative who is in the same business. Hence, you would not pursue further.
- (2) Try to convince the client to buy your product by highlighting how your product is unique in its features and better than the competitor's product.
- (3) Ask the client to go for a trial use of your product without paying for it and then decide whether to buy your product or the competitor's product.
- (4) Suggest to the client that you can have a stand-by arrangement and if he needs a second vendor or a backup, you will be there to help him meet his commitment.

Figure 1. Example item from the Sales Situational Judgment Test (SJT).

Note. Option 1 was the consensus 'avoid' response whereas option 3 was the consensus 'approach' response.

likely to carry out behaviors that they believe are effective anyway. Nevertheless, Ployhart and Ehrhart (2003) conducted a study comparing the psychometric characteristics of results associated with six different sets of instructions. What they found was that 'should do' prompts exhibited weaker reliability and validity than those that asked what participants 'would do'. It is worth noting, however, that their approach to scoring each of these six responses was done in a traditional manner (explained further in the next section), which may have impacted their conclusions.

A second subtle distinction among SJTs is the method by which response options are generated. Stemler and his colleagues have proposed that response options should be generated based on a theoretical framework (Stemler, Elliott, Grigorenko, & Sternberg, 2006; Stemler & Sternberg, 2006). However, the vast majority of research in the field has taken an empirical, but atheoretical, approach to the development of response options in which participants are first interviewed and asked to comment on a variety of different ways one might respond to a particular scenario. Their different approaches are then used as response options. As a result, one scenario may have 10 potential responses while another scenario may have only six potential response options and these are rarely tied back to any meaningful theory.

Finally, there is tremendous variation in the way that participants have been asked to respond to the prompts. In different SJTs, participants may be asked to select the best and worst responses, the most effective response, what they should do, what they would do, what they would most and least likely do, to rank-order their responses, or to rate the effectiveness of each response. A pilot study by Cucina, Vasilopoulos, Leaman (2003,

April) investigated the validity of computing the 'worst' response and the 'best' response separately. They found that the 'worst' response scores predicted the criterion better, whereas the 'best' response score correlated more highly with personality scores. There was also some speculation as to what the two scores, often designated as most likely and least likely scores, measure. Similarly, Elliott, Stemler, Grigorenko, Sternberg, and Hoffman (2011) conducted a study of teachers' tacit knowledge and found that the ability to identify the 'worst response' was a statistically significant predictor of teacher effectiveness whereas the ability to identify the 'best response' was not. Despite substantial variation in approaches to constructing and scoring SJTs, a broad array of past research has demonstrated their predictive validity within the context of personnel selection (Clevenger et al., 2001; McDaniel et al., 2007; Sternberg et al., 2000).

Although SJTs represent an exciting new line of assessment in the context of personnel selection, they do suffer from certain limitations. For example, it is challenging to create job-relevant SJTs that are standardized across different jobs and companies. Their domain-specific nature typically requires the development of new SJTs for each unique job position. Furthermore, there is currently not a clear understanding of exactly why SJTs tend to be predictive of job performance.

One reason for the lack of clarity on why the SJTs are predictive may be due to the general belief, probably rightly so, that the correct way of handling a situation will change from one job to another, one culture to another, and one organization to another. For instance, a process-driven organization may consider different approaches of handling a particular sales situation than would a goal-driven organization. Similarly, a hierarchical organization

Table 1. Ten approaches to scoring

Score	Description
Traditional Avoid (-1, 0, 1)	In this approach, an individual who correctly identifies the 'avoid' answer (as identified by group consensus) receives a score of '1'. An individual who selects the group-identified 'avoid' answer as their preferred 'approach' answer receives a '-1'. All other choices receive a score of '0'
Traditional Approach (-1, 0, 1)	In this approach, an individual who correctly identifies the 'approach' answer (as identified by group consensus) receives a score of '1'. An individual who selects the group-identified 'approach' answer as their preferred 'avoid' answer receives a '-1'. All other choices receive a score of '0'
Traditional Total (-1, 0, 1)	In this approach, the scores from the Avoid only and Approach only variables are added together to form a single score, perhaps known as the capacity to identify the 'correct' response
Match Avoid (1, 0)	In this approach, an individual who correctly identifies the 'avoid' answer (as identified by group consensus) receives a score of '1'. All other choices receive a score of '0'
Match Approach (1, 0)	In this approach, an individual who correctly identifies the 'approach' answer (as identified by group consensus) receives a score of '1'. All other choices receive a score of '0'
Match Total (1, 0)	In this approach, the scores from the Avoid only and Approach only variables are added together to form a single score, perhaps known as the capacity to identify the 'correct' response
PenaltyAvoid (-1,0)	In this approach, an individual who selects the group-identified 'avoid' answer as their preferred 'approach' answer receives a '-1'. All other choices receive a score of '0'
PenaltyApproach (-1,0)	In this approach, an individual who selects the group-identified 'approach' answer as their preferred 'avoid' answer receives a '-1'. All other choices receive a score of '0'
PenaltyTotal (-1,0)	In this approach, the scores from the Avoid only and Approach only variables are added together to form a single score.
Cross Total: PenaltyApproach + Match Avoid	In this approach the scores from Avoid score (1,0) and Opposite Approach score (-1,0) are added together to form a single score.

may propose different ways to handle situations with one's boss and team than would an organization with a flat structure. We believe that a partial answer to the question of why SJTs are predictive may be found by scoring the ability to recognize bad responses separately from the ability to recognize good responses. Furthermore, we propose that scoring SJTs in this manner provides a solution to the problem of standardizing SJTs.

2.4. Theoretical framework

The theoretical rationale for the current study comes from the research on achievement motivation. In the vast literature on achievement motivation, there is a distinction between mastery goals and performance goals (Dweck, 1986), where, 'performance goals focus on the demonstration of competence relative to others whereas mastery goals focus on the development of competence or task mastery'. (Elliot, 1999, p. 169).

Following Sternberg and colleagues (Sternberg et al., 2000), we believe that SJTs measure the tacit knowledge aspect of practical intelligence (Stemler & Sternberg, 2006; Sternberg, 1999). Because tacit knowledge is a cognitive trait, participants taking SJTs are typically being evaluated for their competence relative to others. Consequently, performance on SJTs should invoke a performance orientation in participants.

With regard to performance orientation, Lewin, Dembo, Festinger, and Sears (1944) noted long ago that people have two distinct achievement drives – the desire to achieve success and the desire to avoid failure. Elliot (1999) incorporated this dichotomy into his formal

theory of motivation by articulating a distinction between performance-approach motivation and performance-avoidance motivation. He notes, 'In approach motivation, behavior is instigated or directed by a positive or desirable event or possibility; whereas in avoidance motivation, behavior is instigated or directed by a negative or undesirable event or possibility' (p. 170).

Following Elliot's distinction, we hypothesize that knowing 'what not to do' (i.e., performance-avoid motivation) is, in principle, a different skill from knowing the right way to handle a situation (i.e., performance-approach motivation). In contrast to Elliot, and following Sternberg, we are proposing that our SJTs measure the cognitive dimension, rather than the behavioral aspect, of this motivational process. Stated differently, what our SJTs are assessing is whether an individual knows (i.e., cognitively) which behaviors would likely lead to undesirable events or possibilities (performance-avoidance) and which behaviors would likely lead to desirable events or possibilities (performance-approach).

2.5. Methods for scoring SJTs

There is currently no single, widely-accepted approach to scoring SJTs found in the literature. Furthermore, many of the approaches that are used are justified mainly on empirical grounds, rather than theoretical grounds. For each of the studies reported in this paper, we test 10 different approaches to scoring SJTs. Each of the 10 approaches is described further in Table 1. These 10 approaches may be broadly described within the context

of three families of scores and one new cross-family total score.

2.5.1. Match approaches

The first family of three scoring techniques is the 'Match' approaches. The guiding principle for this family is that a group of participants comes to consensus (either a priori or post-hoc) and identifies the 'worst' (i.e., 'avoid') and 'best' (i.e., 'approach') response to a situation. If an individual participant selects the same 'worst' (i.e., 'avoid') response the group does, then there is a match between the individual and the group and the individual is awarded 1 point. If the individual identifies any other response as the 'worst' (i.e., 'avoid'), the individual receives a score of 0 on the match item. The same principle applies for the 'approach' response. And a 'total match' score is computed by simply summing the 'match approach' and 'match avoid' scores. This scoring technique suggests that the individual who scores high on the match is able to appropriately perceive the correct response as such.

2.5.2. Penalty approaches

The second family of scoring techniques is the 'Penalty' approaches. Like the 'Match' approaches, these techniques first assume a consensus 'best' (i.e., 'approach') and 'worst' (i.e., 'avoid') response for each scenario. Then, if an individual selects as his 'best' (i.e., 'approach') response the group-identified 'worst' (i.e., 'avoid') response, the individual will receive a score of -1 on the 'Penalty Approach' variable. Otherwise the individual receives a score of 0. And vice-versa for the 'Penalty Avoid' variable. A third 'Penalty Total' is simply the sum of the previous two variables. This family of scores captures those individuals who are so oblivious to social norms that they would actually choose the exact opposite response as being the ideal selection.

2.5.3. Traditional approaches

Traditional approaches combine elements from the 'Match' and 'Penalty' approaches into a single overall score. Motowidlo, Dunnette, and Carter (1990) introduced this approach by using a score that combines the effect of both the 'Match Approach' response (most likely) and 'Penalty Approach' response chosen by the individual. What we will henceforth call the 'Traditional Approach' score is calculated as follows:

1. The person gets a $+1$ if s/he chooses the designated 'best' (i.e., 'approach') answer as the 'best' (i.e., 'approach') response.
2. The person gets a -1 if s/he chooses the designated 'worst' (i.e., 'avoid') answer as the 'best' (i.e., 'approach') response.
3. The person gets a zero if s/he chooses a distractor.

The 'Traditional Avoid' response is scored in an analogous way. The 'traditional approach' and 'traditional avoid'

scores are then summed to create a 'Traditional Total' score.¹

2.5.4. Cross-total avoid score

Finally, in extending prior approaches to scoring, we propose a new score that does not necessarily rely on an agreed upon notion of a designated 'best' (i.e., 'approach') answer, but only an agreed upon notion of a designated 'worst' (i.e., 'avoid') answer. This is because we believe that there are multiple ways of handling a situation correctly and hence no one best answer. On the other hand, following the work of Elliott et al. (2011), we propose that there is typically one agreed upon 'worst' way to respond to any given situation. In light of this conceptualization, our 'Cross-Total Avoid' response score is computed by adding 'Match Avoid' plus 'Penalty Approach' where:

1. 'Match Avoid' score: the person gets a $+1$ for choosing the designated 'avoid' answer as the 'avoid' response and 0 otherwise.
2. 'Penalty Approach' score: The person gets a -1 for choosing the designated 'avoid' answer as the 'approach' response and 0 otherwise.

One may note that we preserve the idea of the candidate selecting both a 'best' and a 'worst' response; however, in our approach to scoring, we score them only in reference to the agreed upon 'worst' (i.e., 'avoid') answer. The score may be interpreted as meaning that if the candidate correctly identifies the designated 'avoid' answer he is excellent, if he doesn't he is average, but if he thinks the designated 'avoid' answer is the 'approach' response, he is penalized. This is the score that may best measure whether the person knows 'what not to do'.

3. Methods

3.1. Study 1

3.1.1. Purpose

The goal of study 1 was to evaluate the extent to which a newly developed situational judgment test of sales would predict real-world success in sales in a non-western context. Additionally, the 10 different approaches to scoring the SJT, were compared.

3.1.2. Instrumentation

Consistent with the traditional approach to SJT construction, a literature review was conducted to get a sense of critical skills required in the domain of sales (e.g., Dalessio, 1994; Hausknecht & Langevin, 2010; Phillips, 1992, 1993). In addition, in-depth interviews were conducted with subject matter experts (SMEs) in an effort to identify important competencies that distinguished successful salespeople from less successful salespeople. The results of these investigations identified five key

competencies in the domain of sales: (1) goal orientation, (2) analytic thinking and decision making, (3) client expectation management, (4) customer service orientation, and (5) proactive influence.

During the interviews, SMEs were also asked to identify critical incidents (situations) faced in the context of sales along with a range of possible responses to the situations they identified. Because Ployhart and Ehrhart (2003) have argued that responses to SJTs may be influenced by the nature of instructions, we list below the exact instructions that were given to participants. Participants in this study were given the following instructions:

In this module, you will be provided with various situations one faces in the corporate environment. Based on the situation provided, we wish to understand what action you will take in the given situation. You have to choose two options: (a) The most desirable action amongst the options provided; and (b) The least desirable action amongst the options provided.

For each salesperson in the organization, the employer sets particular sales targets for each month. Thus, the dependent variable we were trying to predict was percentage of sales targets that were met by each employee who took our SJT. The sales targets were calculated as the mean of the last two quarters from when the study was conducted.

3.1.3. Methods

The sales SJT was individually administered online via the Aspiring Minds assessment engine. Participants were required to take the assessment at an authorized testing center. The assessment consisted of 22 scenarios, each of which contained 4–5 response options. The SJT took approximately 35 min to complete. Participants were asked to take the test by their employer. Data on the percentage of sales targets achieved was obtained from the reporting manager who gathered information from the company information system.

3.1.4. Sample

Participants in Study 1 were drawn from a 400-person company in India. The test went to only one sales team in the organization ($N = 54$). The test was done as a development exercise to provide team members feedback. All of the sales people on the team with experience levels ranging from 6 months to 2 years at the company were asked by their managers to participate in the assessment. The final sample proportionally represented the sales population in the company based on gender, age, educational qualifications, and performance.

3.1.5. Results

A total of 50 participants (45 males, 5 females) completed the sales SJT for a response rate of 93%. This was representative of the organization's sales force, which had approximately 88% males. The average age of participants was 26 years old ($SD = 3.5$ years) and the participants had worked for the organization an average of 11 months ($SD = 5.1$ months). The highest level of education in the sample was an MBA, with 84% of participants holding an MBA. Table 2 reports the descriptive statistics relative to the 10 different approaches to scoring.

Table 3 presents the correlation coefficients among the 10 different approaches to scoring the sales SJT and the outcome variable of interest (% of sales targets achieved).

The correlations between different methods for scoring the 'approach' responses were highly correlated with one another with r 's ranging from .71 to .94. Further, the different methods for scoring the 'avoid' approach were also highly correlated with one another with r 's ranging from .89 to .97. However, the data reveal that the scores associated with the ability to correctly identify the 'avoid' response were typically only moderately correlated with the scores associated with the ability to correctly identify the 'approach' response, with r 's ranging from .40 to .47.

Table 3 reveals that among the various approaches to scoring, the 'Cross-Total Avoid' approach had the strongest and most statistically significant correlation with the outcome variable ($r = .36, p < 0.05$), followed by 'Match Avoid' ($r = .33, p < .05$), and 'Traditional Avoid' ($r = .28$,

Table 2. Descriptive statistics for Study 1

Variables	Minimum	Maximum	Mean	Std. Deviation
Traditional avoid	−8	14	6.50	5.11
Traditional approach	2	16	9.20	3.52
Traditional total	−6	28	15.70	7.44
Match avoid	1	15	8.70	3.48
Match approach	5	16	10.88	2.75
Match total	7	29	19.58	5.20
Penalty avoid	−9	0	−2.20	1.97
Penalty approach	−5	0	−1.68	1.30
Penalty total	−13	0	−3.88	2.80
Cross total	−3	14	7.02	4.14

Note: $N = 50$.

Table 3. Pearson correlation coefficients among the 10 different approaches to scoring the sales SJT and between SJT scores and % sales targets achieved for Study 1

Variables	Sales target achieved	1	2	3	4	5	6	7	8	9	10
1. Traditional_avoid	0.28*	—									
2. Traditional_approach	0.16	.47**	—								
3. Traditional_total	0.27	.91**	.80**	—							
4. Match_avoid	0.33*	.97**	.44**	.87**	—						
5. Match_approach	0.08	.40**	.94**	.72**	.39**	—					
6. Match_total	0.26	.86**	.79**	.96**	.87**	.79**	—				
7. Penalty_avoid	0.15	.89**	.45**	.82**	.74**	.37**	.69**	—			
8. Penalty_approach	0.25	.42**	.71**	.62**	.36**	.44**	.48**	.44**	—		
9. Penalty_total	0.23	.82**	.65**	.87**	.69**	.46**	.71**	.91**	.78**	—	
10. Cross_total (Match avoid + penalty approach)	0.36*	.94**	.59**	.93**	.96**	.46**	.89**	.76**	.62**	.82**	—

Note: $N = 50$.

*Correlation is statistically significant at $p < .05$.

**Correlation is statistically significant at $p < .01$.

$p < .05$). It is worth noting that although it was on the borderline, the 'Traditional Total' score was not a statistically significant predictor ($r = .27$, ns).

To further investigate the incremental validity of the 'avoid' scores over the 'approach' scores, we ran a series of regression analyses. The data in Table 4 show that

Table 4. Incremental validity results from Study 1 for predicting percent of sales targets achieved

Predictor	ΔR^2	β
<i>Match approaches</i>		
Step 1	0.01	
Match approach		0.08
Step 2	0.11+	
Match approach		−0.05
Match avoid		.35**
<i>Traditional approaches</i>		
Step 1	0.03	
Traditional best		0.16
Step 2	0.08	
Traditional best		0.03
Traditional worst		.27+
<i>Penalty approaches</i>		
Step 1	0.06+	
Penalty best		.25+
Step 2	0.07	
Penalty best		0.23
Penalty worst		0.05
<i>Cross totals</i>		
Step 1	0.02	
Opposite cross total: match approach + penalty avoid		0.13
Step 2	0.15*	
Opposite cross total: match approach + penalty avoid		−0.23
Cross total: match avoid + penalty approach		.52**

Note: Outcome variable = percent sales targets achieved. + $p < .10$;

* $p < .05$; ** $p < .01$.

when 'Match Approach' was entered into a regression and then followed by 'Match Avoid', the 'Match Avoid' score provided statistically significant incremental validity over the 'Match Approach' score. This provides further evidence in favor of treating each approach to scoring as if it were measuring a separate skill.

When the same procedure was run for 'Traditional' scores and 'Penalty' scores, we observed no significant incremental validity. This makes some conceptual sense because each of these scores intermingles the skills of identifying approach and avoid. When 'Opposite Cross Total: Match Approach + Penalty Avoid' was put into a regression equation first followed by 'Cross Total: Match Avoid + Penalty Approach', we find that 'Cross Total' adds statistically significant incremental validity to the regression equation. This again argues for the importance of the ability to identify the 'worst' answer since 'Match Avoid' is the capacity to correctly identify the 'avoid' answer and in the 'Penalty Approach' score, the person receives an additional penalty (a score of -1) for choosing the designated 'avoid' answer as the 'approach' option and a 0 otherwise. All in all, the scores associated with the ability to appropriately identify the 'avoid response' exhibited statistically significant incremental validity over the scores associated with the ability to identify the 'approach response'.

3.1.6. Discussion

The results from Study 1 reveal three noteworthy findings. First, the results of the intercorrelations of scoring methods show that the 'approach' scores and 'avoid' scores were only moderately correlated, suggesting these two scores are measuring distinct constructs.

Second, the only statistically significant predictors of the outcome variable came when one considers the ability to correctly identify the 'avoid' response as a skill that is separate from the ability to correctly identify the 'approach' response. Furthermore, this prediction

becomes better if we do not rely on the concept of a designated correct 'approach' response but only score responses on the basis of a designated 'avoid' response. Both the 'Match Avoid' response and 'Penalty Approach' response graded this way show higher correlation as compared to their counterparts. Their sum outperforms all other scoring methods. When run in a hierarchical regression analyses, the 'avoid' scores were added after the 'approach' scores, the 'avoid' scores added statistically significant incremental validity to the predictive equation.

The third major finding is that the sales SJT was a statistically significant predictor of actual sales within the context of a mid-sized technology firm in India, providing support for the cross-cultural utility of the SJT methodology in the workforce context in India. The results of this study show promise for the predictive validity of the instrument; however, we were also interested in understanding whether the sales SJT would demonstrate incremental validity over and above other important constructs. In order to further investigate these findings, we conducted a second, follow-up study.

3.2. Study 2

3.2.1. Purpose

Study 2 focused on evaluating the incremental validity of a short form of the sales SJT. Specifically, we examined whether the sales SJT was predictive of sales success after taking into account logical reasoning, personality, and insurance readiness. As in Study 1, the traditional and alternate forms of scoring were compared.

3.2.2. Instrumentation

The short-form sales SJT was administered along with a personality inventory, a test of logical reasoning, and a test of insurance readiness. The SJT assessments were not expected to correlate significantly with the personality inventory or logical reasoning test; however, a moderate correlation was expected with the test of insurance readiness.

3.2.2.1. Personality inventory (AMPI). Aspiring Minds has developed a personality test based on the Five-Factor Model of Personality (McCrae & Costa, 1987). The process for developing the Aspiring Minds Personality Inventory (AMPI) was as follows. An item pool was created according to the Big Five model. All items had simple to understand English and statements which matched the Indian situational and cultural context. All items were reviewed for understandability and clarity of language. Each item developed was allocated to one of the sub-traits by three psychometricians. The final sub-trait for each item was allocated according to their consensus. After that, the items were piloted on a sample of approximately 20,000 entry-level job-seekers stratified according to the diversity in the population.

The AMPI test is a self-report measure that takes approximately 20 min to complete. It includes 80 items measuring the dimensions of Openness to Experience ($\alpha = .70$), Conscientiousness ($\alpha = .71$), Extraversion ($\alpha = .72$), Agreeableness ($\alpha = .72$), and Emotional Stability ($\alpha = .74$). In addition, we include a factor called Polychronicity ($\alpha = .77$), which is the extent to which people prefer to be engaged in two or more activities at the same time and believe that doing so is effective (Conte & Jacobs, 2003; Kantrowitz, Grelle, & Beaty, 2012). The AMPI has been translated in two Indian languages, Hindi and Gujarati. It has been delivered to various populations including those of entry-level engineers, post-graduates, three-year course graduates, rural population and across all Indian states. Its scales show resilience and high reliability throughout. Inter-scale correlations show all correlations to be less than 0.35, in line with global studies. The instrument shows low correlations with English and cognitive skill instruments clearly showing that it measures an orthogonal dimension, whereas with situational-judgment tests, it shows moderate correlations. AMPI is currently being deployed as a selection filter and for internal workforce evaluation. Being India's most deployed Personality tool, over quarter a million candidates have been evaluated on AMPI for various job roles across different sectors. It is also useful in being able to predict success in roles spanning from sales, customer service, relationship management, collections, technical support, managerial roles, and leadership (<http://www.aspiringminds.in/casestudies.html>).

The AMPI has been validated against the NEO-FFM on a sample of 314 final year engineering students from three colleges. Both the tests were delivered to all students and in random order. Due to various reasons as documented in 'Challenges and Concerns with International Personality Assessments' (available on Aspiring Minds' website: <http://aspiringminds.com>), only three out of the five scales of NEO-FFM were reliable. Both the Agreeableness and Openness-to-experience scales of NEO-FFM show reliability of less than 0.50. Even after removing a few questions, the reliability of the scales did not improve. On the other hand, all scales of AMPI show a reliability of more than 0.71 on the sample population.² The inter-correlation between AMPI scales and the three reliable scales of NEO-FFM is more than 0.57, with the corrected correlation (taking into consideration the reliability of the instruments) being greater than 0.78.

3.2.2.2. Logical reasoning (AMCAT). The Logical Reasoning test used in this study was developed internally and is a proprietary component of the Aspiring Minds Computer Adaptive Testing (AMCAT) program. The test assesses the capacity of an individual to interpret things objectively, to be able to perceive and interpret trends to make generalizations and be able to analyze assumptions behind an argument/statement. These abilities are

primary for the success of a candidate in most industries. Specifically, the AMCAT logical reasoning test is divided into three sections: (1) *Deductive Reasoning*: Assesses the ability to synthesize information and derive conclusions; (2) *Inductive Reasoning*: Assesses the ability to learn by example, imitation or hit-and-trial. This also provides an indication of how creative the individual is; and (3) *Abductive Reasoning*: Assesses the critical thinking ability of an individual to see through loopholes in an argument or group of statements. All three abilities are tested using both numeric and verbal stimuli. Various case studies have shown AMCAT Logical Ability to strongly correlate to technical trainability, soft-skill trainability and process trainability (e.g., http://aspiringminds.in/researchcell/featured_profiles/software-engineer-fresher-entry-level.html). Logical ability also correlates with sales and support related role performance. The test consists of 24 items and requires 35 min to complete. The Cronbach's alpha for the test is .77. The standard error for this IRT-based adaptive test is 48 points on a 100–900 scale. The standard error is based on theoretical estimates arrived at by computer simulations. The IRT parameters of the items in the test are set according to the attempt of AMCAT by graduates across India with a sample of more than 50,000 test takers.

3.2.2.3. Insurance readiness. The insurance readiness test was developed internally by AM and is an elementary insurance module that checks for the candidate's basic knowledge of insurance concepts as well as their practical application in the work place. The module assumes that the candidate does not have formal educational or professional training in insurance. Hence, the module is a good evaluation tool to check how 'ready' the candidate is for the insurance industry and is a surrogate measure of their interest or desire to be in the industry. It covers the following three topics: (1) overview of insurance, (2) risk and insurance principles, and (3) types of insurance. The test consists of 18 items and takes approximately 12 min to complete. The Cronbach's alpha has been shown to be .77.

3.2.2.4. Success at sales. For each salesperson in the organization, the employer sets particular sales targets for each moth. Thus, the dependent variable we were trying to predict was percentage of sales targets that were met by each employee who took our SJT.

3.2.3. Methods

The assessments were individually administered online via the Aspiring Minds assessment engine. Participants were required to take the assessment at an authorized testing center. A short version of the sales SJT assessment was administered. The test consisted of 12 items and took approximately 22 min to complete. Participants were not

given any incentives for their participation, but were asked by their managers to take the assessment.

3.2.4. Sample

All of the participants were insurance salesmen selling insurance directly to their customers. Aspiring Minds asked their organizational client to share a stratified sample of candidates based on performance and parameters of gender, educational degree, experience, etc. The final sample proportionally represented the sales population in the company based on gender, age, educational qualifications, and performance. In all, a total of 146 individuals participated in Study 2.

3.2.5. Results

A total of 146 participants (84% male, 16% female) completed the abbreviated sales SJT for a response rate of 73%. The gender demographics were representative of the sales personnel selling insurance policies in a large insurance company, which had approximately 80% males. The mean age of participants in our sample was 26 years old ($SD = 3.5$ years) and the participants had worked for the organization an average of 11 months ($SD = 5.1$ months). Approximately half of the sample had worked at the organization for less than 1 year (48%) whereas the other half had worked there for more than one year (52%). Fully 77% of the sample held a Bachelor's degree, with 14% holding an MBA and 9% holding another type of degree. The results were scored in the same manner as Study 1. Table 5 reports the descriptive statistics on each of the variables included in Study 2.

Table 6 provides the intercorrelations among the 10 scoring approaches.

Table 7 provides the correlation results between the various scoring approaches, the additional measures, and the dependent variable of percent of sales targets met. Similar to the results of the previous study, 'Match Avoid' scores and 'Penalty Approach' scores showed the highest correlations with the dependent variable ($r = .26$, $p < 0.01$, and $r = 0.14$, ns , respectively) as compared to other approaches to score the 'avoid' and 'approach' response. The 'Traditional Total' also showed a significant correlation with the outcome ($r = .17$, $p < .05$), however this correlation was weaker than our new approaches to scoring. The 'Cross Total', summing 'Match Avoid' and 'Penalty Approach' outperforms all other scoring methods with respect to correlation with output ($r = .27$, $p < .001$).

We further wanted to investigate the predictive power of the 10 scores over and above the other instruments used in the study. Table 8 presents the results of an incremental validity study. When personality factors, logical reasoning, and insurance readiness are used in a prediction equation, these variables account for a total 5% of the variability in percentage of sales targets reached. This number is surprisingly low. None of the traditional scoring methods show a significant correlation with the

Table 5. Descriptive statistics for Study 2

Variable	Minimum	Maximum	Mean	Std. deviation
Extraversion	−5.06	2.32	.36	1.16
Conscientiousness	−4.18	2.00	.51	.97
Emotional stability	−2.52	3.19	.00	1.22
Openness	−6.42	1.63	−.18	1.26
Agreeableness	−5.28	1.71	.24	1.01
Polychronicity	−2.28	2.90	.33	1.02
Logical reasoning	235.00	675.00	460.00	85.00
Insurance readiness	189.00	811.00	446.00	114.00
Perf (% targets reached)	5.00	426.00	63.00	65.00
Traditional avoid	−6.00	7.00	.12	2.05
Traditional approach	−4.00	5.00	.41	1.87
Traditional total	−7.00	10.00	.53	3.10
Match avoid	.00	7.00	1.82	1.28
Match approach	.00	5.00	2.05	1.35
Match total	.00	10.00	3.88	2.02
PenaltyAvoid	−6.00	.00	−1.70	1.27
PenaltyApproach	−5.00	.00	−1.64	1.11
Penalty total	−9.00	.00	−3.34	1.79
Cross total	−4.00	7.00	.18	1.78

Note: $N = 146$.

output. As expected, when we add the 'Match Avoid' and the 'Cross Total' score to the regression equation, we get a total significant correlation in each of the case. The prior explains a total variance of 11% and the latter takes it to 13%, more than doubling the predictive power.

3.2.6. Discussion

The results of Study 2 show that not only does the sales SJT demonstrate predictive validity for sales success, but it also demonstrates incremental validity over and above tests of insurance readiness, logical reasoning, and personality. The regression results reveal that when the SJTs are scored only with regard to a designated 'avoid' answer, the variable more than doubles the ability to predict who will be successful in sales. Scoring the variable in such a way explains a statistically significant amount of variance in the overall regression equation.

As noted, however, the correlations between the other cognitive and personality variables and the criterion were themselves surprisingly low, with none of

these variables exhibiting a statistically significant correlation with the outcome variable. There are two potential explanations for this result. First, prior research has shown that objective sales metrics tend to show low correlations with predictor variables whereas the correlations tend to be stronger with subjective measures, such as managerial ratings of employee performance (Skyrme, Wilkinson, Abraham, & Morrison, 2005). Our current study replicates this finding. Next, the lower correlations could be due to the fact that the company may already be selecting employees using personality, insurance readiness measures and as such, there would be restricted range on these variables within the company which would attenuate correlation and regression coefficients.

Having provided some evidence in support of the predictive and incremental validity of the sales SJT, we were interested in determining whether the findings would replicate, using a different SJT – one designed for managers – and a different sample.

Table 6. Study 2: intercorrelations among 10 scoring techniques

Variables	1	2	3	4	5	6	7	8	9	10
1. Traditional_avoid	–									
2. Traditional_approach	.25**	–								
3. Traditional_total	.81**	.77**	–							
4. Match_avoid	.80**	.19*	.64**	–						
5. Match_approach	.23**	.81**	.64**	.17*	–					
6. Match_total	.67**	.66**	.84**	.75**	.78**	–				
7. Penalty_avoid	.80**	.22**	.66**	.29**	.20*	.32**	–			
8. Penalty_approach	.14	.70**	.51**	.10	.14	.16	.12	–		
9. Penalty_total	.62**	.60**	.77**	.24**	.26**	.32**	.76**	.70**	–	
10. Cross_total	.61**	.86**	.92**	.67**	.70**	.89**	.32**	.59**	.58**	–

Note: $N = 149$, *Correlation is statistically significant at $p < .05$, **Correlation is statistically significant at $p < .01$.

Table 7. Correlation table among scoring approaches, sales target (outcome), and other predictor variables in Study 2

Score	Perf	Logic	Insurance readiness	E	C	ES	O	A	P
Traditional avoid	0.16*	0.04	0.15	0.09	0.08	0.03	0.17	0.14	0.01
Traditional approach	0.10	0.10	0.10	0.07	0.11	−0.03	0.05	0.02	0.11
Traditional total	0.17*	0.09	0.16	0.10	0.12	0.00	0.14	0.10	0.08
Match avoid	0.26**	0.06	0.06	0.14	0.13	−0.04	0.12	0.15	0.12
Match approach	0.02	0.09	0.14	0.08	0.10	−0.02	0.04	0.06	0.12
Match total	0.18*	0.10	0.13	0.14	0.15	−0.04	0.10	0.14	0.15
Penalty avoid	0.00	0.01	0.18	0.01	0.01	0.09	0.15	0.07	−0.10
Penalty approach	0.14	0.07	0.00	0.02	0.07	−0.03	0.04	−0.04	0.05
Penalty total	0.09	0.04	0.13	0.02	0.03	0.04	0.13	0.03	−0.04
Cross-total	0.27***	0.09	0.04	0.11	0.13	−0.05	0.11	0.09	0.11

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Perf, Performance Rating by supervisor; E, Extraversion; C, Conscientiousness; ES, Emotional Stability; O, Openness to Experience; A, Agreeableness; P, Polychronicity.

3.3. Study 3

3.3.1. Purpose

Study 3 focused on evaluating the predictive and incremental validity of a SJT for managers. This SJT test for managers would typically be used for people who are managers in an outsourcing firm, where the outsourced work requires servicing consumer customers. For example, in a call center setup, managers have to simultaneously supervise a customer service team answering direct phone calls and also manage the expectations of the client who has outsourced the work. The traditional and alternative scoring methods are compared in this new context.

3.3.2. Instrumentation

In addition to the managerial SJT, data on English ability, quantitative ability, logical reasoning (all three of the preceding were measured by the Aspiring Minds Computer Adaptive Test – AMCAT), and personality factors (measured via the AMPI) were also collected.

The AMCAT was initially developed and validated in the following way. A domestic call center employing sales agents wished to establish success criteria for hiring these agents. A job analysis was conducted to understand the knowledge, skills, and other abilities

(KSOAs) required for the profile. Based on the analysis, English Language Skills, Logical Ability, and Personality Factors were deemed to be critical. Tests of these skills and preferences were developed and delivered to 56 employees in the company. The tests were developed both in English and the local language of the region. Objective sales figures for three months were shared which were combined to serve as the output variable. Logical ability ($r = 0.25$), Extraversion ($r = 0.37$), and Conscientiousness ($r = 0.29$) showed validity with the output variable (percentage of sales targets achieved). Together the modules provided a validity of 0.49 with performance. This matched the job analysis and global studies conducted for sales agents.

3.3.2.1. English ability. Familiarity with English Language and its various nuances is an essential skill, especially in the current climate of global networking. Ideally, any recruitment should involve a test of skills in handling the English language in ways that promote the objectives of a company and establish desired rapport. The English Comprehension Test is an adaptive test developed by Aspiring Minds used as part of the AMCAT. The Cronbach's alpha for the test is .77. The standard error for IRT-based adaptive test is 48 points on a 100–900 scale. The standard

Table 8. Incremental validity results from Study 2

	Multiple R	R square	Adjusted R square	Standard error
1. AMCAT	0.21	0.05	0.00	0.84
2. AMCAT + TraditionalAvoid	0.26	0.07	0.00	0.83
3. AMCAT + TraditionalApproach	0.24	0.06	0.00	0.84
4. AMCAT + Traditional Total	0.27	0.07	0.01	0.83
5. AMCAT + MatchAvoid	0.33**	0.11	0.05	0.81
6. AMCAT + MatchApproach	0.22	0.05	0.00	0.84
7. AMCAT + Match Total	0.28*	0.08	0.02	0.83
8. AMCAT + PenaltyAvoid	0.22	0.05	0.00	0.84
9. AMCAT + PenaltyApproach	0.27	0.07	0.01	0.83
10. AMCAT + Penalty Total	0.23	0.05	0.00	0.84
11. AMCAT + Cross Total	0.35***	0.13	0.07	0.81

Note: $N = 145$, * $p < .05$, ** $p < .01$, *** $p < .001$.

error is based on theoretical estimates arrived at by computer simulations. The IRT parameters of the items in the test are set according to the attempt of AMCAT by graduates across India with a sample of more than 50,000 test takers.

3.3.2.2. Quantitative ability. The Quantitative Ability test used in this study was developed internally and is a proprietary component of the Aspiring Minds Computer Adaptive Testing (AMCAT) program. The test consists of approximately 25 questions and takes approximately 35 min to complete. The Quantitative Ability section measures the candidate's numerical ability and accuracy in mathematical problems. The questions range from purely numeric calculations to problems of arithmetic reasoning, percentage analysis and quantitative analysis. Specifically these are divided into three sections:

Basic Numbers. This section tests whether the candidate has an understanding of basic number systems, that is, fractions, decimals, negative, positive, odd, even numbers, rational numbers, etc. The candidate should know how to do basic operations on these numbers.

Number Theory. This section requires a candidate to apply the concepts related arithmetic reasoning and basic algebra. It assesses the candidate's understanding on factors/divisibility and his/her ability to perform basic algebraic operations.

Applied Mathematics. Apart from operations on numbers, the candidate should know how to convert a real-world problem into equations, which could be solved to find an unknown quantity. Students need to be competent in reading and using quantitative data, in understanding quantitative evidence and in applying basic quantitative skills to the solution of real-life problems in order to perform effectively as professionals and citizens. To assess the same, the candidates are tested on Word Problems representing various real world scenarios.

The Cronbach's alpha for the test is .78. The standard error for IRT-based adaptive test is 52 points on a 100–900 scale. The standard error is based on theoretical estimates arrived at by computer simulations. The IRT parameters of the items in the test are set according to the attempt of AMCAT by graduates across India with a sample of more than 50,000 test takers.

Logical Ability. See Study 2 description.

Personality. See Study 2 description.

Performance Ratings. The dependent variable in the study was the performance ratings ('1' = low performer, '2' = medium performer, '3' = high performer) that were provided to individuals by their supervisors during the course of their annual performance evaluations.

These evaluation scores are part of the company's internal rating system.

3.3.3. Methods

The assessments were individually administered online via the Aspiring Minds assessment engine. Participants were required to take the assessment at an authorized testing center. The managerial SJT consisted of 24 items and took approximately 30 min to complete. Participants were not given any incentives for their participation, but in some cases participation was required by the participants' company. The SJT was administered as part of a larger test battery that also included tests of English, Quantitative Ability, Logical Reasoning, and Personality. The entire test battery took 2.5 hr.

3.3.4. Sample

The sample of participants was drawn from one of the largest business process outsourcing units in India. A group of the middle managers in the organization were asked by their supervisors to participate in the assessment. The sample was selected by the employer such that it proportionally represented the manager population based on gender, age, educational qualifications, and performance.

3.3.5. Results

A total of 49 participants (84% male, 16% female) completed the SJT for a response rate of 86%. The gender demographics were representative of the mid managers supervising different BPO processes in the organization, which had approximately 76% males. The mean age of participants in our sample was 31.5 years old ($SD = 4.4$ years) and the participants had worked for the organization an average of 22 months ($SD = 26.6$ months). Fully 77% of the sample held a Bachelor's degree, with 14% holding an MBA and 9% holding another type of degree. Table 9 reports the descriptive statistics for all variables associated with Study 3.

Table 10 presents the correlations among the 10 different scoring approaches for the Study 3 sample.

Table 11 presents the results of the regression analyses. Taken together, the standardized AMCAT test battery, consisting of assessments of English, quantitative ability, logical reasoning, and personality factors were used in a regression equation and predicted a total of 22% of the variance in supervisor's ratings of performance during an annual review. When the variable indicating managers' ability to identify a bad response ('Match Avoid') is added into the predictive equation after the standard AMCAT battery, it alone explains an extra 12% of the variance in supervisor ratings (for a total of 34% variance explained) and yields a statistically significant R square change. None of the other methods for scoring the managerial SJT

Table 9. Descriptive statistics for Study 3

Variables	Minimum	Maximum	Mean	Std. deviation
Extraversion	-2.56	.96	-.55	.86
Conscientiousness	-3.04	1.90	-.10	.99
Neuroticism	-1.76	2.77	.12	.91
Openness	-2.03	.82	-.13	.62
Agreeableness	-1.22	1.59	.43	.66
Quantitative	255.00	655.00	436.00	103.00
Logical reasoning	100.00	684.00	416.00	137.00
English	385.00	755.00	575.00	93.00
Rated performance	1.38	3.00	2.19	.34
Traditional avoid	-5.00	14.00	5.41	4.57
Traditional approach	-1.00	13.00	6.65	3.11
Traditional total	-1.00	24.00	12.06	6.37
Match avoid	1.00	15.00	8.53	3.24
Match approach	4.00	14.00	9.08	2.38
Match total	7.00	28.00	17.61	4.33
PenaltyAvoid	-8.00	.00	-3.12	1.86
PenaltyApproach	-8.00	.00	-2.43	1.54
Penalty total	-14.00	-1.00	-5.55	2.81
Cross total	-2.00	15.00	6.10	3.99

Note: $N = 49$.

resulted in statistically significant R square change values when added to the regression equation.

3.3.6. Discussion

The results of Study 3 show that the managerial SJT is a significant predictor of rated performance of managers in India. In addition, the managerial SJT demonstrated substantial incremental validity. Specifically, when used in conjunction with the AMCAT test battery the managerial SJT explains a total of 34% of the variance in supervisor ratings of managers performance, and this instrument alone increases the predictive power of the regression equation by more than 50%.

The results of Study 3 are consistent with the results of the prior studies with respect to the superiority of the participant's ability to correctly identify when to invoke the performance-avoid motivation ('avoid' score, see Table 12 for a comparison across studies). The results reveal one instance in which 'Match Approach' does the best among the methods for scoring the 'approach'

response. This finding is in contrast to the superior performance 'Penalty Approach' in the first two studies. This finding explains why the 'Match Avoid' score outperforms the 'Cross Total' in this study (Study 3). Nevertheless, the superiority of 'Match Avoid' among all other scores is observed here as well.

4. Discussion

The current paper set forth two main objectives. One of these objectives was to evaluate the cross-cultural validity of SJTs in a non-western context. Past research has demonstrated the predictive value of SJTs in western contexts (Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001; McDaniel et al., 2001). Yet, few studies have examined the utility of the technique in non-western samples, and even fewer have done so in India, the country with the world's second largest population. We investigated the cross-cultural validity of SJTs by running three studies that used two different situational judgment tests – one for sales and another for management – within the context of three unique samples: (1) relatively new salespeople in a mid-sized Indian company, (2) insurance salesmen selling directly to customers, and (3) middle managers in a large outsourcing firm in India. The results of these studies provide evidence in favor of both the predictive and incremental validity of SJTs within the Indian context.

With regard to our second main objective, we set out to argue for the superiority of a new approach to scoring SJTs that infused elements of Elliot's (1999) performance-approach and performance-avoid motivational theory with Sternberg and colleagues' (2000, 2006) cognitive concept of tacit knowledge associated with the theory of successful intelligence (1999). Specifically, we hypothesized that the cognitive ability to recognize when it is best to invoke a performance-avoid strategy would be the optimal predictor of key job outcomes. The results from all three studies reported here support our hypothesis and suggest that the ability to identify the 'worst' (i.e., 'avoid') answer is a systematically different skill than the

Table 10. Study 3: intercorrelations among 10 scoring techniques

Variable	1	2	3	4	5	6	7	8	9	10
1. Traditional_avoid	–									
2. Traditional_approach	.35*	–								
3. Traditional_total	.89**	.74**	–							
4. Match_avoid	.94**	.28	.81**	–						
5. Match_approach	.23	.88**	.59**	.17	–					
6. Match_total	.83**	.69**	.93**	.84**	.68**	–				
7. Penalty_avoid	.82**	.38**	.77**	.58**	.26	.57**	–			
8. Penalty_approach	.36*	.67**	.58**	.30*	.22	.35*	.36*	–		
9. Penalty_total	.74**	.62**	.83**	.55**	.29*	.57**	.86**	.79**	–	
10. Cross_total	.91**	.48**	.89**	.93**	.22	.82**	.61**	.63**	.75**	–

Note: $N = 49$, *Correlation is statistically significant at $p < .05$, **Correlation is statistically significant at $p < .01$.

Table 11. Incremental validity results from Study 3

Scores	Multiple R	R square	Adjusted R square	Standard error
1. AMCAT	0.47	0.22	0.06	0.33
2. AMCAT + TraditionalAvoid	0.55*	0.30	0.14	0.31
3. AMCAT + TraditionalApproach	0.45	0.21	0.02	0.33
4. AMCAT + Traditional Total	0.52*	0.28	0.11	0.32
5. AMCAT + MatchAvoid	0.58*	0.34	0.19	0.3
6. AMCAT + MatchApproach	0.46	0.22	0.03	0.33
7. AMCAT + Match Total	0.57**	0.33	0.17	0.31
8. AMCAT + PenaltyAvoid	0.47	0.22	0.05	0.33
9. AMCAT + PenaltyApproach	0.45	0.20	0.02	0.33
10. AMCAT + Penalty Total	0.45	0.21	0.02	0.33
11. AMCAT + Cross Total	0.53*	0.28	0.11	0.32

Note: $N = 49$.

ability to identify the 'best' answer (i.e., 'approach'). Further, the data suggest that it is not even necessary to have a consensually designated 'best' (i.e., 'approach') response, but only a designated 'worst' (i.e., 'avoid') answer in order for the scoring to show its strongest predictive validity. When scored as its own variable, the ability to identify an 'avoid' response is a statistically significant predictor of job performance. By contrast, when traditional scoring methods are used that conceive of the ability to identify the 'best' response and 'worst' response as polar ends of a single ability continuum, these approaches yield lower correlations with outcome variables and in many cases are not statistically significant. Interestingly, when the ability to identify an 'approach' response is used as its own variable, regardless of how it is operationalized, it is not a significant predictor of job performance. By contrast, the ability to correctly identify an 'avoid' response exhibits the highest validity in predicting job performance as well as incremental validity over and above English ability, logical reasoning, quantitative ability, personality, and domain-specific knowledge.

We would like to discuss one nuance here. We operationalize the ability to identify the worst answer in the context of two major scoring types. The first is 'Match Avoid' where the candidate gets a +1 if s/he chooses the avoid answer that corresponds with the designated avoid answer. The other, 'Cross Total', is the sum of 'Match

Avoid' and 'Penalty Approach'. Here the candidate gets a +1 if s/he correctly identifies the designated 'avoid' answer and gets penalized by a -1 if s/he chooses the designated 'avoid' answer as their 'approach' response. In Studies 1 and 2, the 'Cross Total' score shows higher validity than the 'Match Avoid' score. One advantage of the 'Cross Total' score is that it has a higher variance than 'Match Avoid' alone. Thus, while the 'Match Avoid' score is conceptually more straightforward to understand and explain, the 'Cross Total' method has certain statistical advantages that may be preferable.

Our finding that the ability to correctly identify an 'avoid' response is a significant predictor of performance is in line with the results of prior research using SJTs for teachers in the United Kingdom (Elliott et al., 2011) that has shown that teachers years of experience is significantly related to the ability to identify a bad response to a situation, but not a good response. It is worth highlighting the fact that the ability to identify a good response was not a significant predictor of either job performance or years of experience whereas the ability to identify a bad response was. One may note this is an indication that our theoretically-based results here appear to replicate in the context of a developed economy (the UK) as well as a developing economy (India).

We speculate that one reason for our findings may be the fact that what constitutes a 'good' (i.e., 'approach')

Table 12. Summary of Study 1–3 correlations between outcomes and scoring approaches

Variables	Study 1 ($N = 50$)	Study 2 ($N = 146$)	Study 3 ($N = 49$)
Traditional avoid	0.28*	0.16*	0.33*
Traditional approach	0.16	0.10	0.14
Traditional total	0.27	0.17*	0.31*
Match avoid	0.33*	0.26**	0.38**
Match approach	0.08	0.02	0.17
Match total	0.26	0.18*	0.38**
Penalty avoid	0.15	0.00	0.14
Penalty approach	0.25	0.14	0.03
Penalty total	0.23	0.09	0.11
Cross total	0.36*	0.27***	0.32*

Note: * $p < .05$, ** $p < .01$, *** $p < .00$.

response to a particular scenario may change from organization to organization, and indeed may be what gives each organization its unique climate. It may also be true that handling the situation one way or the other may not yield very different outcomes, until the way to handle it is entirely wrong. There could be multiple ways to handle the situation, which on average, will work equally well. However, the wrong way (i.e., 'avoid' response) to do things in a field is likely to lead to a negative outcome all the time and be relatively invariant across organizations, jobs and culture. This may provide some key to the puzzle of why SJTs are predictive and may also provide some hope for the future development of SJTs.

The results of our study may bring some clarity to the debate within the field of emotional intelligence research on whether using 'consensus-based' approaches to scoring or 'expert-based' approaches to scoring are superior. Zeidner, Matthews, and Roberts (2009) have speculated that consensus-based scoring, '...is likely to lead to special problems at the top end of the scale...' (p. 60), but that it may be useful for the lower end of the scale. The results from our studies confirm this intuition and bring some theoretical rationale to bear on the matter. It is likely that it is not so much the consensus-orientation of the scoring, but rather the focus on the 'avoidance' response that is of most value in the context of SJT scoring. This is likely because, as we have pointed out previously, it is easier for both experts and non-experts to come to agreement on avoidance responses whereas there are many possible 'best' answers across different organizations (i.e., there are different organizational norms). As the three studies in this paper demonstrate, the ability to correctly identify that 'worst' response tends to be a significantly better predictor of both objective (% of sales targets met) and subjective (supervisor ratings of performance) outcomes.

Different approaches to scoring can lead to different conclusions about the utility of SJTs. Those that focus on the search for 'best' scores using either 'expert' or 'consensus' methods may be less likely to find significant predictive value in SJTs than those focusing on the correct identification of 'worst' scores using either expert or consensus based scoring.

One of the major challenges in SJT research is the fact that each new domain requires the development of a new SJT. If the major contributions that SJTs make in terms of predictive power is that from the low end (i.e., what 'not' to do - 'avoid' response) rather than the high end (i.e., what 'to' do - 'approach' response) then this has important implications. It suggests that the development of SJTs can be relatively standardized in a way that has not been tractable to date.

The finding that the ability to identify bad responses is a systematically better predictor of performance outcomes than the ability to identify a good response may contribute to the theoretical advancement of the field as well.

Recently, there has been an explosion of research suggesting that 'dark traits' – those characteristics that are particularly undesirable for an individual to have in an employment context – are useful for predicting job performance (Harms, Spain, & Hannah, 2011; Hogan & Hogan, 2001; Jonason, Slomski, & Partyka, 2012; Paulhus & Williams, 2002; Veselka, Aitken Schermer, & Vernon, 2012). Specifically, traits such as Machiavellianism, narcissism, and psychopathy tend to be characteristic of individuals who do not perform well on the job. It would be interesting to merge the findings of that literature with SJT methodology. Stemler and colleagues (Stemler & Sternberg, 2006; Stemler et al., 2006) have suggested that SJTs should always be constructed based on theory and dark trait theory provides an interesting opportunity to align with the finding that the 'avoid' response tends to be the most predictive of performance. If one were to integrate the concept of dark traits into SJT construction, it is likely that 'avoid' response could be made to represent a behavioral manifestation of a 'dark trait'. Thus, one direction for future research would be to integrate the theoretical concept of dark traits into the generation of response options for SJTs. Our prediction is that this 'dark trait' would form the agreed upon 'avoid' option, which would, in turn, be a significant predictor of job performance.

The results of the current study raise some questions as to why other researchers, who have invoked different approaches to scoring, also find SJTs to be significant predictors of performance. One potential explanation for this finding is that the past research findings are mainly being driven by the part of the score that corresponds to the identification of the 'avoid' response. It would be a relatively easy matter, and an instructive point, for past researchers to break their scores into separate 'approach' and 'avoid' variables and re-run their data and evaluate the results. We predict that the 'avoid' element would be significantly predictive whereas the 'approach' element would have a weaker correlation to the output, but of course, further investigation into this conjecture is warranted.

Finally, recall the study by Plohart and Ehrhart (2003) in which the psychometric characteristics of results associated with six different instructions ('would do' vs. 'should do') were compared and it was found that the 'would do' instructions had superior psychometric characteristic. One quirk of that study is that all six approaches to scoring were the same in that they each treated the capacity to identify a good response and the capacity to identify a bad response as polar ends of a single ability spectrum. This was operationalized by invoking the penalization and summation approach to scoring (i.e., the 'traditional approach' we note above). Our data from this paper suggest that the instructions themselves may not have been the issue driving their finding. Rather, their approach to scoring may have been driving the null result. That is, perhaps the 'would do' instructions are more useful for capturing the ability to identify an 'approach' response whereas the 'should do'

instructions are more useful for capturing the ability to identify a 'avoid' response. It would be interesting to see a further investigation of their data, re-scored to conceptualize the ability to identify a bad (i.e., 'avoid') response separated from the ability to identify a good (i.e., 'approach') response, in order to test our hypothesis.

4.1. Limitations

There are some potential limitations to the studies reported herein. First, because all of the samples were drawn from India, the data reported herein represents something of a contrast to most of the extant literature on SJTs, which is largely derived from western samples. Consequently, it is possible that there is a cultural effect driving the finding that the identification of an 'avoid' answer is more important than the identification of an 'approach' response.

Next, although Study 2 used three predictor variables that were theoretically expected to correlate with an employee's ability to reach sales targets, the combined variance explained by those variables was strikingly low. However, there are two potential explanations for this finding. First, prior research has shown that objective sales metrics tend to show low correlations with predictor variables whereas the correlations tend to be stronger with subjective measures, such as managerial ratings of employee performance (Skyrme, Wilkinson, Abraham, & Morrison, 2005). Our current study replicates this finding. The reason for lower correlations of performance with objective measures is partially because objective indicators are typically influenced by factors that are outside of the direct control of the participant. Consequently, such measures may only reflect one aspect of the performance construct.

In addition, the lower correlations found in Study 2 could be due to the fact that the company may already be selecting employees using personality, insurance readiness measures and as such, there would be restricted range on these variables within the company which would attenuate correlation and regression coefficients. Consequently, it is difficult to say for sure how much the increased incremental validity added by the SJT is due to the weak performance of those particular predictor variables. Nevertheless, the results from Study 3 provide us with some confidence in this regard as the incremental validity was even stronger when the explained variance from other predictors was stronger.

4.2. Conclusion

In conclusion, we see a bright future ahead for the use of SJTs. Some of the pieces of the puzzle of why they are predictive are beginning to come together. The SJT methodology has many advantages and has demonstrated validity in western and non-western contexts. A standardized

approach to scoring and development seems within reach and the infusion of new theories of personality into the construction of SJTs may help to enhance their utility and provide further clarity on the reasons for their predictive power.

Notes

1. There is other work around how to score SJTs primarily when each option is graded on a scale. That is beyond the scope of this paper but does not change the primary findings of the paper.
2. Note that the AMPI was subsequently administered to a sample of 88 US college students. Six factors, accounting for 39% of the total variance, emerged from a principal components analysis with varimax rotation. Each factor accounted for 5% or more on its own and the items from the AMPI loaded in accordance with theoretical predictions from the Five-Factor + Polychronicity model. The scale reliabilities from the AMPI for the US sample were: (O = .66, C = .63, E = .62, A = .65, N = .77, P = .77).

References

- Arvey, R. D. & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology*, 35, 281–322.
- Bass, B. M. (1990). From transactional to transformational leadership: Learning to share the vision. *Organizational Dynamics*, (Winter), 19–31.
- Burns, J. M. (1978). *Leadership*. New York: Harper & Row, OnursalArkan.
- Central Intelligence Agency World Factbook (2014). India. Available at <https://www.cia.gov/library/publications/the-world-factbook/geos/in.html>
- Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Schmidt-Harvey, V. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, 86, 410–417.
- Conte, J. M., & Jacobs, R. R. (2003). Validity evidence linking polychronicity and big five personality dimensions to absence, lateness, and supervisory performance ratings. *Human Performance*, 16(2), 107–129.
- Cucina, J. M., Vasilopoulos, L., & Leaman, J. A. (2003, April). *The bandwidth-fidelity dilemma and situational judgment test validity*. Paper presented at the 18th annual meeting of the Society for Industrial and Organizational Psychology (SIOP), Orlando, Florida.
- Dalessio, A. T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology*, 9(1), 23–32.
- Dweck, C. (1986). Motivational processes affecting learning. *American Psychologist*, 41, 1040–1048.
- Elliot, A. J. (1999). Approach and avoidance: Motivation and achievement goals. *Educational Psychologist*, 34(3), 169–189.
- Elliott, J. G., Stemler, S. E., Grigorenko, E. L., Sternberg, R. J., & Hoffman, N. (2011). The socially skilled teacher and the development of tacit knowledge. *British Educational Research Journal*, 37(1), 83–103.

- Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, 51(4), 327–358.
- Harms, P. D., Spain, S. M., & Hannah, S. T. (2011). Leader development and the dark side of personality. *The Leadership Quarterly*, 22, 495–509.
- Hausknecht, J. P., & Langevin, A. M. (2010). 'Selection for Service and Sales Jobs'. Articles and Chapters. Paper 190. Available at <http://digitalcommons.ilr.cornell.edu/articles/190>
- Hogan, R., & Hogan, J. (2001). Assessing leadership: A view from the dark side. *International Journal of Selection and Assessment*, 9, 40–51.
- Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, 79, 184–190.
- Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment*, 21, 264–276.
- Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, 29, 340–362.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96, 72–98.
- Jonason, P. K., Slomski, S., & Partyka, J. (2012). The dark triad at work: How toxic employees get their way. *Personality and Individual Differences*, 52, 449–453.
- Kaplan, R. M., & Saccuzzo, D. P. (2013). *Psychological testing: Principles, applications, and issues*. (8th ed.) Belmont, CA: Wadsworth.
- Kantrowitz, T. M., Grelle, D. M., & Beaty, J. C. (2012). Time is money: Polychronicity as a predictor of performance across job levels. *Human Performance*, 25(2), 114–137.
- Lewin, K., Dembo, T., Festinger, L., & Sears, P. S. (1944). Level of aspiration. In J. M. Hunt (Ed.), *Personality and the behavioral disorders* (Vol. 1, pp. 333–378). New York: Ronald Press.
- 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.
- McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1), 81–90.
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, 60, 63–91.
- McDaniel, M. A., Morgerson, 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*, 80, 730–740.
- Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, 75(6), 640–647.
- Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). A theoretical basis for situational judgment tests. In J. A. Weekley and R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application*. (pp. 57–82). Mahwah, NJ: Lawrence Erlbaum Associates.
- Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of Research in Personality*, 26, 556–563.
- Phillips, J. F. (1992). Predicting sales skills. *Journal of Business and Psychology*, 7(2), 151–160.
- Phillips, J. F. (1993). Predicting negotiation skills. *Journal of Business and Psychology*, 7(4), 402–411.
- 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), 1–16.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2), 262–274.
- Schneider, B., & Konz, A. M. (1989). Strategic job analysis. *Human Resource Management*, 28, 51–63.
- Skyrme, P., Wilkinson, L., Abraham, J. D., & Morrison, J. D. (2005). Using personality to predict outbound call center job performance. *Applied H.R.M. Research*, 10(2), 89–98.
- Stemler, S. E., Elliott, J. G., Grigorenko, E. L., & Sternberg, R. J. (2006). There's more to teaching than instruction: Seven strategies for dealing with the practical side of teaching. *Educational Studies*, 32(1), 85–102.
- Stemler, S. E., & Sternberg, R. J. (2006). Using situational judgment tests to measure practical intelligence. In J. A. Weekley and R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application*. (pp. 107–131). Mahwah, NJ: Lawrence Erlbaum Associates.
- Sternberg, R. J. (1999). The theory of successful intelligence. *Review of General Psychology*, 3, 292–316.
- Sternberg, R. J., & Horvath, J. A. (Eds.). (1999). *Tacit knowledge in professional practice*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., et al. (2000). *Practical intelligence in everyday life*. Cambridge, UK: Cambridge University Press.
- Veselka, L., Aitken Schermer, J., & Vernon, P. A. (2012). The dark triad and an expanded framework of personality. *Personality and Individual Differences*, 53, 417–425.
- Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology*, 32, 1236–1247.
- Weekley, J. A., & Ployhart, R. E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance*, 18, 81–104.
- Weekley, J. A., & Ployhart, R. E. (Eds.). (2006). *Situational judgment tests: Theory, measurement, and application*. The Society for the Industrial and Organizational Psychology: Organizational Frontiers Series. Mahwah, NJ: Lawrence Erlbaum Associates.
- Zeidner, M., Matthews, G., & Roberts, R. D. (2009). *What we know about emotional intelligence: How it affects learning, work, relationships, and our mental health*. Cambridge, MA: MIT Press.