EVALUATION OF THE SITUATIONAL JUDGMENT TEST

Lane A. Conner, B.S.

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

May 2007

APPROVED:

Charles A. Guarnaccia, Major Professor
Linda Marshall, Committee Member and Chair of the Department of Psychology
Rodger Ballentine, Committee Member
Sandra L. Terrell, Dean of the Robert B. Toulouse School of Graduate Studies

Conner, Lane A., Evaluation of the Situational Judgment Test. Master of Science (Psychology), May 2007. 43 pp., 12 tables, references, 21 titles.

This research attempts to confirm the reliability and construct validity of a personnel selection instrument called a Situational Judgment Test (SJT) through reliability analysis and factor analysis. The existing literature on SJTs is reviewed, including the advantages of using SJTs in personnel selection as well as the debate on whether SJTs measure a single construct or whether they can be multidimensional depending on the content. The specific SJT in this research was theoretically developed and received expert ratings to assess four general constructs: problem solving, planning, priority setting, and leadership. No support from alpha internal consistency reliability analysis was found for the assembly of these items into the four a priori subscales, thus assembly of these items into the theoretical subscales and scales was not supported.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Industrial/Organizational (I/O) psychologists and Human Resources (HR) professionals use numerous types of assessments to make personnel selection decisions in organizations. These instruments vary greatly in form, cost of administration, validity and reliability. Obviously, the goal is to use assessment procedures that reliably select individuals from a pool of job applicants who will succeed in the position they are filling. That is, they are attempting to maximize predictive validity and face validity all while minimizing costs and adverse impact (adverse impact occurs when a selection method is shown to select certain demographic groups at higher rates than other demographic groups, Gatewood & Field, 2001).

Situational Judgment Tests

Situational Judgment Tests (SJTs) have received attention as useful selection solutions for addressing many of these criteria. In McDaniel, Morgeson, Finnegan, Campion, and Braverman's (2001) meta-analysis of SJTs, they "broadly define a Situational Judgment Test as any paper-and-pencil test designed to measure judgment in work settings." In these tests, "a scenario is described and the respondent must identify an appropriate response from a list of alternatives" (p 730). They go on to include measures that "do not present situations, but rather require respondents to indicate their level of agreement with statements concerning the appropriateness of various work-related behaviors" (p. 730). Although the latter does technically assess an individual's judgment of a situation in a work setting, the former best represents is what is traditionally thought of when SJTs are developed and validated for use in personnel selection.

Situational judgment tests usually present a scenario (approximately one paragraph in

length) describing a problem situation (typically in a work setting) in which multiple courses of action are plausible. After reading the scenario, the participant is asked a series of questions about the scenario. For each question the participant is given a list of plausible behavioral responses, each of which is correct to varying degrees. The participant is then asked to indicate either which response he or she believes to be best or which response he or she would most likely choose. Some SJTs also require the participant to indicate the worst possible response as well (Weekley & Ployhart, 2005). SJTs are typically scored by having a group of subject matter experts (SMEs) decide, a priori, about the relative effectiveness of each of the various behavioral response options and subsequently assigning ordinal weights to each response option. Alternatively, scoring methods can collect central tendency statistics of participant responses and assign effectiveness weights based on these responses. Finally, empirical-scoring methods based on participant responses and performance criterion can be used as well (McDaniel & Nguyen, 2001).

With technological advancements in assessment administration methods, SJTs are not exclusively administered in paper-and-pencil format as the previous definition suggests. In fact, Chan and Schmitt (1997) found video-based SJTs were significantly more face valid and showed significantly lower levels of adverse impact than the traditional paper-and-pencil version of the same SJT. Also, Konradt, Hertel, and Joder (2003) developed and concurrently validated (r = .33) a web-based SJT, which also yielded strong face-validity and acceptance.

Advantages of SJTs

There are numerous reasons SJTs have become commonplace in personnel selection. First, real work is multidimensional in nature. An individual's actual job performance is a heterogeneous criterion comprised of multiple behaviors across multiple settings. Therefore an

instrument such as a SJT, which is by nature a heterogeneous measure, is likely to more thoroughly tap the criterion conceptual space (Chan & Schmitt, 2002) than a homogenous measure such as a test of cognitive ability. There is not just theoretical, but data-driven evidence for the utility of SJTs as predictors of job performance as well. A meta-analysis of Situational Judgment Tests by Clevenger, Pereira, Wiechmann, Schmitt, and Schmidt-Harvey (2001) reported a criterion related validity coefficient of .34. While reported validities for SJTs typically fall in the .20 to .40 range, McDaniel et al. (2001) estimate the true validity of SJTs in predicting performance related criteria to be near .56 when corrected for range restriction and criterion unreliability. This is, as Clevenger et al. (2001) point out, among the best validity coefficients of all the selection instruments available.

Another reason SJTs are useful personnel selection instruments is, as Motowidlo, Dunnette, and Carter (1990) point out, that they can approximate the predictive validity of more "high-fidelity" assessments such as work samples and assessment centers at a fraction of the cost. High fidelity simulations use realistic materials and equipment to replicate a task and allow participants to respond almost exactly as they would were they to encounter the situation on the job. Such measures can be quite costly and time consuming to develop and implement. Based on their findings, Motowidlo et al. (1990) concluded that "low-fidelity" simulations such as SJTs can be empirically valid and can capture the predictive potential of more high-fidelity behavioral samples. Thus, when the time and cost requirements for the development and validation of a work sample or an assessment center prevent them from being a realistic option, human resource professionals may turn to the considerably less expensive low-fidelity situational judgment test without losing any predictive ability.

In addition, measures of situational judgment are impressive options for inclusion in

personnel selection batteries because they consistently have lower levels of adverse impact and higher levels of face validity than other commonly used selection instruments such as tests of General Cognitive Ability (GCA). For example, the SJT developed by Motowidlo et al. (1990) did not yield statistically different scores between black and white individuals on any of the samples examined, nor were the differences between men and women significant. Pulakos and Schmitt (1996) created a SJT in which African American individuals scored 0.41 standard deviations lower than Caucasian individuals. While these differences are still less than ideal, they are considerably better than the one standard deviation difference typically found between African Americans and Caucasians on tests of GCA. Similarly, Weekley and Jones (1999) also found support for their hypothesis that the differences between Caucasian and minority scores on their SJT would be smaller than the differences in scores on the cognitive ability tests.

Finally, as mentioned, there is both theoretical and empirical evidence that SJTs have higher face validity than traditional assessments of individual differences. As Clevenger et al. (1990) point out, the item construction process involved in creating SJTs yields item content which has higher face validity than other individual difference tests (e.g., abstract reasoning problems found on a test of GCA). Items are either developed from actual critical incidents which have occurred in the past or based on SMEs' identification of scenarios which are plausible and realistic. Motowidlo et al. (1990) found SJTs have more face validity and appear to be a more appropriate and job-relevant selection tool to job applicants than traditional individual difference measures

What SJTs Measure

A review of the situational judgment literature reveals there are two basic stances regarding what SJTs are and what they measure. One stance, led by Sternberg, Wagner, and

colleagues (e.g., Sternberg & Hedlund, 2002, Sternberg, Wagner, Williams, & Horvath, 1995, Wagner and Sternberg, 1985) contends that SJTs measure a construct in and of itself, independent of any other individual difference variables. Specifically, they refer to this SJT construct as tacit knowledge. On the other hand, many scholars such as McDaniel and colleagues contend that SJTs are an assessment method similar to selection interviews and assessment centers which can and do measure various job-related constructs, depending on the content of the test (see Chan & Schmitt, 2002, Clevenger et al., 2001, McDaniel et al., 2001, McDaniel & Nguyen, 2001, Smith & McDaniel, 1998). Each of these positions will be examined in turn.

Tacit knowledge perspective. According to Sternberg and Hedlund (2002), tacit knowledge is actually a subcomponent of the more general term "practical intelligence," although the two terms are often confused with one another (explicating the relationship between practical intelligence and tacit knowledge is beyond the scope of this discussion; for more on this topic, see Stemler and Sternberg, 2006). Tacit Knowledge is defined as "knowledge that reflects the practical ability to learn from experience and to apply that knowledge in pursuit of personally valued goals," (p 145; Sternberg & Hedlund, 2002). Most importantly, tacit knowledge is supposedly unrelated to general cognitive ability (GCA). Sternberg et al. (1995) contend there are three defining characteristics which differentiate tacit knowledge from traditional academic knowledge. First, these authors argue that tacit knowledge is acquired with little or no help from others. The individual is not directly instructed as to what should be learned. Second, tacit knowledge is procedural in nature ("knowing how" as opposed to "knowing that"), it guides an individual's actions without being easily articulated. Finally, it has direct relevance to an individual's goals. This type of knowledge which is based on practical experience will supposedly be more helpful to the attainment of personal goals than academic knowledge which

is learned in a classroom (Sternberg et al., 1995).

In research such as Sternberg and Hedlund (2002) and Stemler and Sternberg (2006) which supports the argument SJTs are measures of tacit knowledge and are largely independent of intelligence and other individual difference variables, Sternberg and colleagues cite only certain research findings (e.g., Wagner 1987, Wagner & Sternberg 1985, Wagner & Sternberg 1990) to support their viewpoint. Much of this work is methodologically questionable, using small sample sizes (e.g., $N = 22$ and $N = 45$ in two studies) where range restriction is highly probable, and uses questionable criterion as outcome variables. For example, samples were drawn from Yale University undergraduates, academic psychologists, and business executives, all of which are samples where range restriction is highly likely. Further, Sternberg and colleagues (Sternberg & Hedlund, 2002, Sternberg & Wagner 1990, Stemler & Sternberg, 2006,) argued that SJTs are generally not correlated with personality. However, of the personality measures included in the 1990 study, none of the criterion were measures of the Five Factor Personality model (e.g., Costa & McCrae, 1992), the most prevalent personality model in use today.

It should be noted considerable research contradicts the findings that measures of tacit knowledge are unrelated to GCA. For example, McDaniel et al's (2001) meta-analysis found 80 correlations between scores on SJTs and scores on measures of GCA. The only SJTs included in Stemler and Sternberg's (2006) discussion of the validity of tacit knowledge as the construct being measured by SJTs are the so-called "Tacit Knowledge SJTs" (TKSJTs), such as the Tacit Knowledge Inventory for Managers (Wagner & Sternberg 1991). Sternberg and colleagues' do not include this research on SJTs although no meaningful methodological difference can be found between TKSJTs and all of the "regular" SJTs included in the McDaniel et al. (2001)

meta-analysis. McDaniel et al. (2001) and McDaniel and Whetzel (2005) both took multiple sample items from various SJTs, (e.g., the Teamwork-KSA test, Stevens & Campion, 1999) and did side-by-side comparisons with multiple items from the Tacit Knowledge Inventory for Managers (for the actual item-to-item comparisons, see the aforementioned articles). McDaniel et al. (2001) conclude that while SJT items are not completely identical to TKSJT items, they do measure similar content using similar methodology. McDaniel and Whetzel (2005) conclude Sternberg's tacit knowledge items are ultimately not anything new; they are simply one application of situational judgment items.

It should also be mentioned that when McDaniel et al. (2001) removed the TKSJT correlation coefficients from the meta-analytic data, the results were virtually identical—that is, removing the supposedly low TKSJT correlation coefficients from the meta-analysis calculations did not lower the overall correlation between SJTs and GCA. Finally, in Sternberg and Hedlund (2002) the authors themselves acknowledge that the "measurement instruments used to assess tacit knowledge…have been characterized in the literature as situational judgment tests," (p 147); however, in the very next sentence authors state that situational judgment tests are synonymous with "tacit-knowledge (TK) tests" (p. 147), although it appears the only major difference between TK tests, and other SJTs is TK tests are developed by either Wagner or Sternberg.

In Weekly and Ployhart's 2006 book (which is essentially the compilation of all that is currently known about SJTs) Sternberg and in this case, Stemler (Stemler & Sternberg, 2006) address some of the contradictions between their findings and other research. For example, when addressing the fact that their studies of tacit knowledge have not corrected for range restriction despite the fact that their samples represent a small portion of the population and thus are not

generalizable to the entire population, Stemler and Sternberg (2006) merely state that they believe "the assumptions underlying corrections are somewhat dubious" (p. 120). Further, when addressing the previously mentioned issue that when they asserted measures of tacit knowledge are not correlated with personality (in Wagner and Sternberg, 1990), the criterion did not include the "full range of personality measures one might use," (as mentioned, they did not include any measure of the Five Factor Model (e.g., Costa & McCrae, 1992), the prevailing model used in personality theory today), they acknowledge that "it is *quite possible* that other measures of personality would yield correlations higher than those few we have used" [italics added] (p. 121). In fact, McDaniel and Nguyen's (2001) meta-analysis of the literature on the relationship of the Big Five personality variables with SJTs found correlations of .31 (Emotional Stability), .26 (Conscientiousness), .25 (Agreeableness), .09 (Extroversion), and .06 (Openness).

With regard to the fact that Sternberg and colleagues in their research do not address the .46 correlation found between SJTs and GCA in the McDaniel et al. (2001) meta-analysis, Stemler and Sternberg (2006) explain the issue by stating that while their conclusions are consistent with some results in the literature, they "*appear at face value*, to be less consistent with other results, for example, of McDaniel et al. (2001). We believe that some of the inconsistencies are *surface inconsistencies*" [italics added] (p. 121). It may be that this one statement does not effectively, nor sufficiently address such remarkably contradictory findings

Measurement method perspective. Scholars such as McDaniel who argue SJTs are a measurement method simply point to the moderate to strong correlations of SJTs with various individual difference variables as their argument that SJTs cannot be a construct unto themselves independent of correlations with other constructs. Research attempting to unearth the constructs being measured by SJTs has typically focused on three specific constructs, specifically GCA, job

knowledge/experience, and various aspects of personality, and has produced generally positive results. First of all, as mentioned, the McDaniel et al. (2001) found a corrected mean observed correlation between SJTs and measures of GCA of .46 (uncorrected for item attenuation and range restriction .36).

SJTs have also been found to be correlated with job knowledge and job experience. Weekley and Jones (1999) found significant correlations between scores on SJTs and job experience of .26 and .16 in two separate studies. As McDaniel and Nguyen (2001) note, measures of job knowledge are usually operationalized as measures of job experience. It comes as no surprise then these two related constructs have both been found to be positively correlated with scores on SJTs.

As mentioned, the McDaniel and Nguyen (2001) meta-analysis also calculated the average relationships found between SJTs and the various Big Five personality dimensions and found correlations of .31 (Emotional Stability), .26 (Conscientiousness), .25 (Agreeableness), .09 (Extroversion), and .06 (Openness). For this meta-analysis, it should be noted only three correlations between Openness and SJTs could be found, whereas the calculations for the other four personality constructs were based on 8 to 13 correlation coefficients. More recently, Chan and Schmitt (2002) found correlations of .23, .24, and .29 with Conscientiousness, Extraversion, and Agreeableness, and found that Neuroticism correlated -.20 with their SJT.

Incremental Validity of SJTs

While the relationship between SJTs and various constructs has been well documented, what is perhaps more important to note is the fact that SJTs have been shown to predict job performance over and above the individual difference variables discussed above (e.g., Clevenger et al., 2001, Chan & Schmitt, 2002, Weekley & Jones, 1999). For example, Clevenger et al,

(2001) found that in two of their three experiments, the SJT added significant incremental validity over GCA, job knowledge, job experience, and Conscientiousness combined (partial r = .17 and .14). Chan and Schmitt (2002) went one step further, including all Big Five personality variables and found the SJT used in their study provided substantial incremental validity in predicting performance relative to the validity offered jointly by cognitive ability, all Big Five personality traits and job experience (partial r = .21).

Summary and Conclusions

First of all, SJTs do seem to have utility as personnel selection instruments; their competitive criterion related validity, appealing cost to benefit ratio, lower adverse impact figures and higher applicant approval relative to other selection methods suggest SJTs deserve the attention they are receiving from both Industrial/Organizational psychologists and human resource professionals. In general, there are two positions as to what SJTs are and what they measure. One position, held by scholars such as Sternberg and Wagner, is that SJTs are a unique construct unto themselves; that is, that SJTs measure one construct, tacit knowledge, which is supposedly not correlated with any other individual difference variables. The preponderance of the research however, seems to support the other position held by scholars such as McDaniel, that Situational Judgment Tests are measurement methods which, depending on the actual content of the test, measure a variety of constructs. Specifically, research has shown SJTs to be moderately to strongly correlated with individual difference variables like General Cognitive Ability, various Big Five personality variables, and job knowledge and experience.

The reader must keep in mind that the research discussed in this section only refers to Situational Judgment Tests in a collective sense. What is actually being assessed by any given SJT is largely dependent on the content of the situations and items that compose that test. That

being said, this author tends to agree with Schmitt and Chan (2006) who argue that just as the structured interview, by its nature, is always to some extent going to measure an individual's oral communication skills and personal composure, so too are SJTs likely to assess other constructs a majority of the time. Schmitt and Chan (2006) propose that the underlying construct(s) universally assessed by SJTs can be termed "practical intelligence." The author disagrees with this contention. Rather, based on the research reviewed, it seems likely that all SJTs tap an individual's general intelligence and aspects of their personality, at least to some extent.

The incremental validity SJTs provide above and beyond GCA, personality and experience suggest that SJTs are at least to some extent tapping some additional construct beyond the typical individual difference variables. Here again, Chan and Schmitt (2006) contend this additional construct is best termed practical intelligence. As previously mentioned, the author disagrees with the use of the term practical intelligence in association with SJTs. According to Sternberg and Hedlund (2002), practical intelligence, as previously mentioned, is supposedly composed of an individual's tacit knowledge (amongst other things), which they contend is unrelated to GCA. As discussed earlier, the correlation between SJTs and GCA has been well established. Future research should seek to identify the additional construct SJTs assess and accurately and label it.

Proposed Research

This research attempts to further the understanding of SJTs by developing and validating a situational judgment measure that assesses job-related constructs other than general cognitive ability and job knowledge/ experience. This SJT was developed for use as a selection instrument for the retail sector; its scenarios and constructs assessed are "generic" enough to be implemented in multiple retail settings for multiple companies. Thus, the purpose for this

research and development of this SJT was to create an instrument that could assess constructs that are relevant for nearly all retail environments, regardless of the specific products or services being sold. The authors (a consulting firm in the Southeastern United States) of this Situational Judgment instrument worked in conjunction with multiple large retail organizations to identify those dimensions essential for success in retail settings, regardless of the specific retail industry. Ultimately, four essential constructs were identified: problem solving, planning, priority setting, and leadership.

For the retail organizations, the benefit of participating in this process is that if a valid tool can be developed, they will be able to use the validated SJT as a selection instrument. The goal for the consulting firm that led in the development of this instrument is to create a valid off-the-shelf SJT that can be sold for profit to other retail organizations.

Certain aspects support the development of a generic SJT broadly applicable to multiple retail organizations. In order to develop the content for situational judgment tests, many instrument developers first perform detailed job analyses and identify critical work situations, which yield SJTs that are only applicable in the specific job for which the scenarios are based on. The development of a SJT which uses generic work situations yet still assesses specific work-related competencies could hypothetically keep human resource professionals and industrial and organizational psychologists from having to invest considerable time and money in developing a SJT for just one specific job, yet still maintain the predictive validity situational judgment measures offer. Additionally, based on McDaniel et al. (2001) meta-analytic findings, tests with less detailed questions had higher validity than SJTs with more detailed questions. Thus, a generic SJT such as the one developed in this research might offer even more predictive validity than traditional, more detailed SJTs.

CHAPTER 2

METHOD

Participants

Archival data from 182 minority college undergraduates currently in retail and sales internship positions were used for this analysis. The Situational Judgment Test (SJT) was initially sent to 722 individuals, yielding a response rate, with the final sample of 182, of 25.2%. This final sample consisted of 50 seniors, 84 juniors, 31 sophomores, 2 freshman, 10 non-graduating seniors, and 5 respondents indicated "other" or did not respond. Although specific ethnicity information was not available, because the sample was collected with the cooperation of a non-profit organization placing minority college students in internship positions, every participant was a member of a minority group protected by the Equal Employment Opportunity Commission, e.g., African American, Native-American, Asian-American, or Hispanic. Age and gender information was not available, but the sample was collected from a predominantly college-age pool of interns.

Measure Development

The situational judgment measure (attached) was developed based on competency models, training content, and substantial input from subject matter experts (SMEs). The overall focus of the assessment was on broad critical thinking, problem solving skills, leadership, planning, and interpersonal skills. The assessment consisted of scenarios that required participants to utilize these skills in situations they might actually encounter on the job. The scenarios, question items, and possible responses were developed based on SME input from the various client organizations in which the non-profit organization placed its interns. The SMEs were either job-knowledgeable experts one to two levels above the target position or high-

performing incumbents. All SMEs had at least two years of job experience in the target job. After target positions and SMEs were identified, 4-6 SMEs conducted 12 in depth interviews with high performers, to allow SMEs to further understand environmental issues, job challenges, and performance requirements associated with the jobs. In addition, these interviews allowed SMEs to determine critical sales and leadership competencies associated with the positions. Finally, these interviews allowed SMEs to obtain critical incidents for scenario and item development. After these interviews, the SMEs developed the actual scenarios and questions. Ultimately, this process led to the development of six scenarios assessing four a priori factors, Leadership Effectiveness, Planning, Problem Solving, and Interactive Skills. Each item was written to theoretically assess one of the above four key competencies identified. Items are linked to these four competencies as shown in Table 1. After the scenarios and question items were reviewed and approved by the non-profit organization's staff, focus groups were conducted with 4-6 new SMEs to finalize answers and response item weights. After all revisions were finalized, the web-based test format was developed and a small pilot test was administered. The test was then sent to participants.

How the SJT is Scored

As previously mentioned, after the scenarios and questions were developed, the second group of SMEs came together to develop four responses for each of the 33 items in the instrument. Each possible item response could be considered "plausible," that is, each answer choice developed by the SME group could be considered a correct answer to varying degrees. For each question, an answer choice was developed that was considered "best," (given a weight of 4 points), another created as "second best" (given a weight of 3 points), one created as "third best" (weight of two), and a final solution which was developed as the "fourth best" solution

(weight of one). Participant total score was calculated by summing the weights of answer choice selected for each question; the higher the score the better the performance on the test.

This research hypothesizes that the factor analysis will result in a four-factor solution representing Problem Solving, Planning, Priority Setting, and Leadership. This theoretical mapping of items onto factors is shown in Table 1. The authors of the SJT attempted to assess these specific factors because these skills were consistently identified by SMEs from the multiple retail organizations as key competencies for success in retail settings, regardless of the specific retail industry.

Procedures

After the instrument was developed, a pilot test was conducted to ensure administration would go as planned. After the pilot test, the instrument was administered via the internet. Sales and retail interns were contacted by email and asked to complete the instrument within a two-week period. To gather more respondents, incentives were offered and the deadline was extended. Additionally, a second administration of the instrument was conducted a few months after the preliminary administration in order to gather more participants. The participants were informed that the instrument would be used for research and data collection purposes only and that in no way would performance on the test impact job performance rating.

CHAPTER 3

RESULTS

Descriptive Statistics

Counts and percentages for rankings of scores for each of the 33 items, which are nested within the six scenarios, were calculated. So too were various descriptive statistics such as means and SDs for the set of items within each scenario. In addition, the four a priori subscale means and SDs were calculated (Leadership Effectiveness, Problem Solving, Planning, and Interactive Skills), along with means and SDs for two overall scores (mean of the four subscales and the mean of all the items), along with the means and SDs for the six scenarios ("Leading a Team," "Coaching," "Improving Performance," "Peer Leadership," "Business Savvy," and "Team Member"). This summary data is contained in Table 2.

Finally, reliabilities for a priori subscales, with item-level reliability analysis (inter-item correlations, corrected item-total correlations, alpha if item deleted) were calculated. These reliabilities along with means and standard deviations are shown in Tables 3, 4 and 5. The alpha internal consistency reliabilities were calculated to evaluate the psychometric integrity of the overall SJT and its subscales and scenarios. The Cronbach's alpha internal consistency reliabilities four subscales of the SJT were very low (Problem Solving, $\alpha = .11$; Planning, $\alpha = .26$; Priority Setting, $\alpha = .16$; Leadership, $\alpha = .30$). The reliabilities for the six scenarios were equally poor ("Leading a Team," $\alpha = .17$; "Coaching," $\alpha = .05$; "Improving Performance," $\alpha = -.02$; "Peer Leadership," $\alpha = .15$; "Business Savvy," $\alpha = .29$; and "Team Member," $\alpha = .18$). The two overall SJT measures had slightly better alpha internal consistency reliability, but well below an acceptable level (subscale mean, $\alpha = .40$; total item mean, $\alpha = .47$).

Inferential Statistics

A lower triangle correlation matrix of item rankings between both the four subscales and then between the six scenarios with p levels was calculated. Small but statistically significant correlations were found between certain subscales and between certain scenarios. Tables 6 and 7 contain this information. In addition, an Exploratory Factor Analysis (EFA) was conducted to see if the theoretical four-factors of the SJT exist as claimed.

Exploratory Factor Analysis

Principal Components Factor Analysis was completed attempting to extract four factors, for the four a priori subscales, as suggested by the theoretical construction of the scale. There was an attempt to extract 6 factors as well, since there were 6 scenarios. In review of the eigenvalues, variance explained table, and the scree plots, no support was found for any factor structure. In particular, no discernible knee was found in the scree plots, in conjunction with no associated break points in eigenvalues or change in explained variance.

In an attempt to discern if any meaningful structure existed the rotated factor matrix was reviewed to judge any apparent validity to the item grouping. To accomplish this qualitative analysis, two knowledgeable individuals on situational judgment tasks, the author and a committee member, evaluated the items grouping of both the four- and the six-factor solutions to discern the presence of a theoretical structure. These two individuals concluded that although no meaningful grouping could be gleaned, in general it was observed that those answer choices which were more collaborative in nature, e.g., collecting input from others before acting, acting in accordance with others, etc. were typically given higher weights than those possible solutions which involved acting without taking collaborative steps. This observation will be explored further in the discussion section.

CHAPTER 4

DISCUSSION

Review of Item Characteristics

Careful analysis of the means, standard deviations, counts, and percentages of the participants' responses revealed considerable information about the items within this SJT. Ideally, the distribution of responses would follow the same format for each item. That is, the four point answer would be identified as the "best answer" by participants more than any other response, the three point answer would be identified as "best" the second most frequently, the two point answer would be identified as "best" the third most, and the one point answer identified as the "best" answer the fewest number of times. This would show the overall participant response pattern agreed with how the SMEs weighted the four response options when creating the instrument. There were some items, which showed the expected distribution of responses. For example, in Scenario 3, Question 1 showed the ideal distribution of responses. The least correct answer was identified as "best" by 18 participants, the third best answer was identified as "best" by 30 participants, the second best answer was identified as "best" 58 times, and the best answer was identified as "best" more than any other, 76 times. Question Four of Scenario Three also showed this same general distribution. In addition to information about participant/SME agreement as to the "best" answer choices, analysis of participant response patterns revealed information about item difficulty. For example, items such as Question 1 in Scenario 1 and Question 1 in Scenario 5 showed disproportionately high numbers of participants choosing the four-point answer. In Scenario 1 Question 1, 127 participants (69.8%) chose the best solution, and in Scenario 5 Question 1, 125 participants (68.7%) chose the designated "best" answer. In both of these instances, the high numbers of participants selecting the "best" answer

suggests these items could be too easy, and the "best" answers could be too obvious. On the other hand, certain items showed response patterns where too few participants selected the answer the SMEs designated as "best." For example, in Scenario 5, Question 3, 113 participants (62.1%) chose the second best answer as "best", whereas only 34 (18.7%) selected the response the SMEs identified as "best." Similarly, in Scenario 3 Question 2, 96 participants (52.7%) chose the answer designated as the three point answer by the SMEs, whereas only 27 participants (14.8%) chose the answer designated the four point answer. Results such as these could mean either the item was too difficult or that in fact, at least according to the preponderance of participants, another solution is a better solution than the one the SMEs determined to be best.

Interestingly, in Scenario 1 on Question 2, only 45 participants (24.7%) chose the SME-designated four point answer, whereas 70 participants (38.5%) chose the answer designated as the worst of all possible answers (the one point answer) by SMEs. Similarly, in Scenario 4, Question 1, the distribution of participant responses was nowhere close to what the expected distribution would be. 36 participants (19.8%) chose the worst answer, 48 (26.4%) chose the third best answer, 87 (47.8%) chose the second best answer, and only 11 (6.0%) chose the supposed "best" answer. In instances such as these, reexamination of the scenario, question, and possible answer responses for possible revisions would be the best course of action.

Review of Subscale and Scenario Characteristics

As mentioned in the Results and shown in Tables 3, 4 and 5, the Cronbach alpha internal consistency reliabilities were extremely poor. The largest alpha was .47 for the overall 33-item average. From this, no support from alpha reliability measures was found for the assembly of these items into the four subscales or the two overall scales. Thus, assembly of these items into the theoretical subscales and scales was not supported.

The finding that the SJT was unable to measure specific subscales is not surprising. The majority of the existing SJT literature has similarly been unable to measure specific individual constructs (Schmitt & Chan, 2006). As Schmitt and Chan (2006) note, this may be due to the fact that when constructing SJT items, authors typically assume that each item loads on one and only one factor, when in fact it is most likely the case that any given item likely taps multiple factors. Also similar to a majority of the SJT research is the fact that while the instrument was unable to assess specific subscales, this instrument as a whole showed a slightly improved reliability figure with regard to its ability to measure some overarching construct seemingly related to an individual's overall judgment in a retail sales setting. To be clear, other SJT measures have in fact achieved much higher overall reliability figures; for example, the McDaniel et al. (2001) meta-analysis included SJTs, which reported alpha internal consistency values up to .94.

In spite of the poor psychometric performance this measure, as constructed, seems to assess knowledge of retail sales situations. In this way, the items in the SJT appear to have some level of face validity. There would still be value in assessing this one construct that could hypothetically be called "Retail Sales Situational Judgment." This initial iteration of the instrument does appear to be a good starting point for the development of a tool for selecting mid-level positions in retail sales.

Despite the apparent face validity of the instrument, the poor alpha internal consistency reliability figures lead to the conclusion that, in its current state, this instrument could not be used to validly assess any of the four subscales or Retail Sales Situational Judgment in general. Further refinement of the instrument (i.e., removing items which hurt reliability, re-wording items or potentially adding certain additional items) to improve reliability could yield a meaningful selection instrument.

Review of Subscale and Scenario Intercorrelations

Based on the reliabilities observed in Table 3, there is not sufficient consistency of the items within a subscale (these alphas range from .11 to .30) to suggest these groupings of items meaningfully hang together as a single unit. This also brings into question whether they assess the construct they are named for and purportedly measure. This leads one to question the merit of exploring the meaning of the correlations between the subscales, as the construction of these subscales is suspect.

Nonetheless, certain subscales show small, yet significant correlations, and the causes for these relationships should be explored. Although these items may not be actually assessing the subscales they were purported to measure, each item within the scale still assesses (at least to some degree) an individual's level of situational judgment within a retail setting. This fact is evident in the .47 correlation found for the overall 33-item reliability figure. Another possible explanation could be method bias. The questions are all posed in the same manner: after reading a paragraph describing a situation in a retail sales setting, the participants are asked a series of questions about the situation they have just read. Because the content is along the same lines on each of the scenarios, significant correlations are likely to be found amongst groupings of questions. Finally it is possible, and in the opinion of the author, quite likely, that these subscales and the items within them are all tapping, at least to some extent, cognitive ability, personality, job experience, or some combination of these three variables, which could also be an explanation for the significant correlations found between subscales.

As noted, it was observed that those answer choices which involved courses of action that were more collaborative in nature were often given higher scoring weights than those answer choices which involved acting without taking collaborative steps. This pattern may have

emerged in the instrument because it is reflective of the value system of the item writers themselves and the perceived importance of working collaboratively by professionals in the retail sales industry.

Limitations

There were numerous limitations inhibiting the success of this study. These limitations can be roughly grouped into two categories: issues with the sample and issues with the scale itself. First, regarding the sample used for analysis, one limitation was the fact that the sample was a somewhat unusual sample in that it consisted entirely of minorities. Furthermore, specific demographic information for each participant was not was not collected. Obviously, an entirely minority sample of college undergraduates is also unrepresentative of the general population of mid-level retail sales employees, which further limits the study. In addition, the relatively small sample size of 182 is less than ideal for the various statistical procedures conducted.

As mentioned, there were scale characteristics, which limited the study as well. For one, certain psychometric properties of the measure limited the statistical analysis and the utility of that analysis. For example, based on the factor analysis, the subscale factors did not group together; therefore, the meaningfulness of the correlations between subscales is suspect and must be viewed with caution. In addition, there were an unequal number of items per subscale. For example, there were three times as many items attempting to tap Leadership than for any other subscale, which suggests a greater emphasis on this facet relative to those that are attempted to be measured by the other subscales. Another limitation is that the scale was developed from a theoretical perspective without any empirical validation during development. In addition, the response options were also given theoretically assigned weights without any empirical validation as well. There was no testing of the order weights assigned to the answer choices to determine if

tests groups selected the various answer choices in distributions that were in agreement with the SME weightings of the answers.

Suggestions for Future Research

Various steps can be taken in the future to improve upon the instrument and study. First, a larger and more representative sample should be used and specific participant demographic information should be collected in order to determine what, if any, adverse impact exists within the instrument. In addition, future research should also obtain participant data such as job performance information for criterion-related validation of the instrument. If possible, it would be beneficial to simultaneously give participants personality and cognitive ability scales in order to explore the relationship between scores on the SJT and these key individual difference variables. In addition, it would be useful to further develop the instrument by empirically evaluating and refining the items, answer choices, and the weights assigned to each answer choice. Special attention should be paid to those items where there is a discrepancy between SME item choice weights and participant response distributions. Finally, it would be beneficial to have even numbers of items per subscale.

Table 1

Matrix of Questions Assigned by Competency

Item	Scenario Title	Problem Solving	Planning	Priority Setting	Leadership
1-1	Leading Team			X	
1-2				X	
1-3					X
1-4		X			
1-5					X
2-1	Coach				X
2-2					X
2-3					X
2-4					X
2-5					X
2-6			X		
3-1	I. Performance				X
3-2					X
3-3					X
3-4					X
3-5					X
4-1	Peer Leader			X	
4-2					X
4-3		X			
4-4			X		
4-5			X		
4-6			X		
5-1	BusinessSavvy			X	
5-2					X
5-3					X
5-4			X		
5-5			X		
5-6					X
6-1	Team Member			X	
6-2		X			
6-3		X			
6-4		X			
6-5			X		

Table 2
Counts and Percentages for Rankings of Scores for Scenario 1

	Counts	%	Mean	SD
Scenario 1 – Leading a Team				
Question 1			3.53	0.80
1. (response a)	4	2.2		
2. (response c)	23	12.6		
3. (response b)	28	15.4		
4. (response d)	127	69.8		
Question 2 –			2.37	1.23
1. (response b)	45	24.7		
2. (response d)	70	38.5		
3. (response c)	48	26.4		
4. (response a)	19	10.4		
Question – 3			3.18	0.83
1. (response c)	34	18.7		
2. (response a)	77	42.3		
3. (response d)	5	2.7		
4. (response 4)	66	36.3		
Question – 4			3.06	1.25
1. (response b)	13	7.1		
2. (response c)	38	20.9		

(*table continues*)

Table 2 (*continued*).

3. (response a)	22	12.1		
4. (response d)	109	59.9		
Question – 5			3.35	1.02
1. (response d)	8	4.4		
2. (response c)	125	68.7		
3. (response a)	37	20.3		
4. (response b)	12	6.6		

Note: All questions and weighted responses are not shown for test security. Responses are in numeric weight order.

Table 3
Counts and Percentages for Rankings of Scores for Scenario 2

	Counts	%	Mean	SD
Scenario 2 – Coaching				
Question 1 –			2.99	.91
1. (response d)	72	39.6		
2. (response c)	38	20.9		
3. (response b)	70	38.5		
4. (response a)	2	1.1		
Question 2 –			3.25	1.13
1. (response a)	28	15.4		
2. (response b)	14	7.7		
3. (response c)	24	13.2		
4. (response d)	116	63.7		
Question 3 –			2.82	1.33
1. (response c)	12	6.6		
2. (response d)	95	52.2		
3. (response a)	52	28.6		
4. (response b)	23	12.6		
Question 4 –			2.84	.97
1. (response a)	21	11.5		
2. (response b)	38	20.9		

(*table continues*)

Table 3 (*continued*).

3. (response d)	51	28.0		
4. (response c)	72	39.6		
Question 5 – How do you best use sherry as example			3.59	.62
1. (response d)	119	65.4		
2. (response b)	7	3.8		
3. (response c)	54	29.7		
4. (response a)	2	1.1		
Question 6 –			3.00	1.09
1. (response b)	38	20.9		
2. (response a)	23	12.6		
3. (response c)	37	20.3		
4. (response d)	84	46.2		

Note: All questions and weighted responses are not shown for test security. Responses are in numeric weight order.

Table 4
Counts and Percentages for Rankings of Scores for Scenario 3

	Counts	%	Mean	SD
Scenario 3 – Improving Performance				
Question 1 –			3.05	1.00
1. (response d)	76	41.8		
2. (response c)	58	31.9		
3. (response b)	30	16.5		
4. (response a)	18	9.9		
Question 2 –			2.65	.93
1. (response d)	96	52.7		
2. (response b)	28	15.4		
3. (response a)	27	14.8		
4. (response c)	31	17.0		
Question 3 –			2.36	1.16
1. (response a)	68	37.4		
2. (response c)	32	17.6		
3. (response d)	12	6.6		
4. (response b)	70	38.5		
Question 4 –			3.03	.92
1. (response c)	61	33.5		
2. (response b)	41	22.5		

(*table continues*)

Table 4 (*continued*).

3. (response a)	11	6.0		
4. (response d)	69	37.9		
Question 5 –			2.97	.89
1. (response a)	18	9.9		
2. (response b)	20	11.0		
3. (response c)	93	51.1		
4. (response d)	51	28.0		

Note: All questions and weighted responses are not shown for test security. Responses are in numeric weight order.

Table 5
Counts and Percentages for Rankings of Scores for Scenario 4

	Counts	%	Mean	SD
Scenario 4 – Peer Leadership				
Question 1 –			2.40	.87
1. (response a)	36	19.8		
2. (response c)	11	6.0		
3. (response d)	48	26.4		
4. (response b)	87	47.8		
Question 2 –			3.43	.84
1. (response d)	110	60.4		
2. (response b)	14	7.7		
3. (response c)	49	26.9		
4. (response a)	9	4.9		
Questions 3 –			3.24	.98
1. (response d)	44	24.2		
2. (response c)	99	54.4		
3. (response a)	23	12.6		
4. (response b)	16	8.8		
Question 4 –			3.35	.92
1. (response d)	48	26.4		
2. (response c)	106	58.2		

(*table continues*)

Table 5 (*continued*).

3. (response a)	14	7.7		
4. (response b)	14	7.7		
Question 5 –			3.22	.87
1. (response d)	79	43.4		
2. (response b)	11	6.0		
3. (response c)	78	42.9		
4. (response a)	14	7.7		
Question 6 –			2.79	.98
1. (response c)	68	37.4		
2. (response a)	42	23.1		
3. (response b)	14	7.7		
4. (response d)	58	31.9		

Note: All questions and weighted responses are not shown for test security. Responses are in numeric weight order.

Table 6
Counts and Percentages for Rankings of Scores for Scenario 5

	Counts	%	Mean	SD
Scenario 5 – Business Savvy				
Question 1 –			3.18	.75
1. (response d)	19	10.4		
2. (response a)	63	34.6		
3. (response c)	94	51.6		
4. (response b)	6	3.3		
Question 2 –			2.91	.82
1. (response b)	75	41.2		
2. (response c)	6	3.3		
3. (response d)	9	4.9		
4. (response a)	92	50.5		
Question 3 –			2.98	.64
1. (response d)	33	18.1		
2. (response a)	34	18.7		
3. (response c)	113	62.1		
4. (response d)	2	1.1		
Question 4 –			2.90	1.10
1. (response a)	32	17.6		
2. (response d)	69	37.9		

(*table continued*)

Table 6 (*continued*).

3. (response c)	58	31.9		
4. (response b)	23	12.6		
Question 5 –			3.54	.78
1. (response b)	124	68.1		
2. (response c)	7	3.8		
3. (response d)	11	6.0		
4. (response a)	40	22.0		
Question 6 –			3.31	1.01
1. (response d)	24	13.2		
2. (response b)	12	6.6		
3. (response a)	120	65.9		
4. (response c)	26	14.3		

Note: All questions and weighted responses are not shown for test security. Responses are in numeric weight order.

Table 7
Counts and Percentages for Rankings of Scores for Scenario 6

	Counts	%	Mean	SD
Scenario 6 – Team Member				
Question 1 –			2.90	.97
1. (response a)	23	12.6		
2. (response b)	26	14.3		
3. (response c)	80	44.0		
4. (response d)	53	29.1		
Question 2 –			2.99	1.16
1. (response d)	36	19.8		
2. (response c)	89	48.9		
3. (response a)	24	13.2		
4. (response b)	33	18.1		
Question 3 –			3.10	1.01
1. (response d)	17	9.3		
2. (response a)	63	34.6		
3. (response b)	80	44.0		
4. (response c)	22	12.1		
Question 4 –			2.97	.93
1. (response a)	11	6.0		
2. (response c)	64	35.2		

(*table continues*)

Table 7 (*continued*).

3. (response d)	47	25.8		
4. (response b)	60	33.0		
Question 5 –			3.32	.95
1. (response d)	115	63.2		
2. (response c)	14	7.7		
3. (response b)	49	26.9		
4. (response a)	4	2.2		

Note: All questions and weighted responses are not shown for test security. Responses are in numeric weight order.

Table 8

Means Standard Deviations and Cronbach's α of the Four Subscales

SJT Subscale	*M*	*SD*	Cronbach's α
1. Problem Solving	3.07	.50	.11
2. Planning	3.16	.41	.26
3. Priority Setting	2.87	.45	.16
4. Leadership	3.07	.28	.30

Table 9

Means, Standard Deviations and Cronbach's α of the Six SJT Scenarios

SJT Scenario	*M*	*SD*	Cronbach's α
1. Leading a Team	3.10	.50	.17
2. Coaching	3.08	.43	.05
3. Improving Performance	2.82	.44	-.02
4. Peer Leadership	3.07	.40	.15
5. Business Savvy	3.20	.41	.29
6. Team Member	3.06	.49	.18

Table 10

Means, Standard Deviations and Cronbach's α of Overall Scales of SJT

Overall Scales	*M*	*SD*	Cronbach's *α*
Four Scale Average	3.05	.25	.40
33-Item Average	3.06	.23	.47

Table 11

Pearson Correlation of Four SJT Subscales

SJT Subscale	1. Prob Solv	2. Planning	3. Prior Set	4. Lead
1. Problem Solving	-			
2. Planning	.15*	-		
3. Priority Setting	.18*	.10	-	
4. Leadership	.16*	.24***	.09	-

Note: $N = 182$

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 12

Pearson Correlation of 6 Scenarios

Scenario	1	2	3	4	5	6
1. Leading a Team	-					
2. Coaching	.09	-				
3. Improving Performance	.03	.25***	-			
4. Peer Leadership	.11	.14	.12	-		
5. Business Savvy	.06	.11	.05	.22**	-	
6. Team Member	.09	.05	-.02	.33***	.25***	-

Note: $N = 182$

*$p < .05$. **$p < .01$. ***$p < .001$.

REFERENCES

Chan, D. & Schmitt, N. (1997). Video-based versus paper-and-pencil method of Assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143-159.

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

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

Costa, P.T. & McCrae, R.R. (1992). *Manual for the revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO FFI).* Odessa, FL: Psychological Assessment Resources

Gatewood, R.D. & Field H.S. (2001). *Human resource selection.* Orlando, FL: Harcourt, Inc.

Konradt, U., Hertel, G., & Joder, K. (2003). Web-based assessment of call center agents: Development and validation of a computerized instrument. *International Journal of Selection and Assessment, 11*, 184-193.

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

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.

McDaniel, M.A. & Whetzel, D.L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33,* 515-525.

Motowidlo, S.J., Dunnette, M.D., & Carter, G.W. (1990). An alternative selection procedure: The low fidelity simulation. *Journal of Applied Psychology, 75*, 640-647.

Pulakos, E.D. & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performace, 9* 241-258.

Schmitt, N. & Chan, D. (2006). Situational judgment tests: Method or construct? In J. Weekley & R. Ployhart (Eds.), *Situational judgment tests: Theory, Measurement and Application*. (pp. 135-155). Laurence Earlbaum Associates, Mahwah, NJ.

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

Stemler, S.E. & Sternberg, R.J. (2006). Using situational judgment tests to measure practical intelligence. In J. Weekley and R. Ployhart (Eds.) *Situational judgment tests: theory, measurement and application*. Laurence Earlbaum Associates, Mahwah, NJ.

Sternberg, R.J. & Hedlund, J. (2002). Practical intelligence, *g*, and work psychology. *Human Performance, 15*, 143-160.

Sternberg, R.J., Wagner, R. K., Williams, W.M, & Horvath, J.A. (1995). Testing common sense. *American Psychologist, 50*, 912-927.

Stevens M.J. & Campion M.A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management, 25,* 207-228.

Wagner, R. K. & Sternberg, R.J. (1985) Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49,* 436-458.

Wagner, R. K. & Sternberg, R.J. (1991). *Tacit Knowledge Inventory for Managers: Users manual*. San Antonio, TX: Psychological Corporation

Weekley, J.A. & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*, 679-700.

Weekley, J.A. & Ployhart, R.E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance, 18,* 81-104.