Abstract

KROKOS, KELLEY JOAN. Situational Judgment: An Investigation of Its Process and Relationship to Scholar Performance. (Under the direction of Mark A. Wilson.)

Considerable disagreement exists regarding the nature of situational judgment and its relationship to performance. The purpose of this research is to address this disagreement. First, this research addresses lack of agreement regarding the nature of situational judgment by proposing that research to date has focused inappropriately on the final test score. More specifically, this research proposes that situational judgment can be shown to be a function of various cognitive processing tasks. A situational judgment inventory with embedded cognitive processing questions was developed to address these issues. The data do not support the models as proposed. After significant modification, situation awareness was the only cognitive processing variable to show promise as a predictor of situational judgment scores. Likely reasons include inappropriate operationalization of the factors.

This research also examines the relationship of situational judgment to performance in a group of university scholarship recipients. Situational judgment was proposed to be a partial mediator between accepted performance predictors and three performance criteria. The data do not support the model as hypothesized. After significant modification, the situational judgment scores were still not predictive of performance. Likely reasons for the lack of predictive validity include the nature of situational judgment, the nature of the sample, and methodological weaknesses. Implications for future research are discussed.

SITUATIONAL JUDGMENT:

AN INVESTIGATION OF ITS PROCESS AND RELATIONSHIP TO SCHOLAR PERFORMANCE

by

KELLEY JOAN KROKOS

A dissertation submitted to the graduate faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

PSYCHOLOGY

Raleigh

2003

APPROVED BY:

_____ _____

_____ _____

Chair of Advisory Committee

DEDICATION

This dissertation is dedicated to my Father, Stan and to my Mother, Jo. From my Father I learned to set high standards and how to set specific goals to achieve them. A cornerstone of his philosophy is having a positive mental attitude, which I have found to be immeasurably helpful in virtually every area of my life, particularly during the more challenging days of my graduate school education.

From my Mother I learned strength and compassion. More importantly, I learned that real strength walks hand in hand with compassion. I also learned about courage from my Mother, both through example and through words. She has always been able to find that delicate balance between loving me enough to worry about me as I pursued high adventures of the athletic, traveling, sailing, and the intellectual kind, and simultaneously giving me the courage to pull away from the dock (both literally and figuratively).

Thank you both for instilling in me the intrinsic and permanent value of education. Thank you for your words of encouragement and for showing me that intellectual curiosity is a life-long adventure. Thank you for the laughter, which was much and often. Finally, thank you for being you. It is clear that whatever I have achieved is because you loved me.

BIOGRAPHY

KROKOS, KELLEY JOAN. Kelley was born in 1963 in Columbia, SC, the daughter of Stan and Jo Krokos. She was reared in Spartanburg with her two sisters, Kimberley and Kristy. She graduated with honors from Spartanburg High School in 1981 and went on to complete her Bachelor of Arts degree in Psychology at Furman University in Greenville, South Carolina in 1985. After many years of coastal and offshore sailing, she began work in 1994 with the preemployment research section of the North Carolina Department of Correction. In order to meld her interests in performance in extreme environments with personnel selection and job performance, she entered the Industrial/Organizational Psychology program at North Carolina State University in 1996. This dissertation marks the completion of the Ph.D. program and the continuation of her personal and professional objective to understand and improve job performance in extreme environments.

ACKNOWLEDGEMENTS

Many have made significant contributions to this research. Dr. Bob Pond, committee member and Principal Investigator of the Scholarship Review and Development Team spent many hours reviewing the SJT and providing creative suggestions and support. Team member April Cantwell also spent considerable time reviewing the SJT and providing needed constructive criticism. This research and my skills are far improved for her influence. Torrey Rieser graciously assisted with data entry. Thanks are also due to the Scholarship Program Director, staff, advisory committee, and the project technical advisory board for their ongoing support of the selection review process. Dr. Wally Borman was particularly helpful.

Thanks are also due to the remaining Ph.D. committee members, without whom I would not have had the fortitude or skill to tackle a project of this magnitude. Dr. Katherine Klein has been a mentor and friend from my earliest days at North Carolina State, always sending opportunities for experience and growth my way. Dr. Donald Drewes sets a high standard of performance, but is consistent and patient in his support; the hours he spent answering questions and discussing ideas are proof that his interest in the development of the individual student is genuine. Finally, I would like to thank my committee Chair, Dr. Mark Wilson for accepting me as a student and for setting such an exemplary example. His ability to hone in on the crux of any matter and his ability to adeptly manage both academic and applied science is greatly admired. Finally, I thank Dr. Wilson for being supportive of my desire to develop a selection instrument for this dissertation. His confidence in me sends me forth well prepared.

Thank you all. My years in graduate school have been enriched by each of you.

TABLE OF CONTENTS

List of Tables

Page

List of Figures

Page

Situational Judgment:
An Investigation of Its Process and Relationship to Scholar Performance

Introduction

Situational judgment tests that measure judgment in job-related situations are becoming increasingly popular (Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; Hanson & Ramos, 1996; McDaniel, Finnegan, Morgeson, Campion, & Braverman, 2001). This surge of interest is likely due to reported successes in their ability to predict performance (McDaniel et al., 2001; Phillips, 1993; Weekley & Jones, 1999). A recent meta-analysis conducted by McDaniel et al. (2001) found an estimated population validity of .34. In addition to demonstrating criterion related validity, situational judgment tests (SJTs) have been shown to have smaller mean subgroup differences than tests of general cognitive ability (Hanson, 1994; Motowidlo & Tippins, 1993; Weekley & Jones, 1997). However, despite accumulating evidence that SJTs "work," there is significant variability in the size of the validities. Criterion related validities from individual studies range from no significant correlation (Smiderle, Perry, & Cronshaw, 1994) to as high as .56 (Stevens & Campion, 1999).

Furthermore, there is considerable disagreement regarding what SJTs are measuring, or more fundamentally, what situational judgment is and how or why it is predictive of performance (McDaniel & Nguyen, 2001; Weekley & Jones, 1999). Some situational judgment researchers suggest that SJTs are measuring a new construct. To support this proposition, they provide results that suggest that SJT scores explain variance in the criterion over and above what is accounted for by general cognitive ability or experience (Clevenger et al., 2001; Stevens & Campion, 1999; Weekley & Jones,

1997). Researchers in other areas of study have also proposed the existence of a new construct related to performance in complex situations. For example, Sternberg, Wagner, Williams, and Horvath (1995) propose the existence of a new construct, tacit knowledge that in contrast to cognitive ability is a sort of practical intelligence that allows one to successfully negotiate everyday situations. Similarly, Dulewicz and Higgs (2000) propose that emotional or social intelligence (EQ) is a non-cognitive skill necessary for the successful negotiation of complex social situations. Like situational judgment, both tacit knowledge and EQ are proposed to explain an additional amount of variance in organizationally important outcomes over and above that accounted for by *g* alone. Although the terminology differs somewhat, these constructs are routinely discussed as related constructs or otherwise included in the situational judgment literature and research (see Robins, 1994; Legree, 1995; McDaniel et al., 2001; Stevens & Campion, 1999). Consequently, this dissertation considers the tacit knowledge and EQ constructs as part of the situational judgment domain.

Not all researchers propose that SJTs are measuring a new construct. Strong observed correlations between SJT scores and scores on cognitive ability tests (McDaniel, Finnegan, Morgeson, Campion, & Braverman, 1997, 2001) have been cited as evidence that SJTs measure little more than general cognitive ability. Job knowledge and experience have also been cited as possible mechanisms underlying SJT scores (Schmidt & Hunter, 1993; Schmidt, 1994). Preliminary research suggests that SJT scores are related to three of the Big Five facets of personality: need for stability, conscientiousness, and agreeableness (McDaniel & Nguyen, 2001). Finally, others have taken a methodological approach and have suggested that SJTs are a measurement

method and hence, that an SJT could be developed to tap any construct of interest (Hanson, Horgen, & Borman, 1998; McDaniel & Nguyen, 2001).

Summary of the Problem and Purpose of the Current Research

In sum, SJTs have been found to be predictive of job performance, but the validities range greatly. In addition, SJTs have resulted in smaller mean subgroup differences than typical *g* tests, but again, the differences vary. This variability in criterion related validity and mean subgroup differences is likely due at least to some extent to the considerable lack of agreement among researchers regarding what situational judgment is and hence what SJTs measure. Furthermore, this lack of agreement regarding the nature of situational judgment is likely the result of an inappropriately narrow focus on a test taker's final SJT score and the concomitant acceptance of situational judgment as a unitary construct. This dissertation hypothesizes that answering an SJT item is a cognitive information processing task that involves accuracy at several levels of processing, including perception and comprehension of situational stimuli, comparison of the information to currently held knowledge about such situations, generation of consequences and meaning of alternatives, and finally, actual decision making. As such, situational judgment is likely a combination of several skills and abilities. Consequently, an understanding of the nature of situational judgment or of what SJTs measure necessarily requires an investigation of the various steps in the process. Such an understanding is a professional necessity if we are to improve upon the predictive validity of SJTs (Ployhart & Ryan, 2000), as well as an ethical and legal necessity if we are to defend the use of such tests (see Standards for Educational and Psychological Testing, 1999).

In order to shed light on the cognitive process of situational judgment, this dissertation will review both the situational judgment and the cognitive information processing literature. First, research regarding the predictive validity of SJTs will be reviewed, highlighting both the successes and failures of SJTs in predicting job performance. A portion of this discussion will be devoted to an examination of the observed differences in subgroup SJT scores. Possible reasons for the variability in the validity coefficients and the subgroup differences will be proposed. Second, the literature regarding the underlying nature of situational judgment will be reviewed, highlighting the significant differences in opinion among various researchers. It will be proposed that answering an SJT item is a cognitive information processing task composed of four processing stages. Third, the cognitive information processing literature will be reviewed, and based on these findings, a cognitive process model of situational judgment will be proposed and research will be proposed to test the model.

Job performance is arguably the raison d'etre of IO psychology; indeed the primary interest in SJTs has been for their role as preemployment selection tools. Consequently, in addition to the need to understand what situational judgment is, there is also a need to understand its relationship to job performance. It is the proposition of this dissertation that situational judgment is a partial mediator between established job performance predictors such as general cognitive ability, experience, and personality and various facets of job performance. This dissertation will review the job performance modeling literature and will propose an integrated model of performance that highlights the role of situational judgment. Research will be proposed to test the integrated model via a concurrent validity study using the most robust statistical analysis tools available.

Literature Review

The Criterion Related Validity of SJTs

A review of the situational judgment literature suggests that SJT scores are related to job performance. A meta-analysis conducted by McDaniel et al. (2001) of 39 SJTs resulted in 102 correlation coefficients from 10,640 participants. The authors report an estimated population validity of .34 (*SD*= .14 and 45% of the variance in the observed distribution due to artifacts). This is likely a conservative estimate of the mean validity as no corrections were made for restriction of range. Note that publicly marketed, government-owned, and tests developed for use by individual firms were included in the analysis, many of which have never been published in a peer-reviewed journal. In addition, readers are cautioned due to small sample sizes for some of the studies included in the meta-analysis.

Despite the encouraging nature of these summary results, criterion related validities from individual studies vary substantially. Weekley and Jones (1997) examined the relationship between SJT scores and job performance of hourly employees. In Study 1, the authors developed a video format SJT, which was administered to a development sample of 684 hourly employees of a discount retailer and a cross validation sample of 787 newly hired hourly employees. The resulting validity coefficients were .33 for the developmental sample and .22 for the cross validation sample. When corrected for criterion unreliability, the cross-validated validity coefficient rose to .34.

In 1999, Weekley and Jones conducted two additional studies of hourly retail store workers and hotel operators. Although the primary hypotheses for these studies pertained to the relationship between cognitive ability and SJT scores, they did report

significant validity coefficients between performance on two paper-and-pencil SJTs and supervisory ratings of job performance. The coefficient between SJT scores and performance reported from Study One using 844 hourly retail workers was .23. The coefficient reported for Study Two using 1040 hotel operators was .16.

Similar results were found by Phillips (1992) who developed an SJT called the Sales Skills Inventory to predict the selling performance of telephone sales operators. A job analysis revealed five major performance dimensions: selling, other customer interaction, coworker interactions, general office work, and operating the computer, with selling being the most important and time consuming. The behaviors associated with the selling dimension were used to develop the Sales Skills Inventory that was administered to 236 currently employed service representatives. The performance dimensions were collapsed to develop three composite performance criteria that were measured via supervisor ratings: job duties, selling, and job overall. The observed relationship between the SJT and job duties was .24, with selling was .18, and with job overall was .16. However, the results by gender provided a much different result. None of the validities were significant for the 145 women in the sample. The validities for the 36 men were as follows: r = .36 for job duties, r = .44 for selling, and r = .44 for job overall. This raises serious concerns given that the sample of 236 was primarily female (61% female, 16% male, and 23% not reporting).

In 1993, Phillips conducted another concurrent validation study in a large telecommunications company to examine the relationship between scores on an SJT that was developed to predict the relationship between the negotiation skills of credit consultants for residential telephone service customers and job performance. A job

analysis revealed five main performance dimensions: negotiating, other collection duties, coworker interaction, general office work, and using the computer, with negotiating being rated by supervisors and incumbents as the most important and most time consuming. The behaviors identified as part of the negotiating dimension were used to develop an SJT called the Negotiation Skills Inventory. The sample consisted of 249 currently employed credit consultants, of which 179 or 78% were women and 70 or 28% were men. The performance dimensions were collapsed into three performance criteria and were measured via supervisor ratings: job duties, negotiation, and job overall. Note that the author controlled for job knowledge by giving each participant a booklet to read before taking the SJT that listed the primary job responsibilities. The author reported strong validity coefficients: $r = .45$ for job duties, $r = .41$ for negotiation skills, and $r = .43$ for job overall. Unlike the results from Phillips' previous study (1992), the results from this study were significant and similar in magnitude for both males and females. A racial comparison revealed that the SJT was highly predictive of performance for the 21 Blacks (validities for the three performance dimensions range from .61 to .69), although the SJT scores were not significantly predictive of performance for the 22 Hispanics. However, caution should be used when interpreting these results due to the small sample sizes.

Motowidlo, Dunnette, and Carter (1990) gathered critical incident data regarding problem solving, interpersonal, and communication skills from a group of incumbents and supervisors from seven telecommunications companies. The authors wrote 64 situations based on these critical incidents, which were then reviewed by incumbent managers who provided information on the best possible behavioral response. Response

alternatives were developed from this data, which were then reviewed by senior managers. The 58 items that survived the final processes were administered in paper-and-pencil format to approximately 117 externally hired managers and approximately 25 internally promoted managers. The number in each sample is approximate due to missing data. Performance criteria were ratings of interpersonal, problem solving, communication and overall effectiveness. For the externally hired incumbents, the validity coefficients for the SJT were: $r = .35$ for interpersonal effectiveness, $r = .28$ for problem solving, $r = .37$ for communication, and $r = .30$ for overall effectiveness. These validities varied somewhat by gender and racial group. However, the validities were both statistically and practically significant for all groups and all dimensions except for the interpersonal effectiveness dimension for Blacks, which could be the result of the small sample size of approximately 23. The SJT was predictive of only two of the four performance dimensions for the internally promoted managers but the coefficients were relatively large: $r = .44$ for problem solving effectiveness, and $r = .34$ for communication effectiveness.

Given the significant length of time required to negotiate the 58-item version of SJT just discussed, the authors empirically developed a shorter version by including only the items with the highest relationship with performance ratings. The validities for the externally hired incumbents rose to .44, .48, .43, and .43 for the interpersonal, problem solving, communication and overall dimensions respectively. The validities for internally hired incumbents rose only slightly ($r = .45$ for problem solving and $r = .34$ for communication). As with the longer version of the SJT, the validities for the shortened version did not reach significance for the interpersonal and overall performance

dimensions for this group.

Stevens and Campion (1999) developed a paper-and-pencil SJT to test previously identified interpersonal and self-management KSAs required in team settings. In Study One, the authors administered the teamwork SJT and a battery of other aptitude tests to 70 currently employed pulp mill workers who were applying for employment at the company's new plant. Performance was operationalized as the average rating of performance across 5 supervisors. The performance data were gathered for purposes of this research only and included three items that measured teamwork performance and two items that measured task work performance. Combining the items yielded an overall job performance score. The resulting correlation coefficients between SJT scores and the performance criteria were relatively high: $r = .44$ with teamwork performance, $r = .56$ with task work performance, and $r = .52$ with overall job performance.

Smiderle, Perry, and Cronshaw (1994) found less impressive results in their investigation of the reliability and validity of the Metro Seattle Video Test, a commercially marketed SJT used for the selection of transit operators. They found no significant correlation between scores on the SJT and the number of commendations received, or a performance composite composed of the number of passenger complaints, absences, and the number of preventable accidents. They did find a statistically significant negative relationship between scores on the SJT and the number of complaints originating from passengers or supervisors ($r = -.12$). However, this means the SJT explained only 1.4% of the variance in the number of complaints received.

Some researchers have found significant relationships between SJT scores and organizationally important but less traditional measures of job performance. For

example, Dalessio (1994) examined the ability of a video format SJT to predict the turnover of insurance agents. A video format SJT was developed and administered to 677 newly hired agents, who were divided into three samples. Dalessio (1994) divided the SJT test scores into quartiles and found that participants in the highest scoring quartile had a 19% increase in average survival rate at the one year tenure mark when compared to those participants in the lowest scoring quartile.

In sum, SJTs are predictive of a variety of organizationally important outcomes. However, the significant variability in the size of the observed validity coefficients suggests that some SJTs may be "better" than others. However, the reasons for these differences are unclear.

Subgroup differences on SJTs.

In addition to being predictive of performance, SJTs have been touted as a particularly useful selection tool in that they generally produce smaller subgroup differences than tests of general cognitive ability. Following is a review of the specifics surrounding SJT subgroup differences.

While tests of general cognitive ability are arguably the most valid predictors of job performance, mean scores among various subgroups generally differ by approximately one standard deviation (Hunter & Hunter, 1984). Consequently, these tests may exhibit significant levels of adverse impact against a variety of protected groups. Some of the enthusiasm regarding SJTs is because although Whites and females tend to score somewhat higher than their subgroup counterparts, the differences in mean scores are either not significant or much smaller than the typical one standard deviation difference found among subgroups on tests of general cognitive ability (Hanson &

Ramos, 1996).

For example, Hanson (1994) developed an SJT to test the supervisory effectiveness of Army personnel. The SJT was administered to 1049 second tour soldiers, of which 89% were male and 11% were female. The racial composition of the sample was 56% White, 33% Black, 6% Hispanic, and 5% other. Results showed a difference between mean scores for men and women of approximately one third of a standard deviation, with men scoring lower. Results also show a significant difference in mean SJT scores among racial groups, with Blacks scoring approximately one third of a standard deviation lower than Whites or Hispanics. While statistically significant, these differences are less than the one full standard deviation difference typical of cognitive ability tests.

Weekley and Jones (1999) hypothesized that the differences in mean SJT scores by race would be smaller than the mean differences typically found in cognitive ability test scores. Results from two studies of approximately 4000 participants supported this hypothesis. In Study One, there was no significant difference between the mean SJT scores for Whites and Hispanics. However, Whites scored significantly higher than Blacks (*ES* = .85). In Study Two, Whites scored significantly higher than Blacks on the SJT (*ES* = .52) and significantly higher than Hispanics (*ES* = .36). Although these differences are statistically significant, they are smaller than the difference of one standard deviation typically found in tests of cognitive ability. In addition, they are smaller than the observed racial subgroup effect sizes in cognitive ability, which ranged from .52 to .94. The authors suggest in the general discussion that minimizing mean subgroup differences is an important benefit of SJTs, but that it is unlikely that any SJT

could ever completely eliminate such differences.

Motowidlo and Tippins (1993) compared the SJT test scores of 403 applicants and found no significant difference between men and women, and a small but statistically significant 2.8-point difference in the mean score among racial groups with Blacks scoring lower than Whites. Similar results were found in Study 2, with no statistically significant difference in mean SJT scores found between men and women. However, the racial subgroups in Study 2 were too small to compare. In the general discussion, the authors summarize that when combined with the weighted sample results from an earlier study using a longer version of the same SJT (Motowidlo, Dunnette, & Carter, 1990), mean test score differences by subgroup average approximately one third of a standard deviation, which again, is significantly less than tests of general cognitive ability.

In sum, available evidence suggests that scores on SJTs are generally related to job performance, but some SJTs are more predictive than others. Furthermore, while SJTs generally show non-significant or small mean score differences by subgroup, some SJTs show greater subgroup differences than others. Possible reasons for the variability in the predictive validity and mean subgroup differences are explored below.

Why the Variability in Validity and Mean Subgroup Differences?

The majority of the evidence regarding SJTs suggests that they are predictive of performance and demonstrate smaller mean score differences among subgroups than tests of general cognitive ability. However, the reported criterion-related validities vary significantly as do the differences in mean subgroup scores. How might these differences be explained? Variability in the criterion related validity and mean subgroup scores of SJTs may be a function of differences in test development, format, research

methodology, research design, or methodological weaknesses. This section reviews what is known about each of these issues.

Differences in SJT Development. There are four primary ways SJT item stems can be developed: via training material, job analysis, a critical incident technique, or on the basis of an existing taxonomy (Hanson, 1994). Some research has been conducted to determine the impact of test development issues on the validity of SJTs. For example, McDaniel at al. (2001) examined the impact of job analysis on the criterion related validity of SJTs. Their meta-analysis compared 5959 scores from SJTs that were developed based on job analyses to 3251 scores from SJTs that were not based on job analysis. They found that SJTs developed on the basis of a job analysis had higher mean corrected population correlations with job performance ($\rho = .38$, $SD = .06$, and 73% of the variance in the observed distribution due to artifacts) than SJTs not based on a job analysis ($\rho = .29$, $SD = .16$, and 46% of the variance due to artifacts).

Another development issue that has been examined is the amount of detail in the situations presented. The recent meta-analysis by McDaniel et al. (2001) examined the detail level of SJT questions as a possible moderator of the relationship between SJT and job performance. The authors hypothesized that highly detailed SJT questions would require greater job knowledge and reading ability, which would in turn likely increase the correlation between the score on the SJT and job performance. The results did not support their hypothesis. Rather, the mean estimated population correlation for detailed questions was slightly smaller ($\rho = .33$ and $SD = .10$ for $N = 2218$ with 46% of the variance in the observed distribution due to artifacts) than for less detailed questions ($\rho = .35$ and $SD = .12$ for $N = 6747$ with 53% of the variance due to artifacts), suggesting that

less detailed situations on an SJT may produce higher correlations with job performance.

Differences in SJT Format. Researchers are beginning to examine and manipulate the format of SJTs to determine its impact on validity and mean subgroup differences. For example, in an attempt to isolate the effect of method of administration on mean subgroup differences, Chan and Schmitt (1997) developed two SJTs designed to be identical in content but that differed in administration method: one was administered in video format and the other was administered in paper-and-pencil format. The authors hypothesized that the video administration method would produce smaller subgroup differences. They further hypothesized that the reduction in subgroup differences would be due to the reduced reading requirement and increased face validity of the video administration method. The White and Black participants were randomly assigned to either the video or paper-and-pencil administration group, resulting in a 2 x 2 (Race x Method) research design. All participants were administered a reading test and a self-report measure of face validity. The results support the hypotheses. First, the authors provided support for the assumption that the two tests were measuring the same constructs by showing that the factor loadings and error variances are invariant across groups. Next, the authors examined the main effects of race and method and found that Whites scored higher than Blacks (d = -.61) and that the video-administration group scored higher than the paper-and-pencil administration group (d = -.42). Interaction effects were also found; the difference in test performance between Whites and Blacks was higher in the paper-and-pencil administration method (d = -.95) than in the video administration (d = -.21), thus supporting their hypothesis that the video format would produce smaller subgroup differences. An interaction was also found between reading

comprehension and method of administration, such that reading comprehension was correlated with test performance on the paper-and-pencil administration method, but not in the video administration method. This provides support for their hypothesis that the video SJT test scores are higher due to a reduced reading requirement. Finally, the authors examined the perceptions of face validity and found that participants who took the SJT via video administration reported higher face validity perceptions than the paper-and-pencil group. Furthermore, a race and method interaction for face validity occurred such that the difference between face validity reported by Blacks and Whites was smaller for the video administration (d = -.11) than for the paper-and-pencil administration (d = -.80). The authors summarize that video administration reduces adverse impact partly through reducing the reading required and partly by increasing motivation through the perception of higher face validity. These results support the work of others who have also suggested that video format may be preferable in test situations where the reading ability of the participants is low (Weekley & Jones, 1997). Unfortunately, the recent SJT meta-analysis by McDaniel et al. (2001) did not include video format SJTs and consequently, no summary evidence is available that estimates the population differences between video and paper-and-pencil administration.

SJTs generally present a situation or situations followed by a list of response alternatives to which the test-taker is instructed to respond. The wording of the instructions represents another format issue of interest. For example, SJT questions may require the test-taker to indicate which behavioral response is the "most effective" or what she or he would "most likely do." Regardless of how the question is worded or whether the situations are presented in video or paper-and-pencil format, the test-taker is

not required to demonstrate the behavior, but rather to simply choose from among the alternatives provided. Consequently, the researcher cannot know whether the test-taker would actually perform the behavioral response if faced with the dilemma. This "should do" versus "would do" discrepancy is an inherent limitation of any SJT that requires the test-taker to record his or her preferences as opposed to demonstrating those preferences.

Robins (1994) provides empirical evidence of this discrepancy by manipulating the question format on a paper-and-pencil SJT. Test-takers were presented with a series of work and life situations and asked to identify both what they thought they "should do" and what they "would do." The correlation between the "would" and "should" questions for both the work and life situations was .77. These results suggest that there may be a difference between what test-takers report is the best answer and what they would actually do if presented with the situation.

Researchers have approached this problem in various ways. Hanson (1994) acknowledges the discrepancy and proposes that the organization should be interested in what the test-taker would actually do. However, Hanson then proposes that asking test-takers what they "would do" may result in contamination of the responses; that is, respondents who are able and willing to be truthful will likely report what they would do, but respondents who are motivated by social desirability concerns would at least in some instances likely report what they thought they should do. In order to reduce this contamination and hence to standardize the responses, Hanson chose to ask participants what they thought they "should do" rather than what they "would do."

McDaniel and Nguyen (2001) also suggest that an SJT that asks a test-taker what she or he would most likely or least likely do can easily be "faked," in that the test-taker

would be likely to choose the more socially desirable response regardless of what action she or he would actually take if faced with the situation. They suggest that a more fake resilient technique is to ask the test-taker to identify the best and worst response to each item. Weekley and Jones (1999) also propose that finding out if the test-taker can identify what should be done is more appropriate.

However, other researchers have proposed that the issue of interest is what the participant would actually do as opposed to what they thought was appropriate. For example, Motowidlo, Dunnette, & Carter (1990) asked participants what alternative they would "most likely" and "least likely" do.

Despite the potentially serious nature of this discrepancy, Weekley and Jones (1999) suggest that it is "probably trivial" (p. 685). In addition, there are two relevant practical issues. First, this discrepancy is theoretically more likely in some work settings than others. For example, there is likely to be a discrepancy between what an applicant for a correctional officer position in a prison facility would choose on a paper-and-pencil test and what she or he would actually do if faced with a dangerous situation that carried some element of personal risk. Theoretically at least, it is much *less* likely that such a discrepancy would exist for jobs that require little or no physical danger or risk.

Second, it seems theoretically reasonable that asking a test-taker to identify the correct course of action from among several alternatives is assessing judgment or knowledge. However, if one is trying to assess whether the test-taker would perform well on a job (and hence provide criterion related validity evidence for the test as a selection tool for job performance) the appropriate approach would be to determine what the test-taker would actually do if faced with the dilemma. This supposition is supported

empirically by Robins (1994). Although Robins used a composite of several situational judgment measures (video, paper-and-pencil, "should" work questions, "would" work questions, "should" life questions, "would" life questions, a "would" composite, and a "should" composite) to test the main hypotheses regarding situational judgment, personality, and performance, evidence of the differential effects of the "would" composite and the "should" composite are provided in the correlation matrix. The validity coefficients between the "would" composite and several important performance criteria were higher than the coefficients for the "should" composite with those same criteria. Specifically, the correlation coefficients for the "would" composite were .23, .26, .16, and .14 with service-oriented sales performance, customer service performance, sales performance, and the number of client compliments. The coefficients between the "should" composite and these same criteria were smaller (.19, .22, .14, and .10 (*ns*) respectively).

In sum, the wording of the instructions regarding how a test-taker should respond to the situational dilemma(s) may have an impact on the SJTs predictive validity. However, it is the proposition of this dissertation that the choice of instructional wording should be determined by the purpose of the administration, the composition of the sample, and the setting. For example, if one is administering the SJT to validate the items, asking test-takers what they would "likely do" should maximize the relationship between the SJT score and performance, hence maximizing the validity coefficient. Furthermore, members of any sample for whom the scores will have no impact (incumbents or applicants if the SJT will have no impact on selection decisions) are much less likely to be tempted to "fake," again making it reasonable for the instructions to ask

what the test-taker would "likely do." Finally, for reasons previously discussed, differences in instruction may be significantly more pertinent in some work settings than others.

Another format condition that has been examined and manipulated is format of the response scale. The traditional response scale consists of a forced choice design that requires the test-taker to choose the most effective or the most "likely to do" response from a list of alternatives. A variant of this approach is to ask test-takers to choose both the most and least effective or "likely to do" alternatives. This approach reveals substantially more information about the test-taker and allows for a number of scoring methods. For example, Hanson and Borman (1993) identified a number of scores that could be derived by combining these two pieces of information with SME ratings of effectiveness for each alternative: the number of "most effective" hits, the number of "least effective" hits, the mean effectiveness rating per SMEs of the response the test-taker chose as most effective, the mean effectiveness rating per SMEs of the response the test-taker chose as least effective, and finally, a score that represents the difference between the mean effectiveness rating for the most and the least effective alternatives. The authors utilize this multiple scoring method in their development and examination of an SJT created to measure supervisory job knowledge of Non-Commissioned Army Officers. All five scores were calculated for the Army SJT. However, an examination of the item-total correlations for each of the scoring strategies yielded modest results. The authors subsequently opted to use the score associated with the difference scoring method because it had the highest median item-total correlation (r = .33). The authors used this score alone for all subsequent analyses on the Army SJT data.

Others have proposed that the most appropriate method is to present the test-taker with multiple response alternatives and ask him or her to rate each one on a Likert-type scale of effectiveness (Legree, 1994; Sternberg, Wagner, & Okagaki, 1993; Wagner, 1987). In an empirical investigation of this issue, Legree (1994) compared the reliability of the scores on the Army's supervisory ability SJT using the traditional forced choice method and an alternative Likert-type scale format. Test-takers in the forced choice condition were asked to choose the most and least effective alternative and these data were then used to compute five different scores according to the methods described above. Test-takers in the Likert type scale format condition were administered the same SJT but were asked to rate the appropriateness of each response alternative on a scale of 1 to 11. Legree hypothesized that scores from the Likert-type scale format would be more reliable than scores from the forced choice method because it yields more data points per test-taker. This hypothesis was supported; SJT scores using the alternative method had higher reliability (.62) than the most reliable of the five scores (the difference weighting score) from the forced choice response scale method (.51). An advantage of the Likert method is that the performance of individual response alternatives can be compared. Legree refined the SJT by eliminating 57 response alternatives that had a negative item-scale correlation. The reliability for the refined scale increased to .84. The Likert type scale format is particularly useful in that the increase in the number of data points (and the resulting increase in reliability) did not come at the expense of lengthening the test.

In a similar investigation of 400 U.S. Air Force recruits, Legree (1995) found somewhat smaller differences in reliability between the forced choice method and the Likert type scale format. The reported reliabilities were .76 for the difference weighting

score in the forced choice condition and .80 for the Likert-type scale condition.

Another important difference in SJTs relates to the strategy used to identify the "correct" answer, which is required in order to score an SJT. McDaniel and Nguyen (2001) report that there are three main scoring strategies. First, a researcher may elicit the assistance of subject matter experts (SMEs) who decide individually or in groups which response alternative is the most effective. Items with little or no SME agreement are deleted or rewritten. Second, an SJT may be pilot tested and the "correct" answers identified based on central tendency statistics. Finally, an empirical method may be employed to determine the "correct" answer. Research comparing the first technique (SME derived key) versus the third technique (empirically derived key) is provided by Weekley and Jones (1997, Study 2) who developed an SJT for entry-level caregivers in a nursing home setting. The development sample consisted of 412 employees in various entry-level caregiver positions and the cross validation sample consisted of 148 applicants. However, the authors created two different scoring keys for the SJT items. The development sample of current employees was used to develop an empirical scoring key. The authors averaged the job performance scores of all the subjects who choose a particular response option. The option that was chosen by the subjects with the highest mean performance was coded as +1 and the option chosen by the subjects with the lowest mean performance was coded as –1. Of these items, 41 performed appropriately for this scoring strategy. The remaining 8 items had one response alternative that was chosen by most of the subjects. These items were scored from 0 to 1. The final empirical SJT score was the sum of the scores from each of these 49 items.

The authors also derived a second scoring key using "customer" opinion. This

process involved asking the family members of the nursing home residents to choose the response they preferred in each of the 49 SJT video items. The option chosen most often by family members was coded as 1 and all other options were coded as 0. The rational SJT score was simply the sum of the scores of the 49 items. Measures of general cognitive ability, experience, and job performance were also collected from both the development and cross-validation samples.

The results suggest that the scoring keys are different. The correlation between the empirical and the rational scoring keys was .53 for the development sample and .48 for the cross-validation sample, suggesting that perspective may account for a great deal of variation in what the "correct" answer is. Despite the difference in scoring keys, both keys were related to job performance. The empirically derived scoring key correlated with performance at .35 in the development sample and .24 in the cross-validation sample. Corrected for criterion unreliability, the empirically derived key correlated with performance at .38 for the cross validation sample. Finally, the empirical key accounted for an additional 5.7% of the variance in performance for the cross-validation sample over and above that accounted for by cognitive ability and experience.

The validities using the customer-based rational scoring key were .14 for the development sample and .33 for the cross-validation sample. Like the empirical scoring key, the rational scoring key accounted for an incremental portion of the variance in performance (9.6%) for the cross-validation sample over that accounted for by cognitive ability and performance.

In addition to differences in validities, Whites scored higher than Blacks using both the empirical and the rational scoring keys. Furthermore, the empirical and

customer based keys were differentially related to race. The effect size for the difference in mean scores using the empirical key was .38 and the effect size for the rational scoring key was .52. The authors suggest that the race of the customers who provided the data for the rational key may be partially responsible for this effect, although this is not verifiable as the race information was not available for this group. Taken together, these results suggest that scoring strategy can have a significant impact on observed validities and on mean SJT score subgroup differences.

In sum, there are significant differences in how SJTs have been developed, administered, and scored. Research suggests that these differences may account for some of the observed range in validities and mean score differences among subgroups. However, there are several issues that remain unclear. First, the SJTs reviewed here generally present the test-taker with unrelated or independent items. However, it may be that items that build upon the previous item and that unfold in a temporally natural sequence may be more naturalistic and hence more valid predictors of performance. While some literature exists for the use of dependent items in written simulations used as criterion measures (see Hanson, 1994), no research was found that examined the impact of using dependent or related items in a selection SJT. It is unknown what impact this format would have on predictive or concurrent validity. Second, the difference between a "most likely to do" score and a "least likely to do" score is unknown. The present research will address these issues.

Differences in Research Methodology. In addition to examining the impact of how SJTs are developed, administered, and scored, one must also consider differences in methodology and research design. For example, there are significant differences in how

job performance has been measured across SJT studies. There are a number of ways to measure job performance that vary along a continuum of relatively objective to relatively subjective. Smith (1976) identifies several different types of objective performance criteria including absences, tardiness, accidents, tenure, promotions, salary, and production measures such as sales in dollars and the number of records entered by data entry personnel. Smith acknowledges that there is a subjective component to even these relatively objective criteria. However, when available and appropriate, these types of on-the-job measures can be a valid and fair method of measuring job performance. Despite the relative objectivity of these types of measures, only a few SJT researchers have examined the relationship between SJT scores and objective performance criteria. These studies have yielded mixed results. In the previously discussed investigation of the Metro Seattle Video Test, Smiderle, Perry, and Cronshaw (1994) found either no statistically significant or no practically significant relationship between SJT scores and complaints, commendations, tenure, or a summary score that included passenger complaints, absences, and accidents. They did find a statistically relationship with length of tenure (r = -.28), but the correlation was in the opposite direction than hypothesized. In another previously discussed study, Dalessio (1994) found that SJT scores were predictive of employee turnover, an organizationally important but less traditional measure of performance.

Performance during training may be used as a measure of job performance. To the extent that there is fidelity between training content, tests, and scored exercises, and the job requirements, training data can be a valid and useful measure of job performance. Some research suggests that a relationship exists between SJT scores and various

measures of training performance. For example, Krokos (1999) used structural equation modeling to examine the relationship between SJT scores and performance in a sample of correctional officers. She found modest structural path coefficients between SJT scores and various training criteria including basic training exam scores (.13) and basic training firearms scores (.45).

While objective on-task performance criteria are available as measures of performance for some jobs, many jobs have a significant cognitive component that does not easily lend itself to objective measurement. In this case, measures of job performance such as ratings or rankings may be required. In a 1994 review of personnel selection and placement, Landy, Shankster, and Kohler reported that supervisory ratings were the most common performance criteria. A perusal of recent SJT validation studies supports this notion. For example, previously discussed studies by Weekley and Jones (1997 and 1999), Stevens and Campion (Study One, 1999), and Phillips (1992 and 1993) used ratings or rankings from one or more supervisors as the only performance measurement method. Summary evidence of the ubiquity of supervisory ratings is provided in a meta-analysis by McDaniel et al. (1997) of 15,234 subjects across 95 studies that examined the relationship between SJT and job performance. Of those 95 studies, 88 employed supervisory ratings or rankings of job performance as the criteria. Limiting performance criteria to one measurement method such as ratings is problematic in that it makes it impossible to determine how much of the observed variability in performance is due to method variance (Campbell, 1990).

Ratings of performance from sources other than supervisors, such as self, peer, and subordinates may also be used. Smith (1976) suggests that ratings from peers are not

only valid, but may be more stable over time than supervisory ratings. Ratings from subordinates appear to have some validity, but they differ so markedly from supervisory ratings as to be measuring a different aspect of job performance. Self-ratings are prone to the obvious biases, but may be useful for creating a positive perception of the appraisal process. A more recent review by Landy, Shankster, and Kohler (1994) provides a mixed perspective. These authors cite evidence that suggests that race effects may be more likely in peer ratings than in supervisory ratings, and that peer ratings seem to be measuring something quite different than supervisory ratings. They conclude that peer ratings should not be considered true measures of performance, but instead as a construct or phenomena to be examined.

Stevens and Campion (1999) conducted the only SJT validation study found that utilized performance data from sources other than supervisors. In Study Two, the authors expanded the criteria used in their previously discussed Study One. They administered the teamwork SJT and a reduced battery of aptitude tests to a sample of 72 currently employed box plant workers, whose job responsibilities were similar to the pulp mill workers. As in Study One, supervisory ratings of performance were gathered for research purposes only and were averaged across multiple raters. However, a 360-degree approach of gathering performance ratings from multiple sources was used that included self-ratings and peer nominations. In addition, more items were rated: five teamwork items, three task work items, and one overall item were rated. Significant correlations were found between scores on the teamwork SJT and several of the supervisor ratings of performance: $r = .21$ with teamwork performance, $r = .25$ with task work performance, $r = .23$ with overall performance. Results of the 360-degree analyses were mixed.

Correlations between scores on the teamwork SJT and the 360 performance criteria were: $r = .23$ for peer nominations of teamwork and $r = .21$ with peer nominations of overall performance. However, the teamwork SJT was not related to any of the self-ratings or peer nominations of task work performance. These results suggest that a 360º approach may be appropriate, at least for some performance criteria.

Despite the prevalence of ratings as measures of job performance, many have criticized them as being the source of systematic unwanted variance. For example, characteristics of the organization, the position, the rater, the ratee, the reason for the rating, and the rating instrument itself have all been identified as possible reasons for the variability in ratings (Landy & Farr, 1980). In a recent empirical investigation, Scullen, Mount, and Goff (2000) examined performance ratings by supervisors, peer, subordinates, and self for 2142 managers. Using a modified correlated uniqueness-CFA technique to separate the trait factors (general performance and performance on specific dimensions), method factors (rater perspective and rater biases), and random error, they found that rater biases accounted for an average of 62% of the variance in performance ratings, rater perspective accounted for an average of 9% of ratings, random error accounted for 11%, and actual performance accounted for only 20% of the variance in ratings (11% for general performance and 8% for performance on specific dimensions). The finding that actual job performance accounts for so little of the variance in ratings for this sample is a concern. However, others are more optimistic. Landy, Shankster, and Kohler (1994) report that when ratings are conducted by trained raters using rating scales that have been developed appropriately, that they are "minimally biased" (p. 282). Others suggest that the reliability and validity of ratings can be improved by increasing

the number of raters (Campbell, 1990) and by training the raters (Woehr & Huffcutt, 1994; Landy & Farr, 1980). In sum, if conducted appropriately, ratings of performance from a variety of sources may be a useful and valid method of measuring performance.

Another difference in research methodology refers to whether the SJT was tested using a concurrent or predictive validity research design. Most of the validation research that has been conducted on SJTs has been conducted using concurrent designs. Of the 10,640 subjects included in a recent meta-analysis by McDaniel et al. (2001), 10,294 were part of concurrent validation studies and only 346 were part of predictive validation studies. The prevalence of concurrent designs is likely due to time required to test job applicants and then wait to gather corresponding performance data. In addition, organizations may be loath to devote the resources to develop and administer a preemployment test and then not utilize the results. However, to the extent that experience contributes to performance on an SJT, concurrent validation with current employees will likely result in somewhat different results than a predictive validity study using applicants. If the goal is to develop an SJT to be used as a preemployment selection test, then a predictive validity design would be more appropriate.

Other researchers have sought to reconcile the needs of science with the needs of the sponsoring organization by combining the two approaches; these researchers have developed an SJT and tested both its concurrent and predictive validity. For example, as previously discussed, Motowidlo, Dunnette, and Carter (1990) developed a full-length and an empirically shortened SJT and tested them via a concurrent validation design using current employees. Several years later, Motowidlo and Tippins (1993) examined the predictive validity of the shortened 30-item version by administering it to 403

applicants in a large telecommunications company over a two-year period. Of the applicants who were ultimately selected for employment, 36 had performance data and could be included in the study. Although the sample size is small, the results suggest that the SJT is predictive of communication effectiveness (r = .33) and of overall job performance (r = .31). The correlations between the SJT and leadership, interpersonal effectiveness, and problem solving effectiveness were not statistically significant but were in the expected direction.

The results of these two studies suggest that SJT criterion related validation studies may result in statistically and practically significant results, regardless of whether a concurrent or predictive design was employed. However, in an effort to estimate the impact of study design on the criterion related validity of SJTs in the population, McDaniel et al. (2001) compared the 346 SJT scores obtained in predictive design studies to the 10,294 SJT scores obtained in concurrent designs. Their results suggest that research design is a moderator of the relationship between SJT scores and job performance, with concurrent designs resulting in higher mean estimated population validities with job performance (ρ = .35, SD = .14, and 45% of the variance in the observed distribution due to artifacts) than predictive designs (ρ = .18, SD = .05, and 91% of the variance due to artifacts).

There is substantial efficiency in developing an SJT and immediately testing its concurrent validity while simultaneously gathering data for a future predictive validity study. This technique maximizes the benefit to the researcher, organization, employees, and the applicants. Given these benefits, this method will be used to examine the efficacy of the SJT developed in this study. However, only the concurrent validity results

will be reported here.

Methodological Weaknesses. In addition to differences in methodology and research design, a perusal of the SJT literature reveals some methodological weaknesses. For example, it has been noted by several researchers (McDaniel et al., 2001; Ree & Earles, 1993; Weekley & Jones, 1999) that previous situational research has used small samples. Sample or subgroup sizes examined in published studies included in this dissertation are as small as $n = 25$ (Motowidlo, Dunnette, & Carter, 1990) and $n = 36$ (Motowidlo & Tippins, 1993; Phillips, 1992).

In addition to small sample sizes, many SJTs are developed specifically for use in a particular organization (Weekley & Jones, 1999). Consequently, much of the literature available describing SJTs is in the form of conference papers or unpublished technical reports (see References from McDaniel & Nguyen, 2001). The actual test items are generally not available for examination. Of those SJTs that are commercially available, some have shown to have technical weaknesses. For example, as previously discussed, Smiderle, Perry, and Cronshaw (1994) found the Metro Seattle Video Test to be lacking in both reliability and validity.

Another methodological weakness in SJT validation studies is that statistical analyses have been conducted almost entirely at the measurement level. The usual index of criterion-related validity is the bivariate correlation between a single SJT score and a single measure of performance. Campbell (1990) has suggested that despite the ubiquity of this index, it is an inappropriate measure of validity because it oversimplifies both the predictor and criterion space. That is, there is more than one useful predictor of performance, and performance is not adequately conceptualized as a unitary construct. In

addition, this index of validity makes it difficult to separate measurement from structure. That is, it is impossible to ascertain whether a low validity coefficient is due to inadequate measurement of the constructs of interest, or to the inadequacy of the constructs to account for the observed relationships.

While measurement level analyses can be an important data analysis component, it is also appropriate to examine how well the constructs have been operationalized and the relationship between the proposed constructs. The statistical technique most suited to answer these questions is structural equation modeling (SEM), which provides indices of the efficacy with which the manifest measures are indeed measuring the latent construct they are proposed to measure, and of the proposed relationships between and among the latent constructs. Despite the superior nature of this technique, it has generally been untapped as a resource for examining the content and process of situational judgment. Various explanations have been proposed. Austin and Villanova (1992) suggest that analyses at the construct level are critically important, but recognize that the statistical techniques and computing power have only recently become available for such analyses. More recently, Borman, Hanson, and Hedge (1997) suggest that it is the increasing complexity of these techniques such as SEM that is preventing researchers from embracing them more fully.

In one of the very few published studies that utilize SEM to examine situational judgment and the relevant predictors and criteria, Borman, Hanson, Oppler, Pulakos, and White (1993) propose a full model that includes ability (as measured by the Armed Services Vocational Aptitude Battery) and experience (self-report) as predictors of job knowledge (SJT) and proficiency (mini-assessment center), which in turn influence

ratings of leading/supervising. In addition, the model includes a path from ability to experience, and a path from knowledge to proficiency. This five-construct model represents one of the most comprehensive situational judgment models to date in that it includes *g*, experience, SJT scores, a measure of job proficiency, and performance ratings. To determine the individual effects of various paths, the authors tested a series of models that are nested in the full model. Data were gathered from 570 second-tour soldiers who are considered to be "beginning supervisors" (p. 445) in that they have a mean of 26.46 months of experience as first-line supervisors. The results suggest that the full model with all five constructs was the best fitting model. The path coefficients led the authors to conclude that ability has a greater influence than experience on job knowledge (as measured by SJT scores), and that both ability and experience have an influence on proficiency (as measured by mini-assessment center).

This is an impressive effort that represents one of the few attempts to examine the relevant constructs at the structural level. However, the model does not include personality. The authors allude to this omission in their conclusion that future investigations should attempt to identify which personal characteristics play a role in determining which employees are offered leadership opportunities. In addition, the model is proposed to test supervisory effectiveness only, as opposed to a full multi-factor model of job performance. The criterion side of their model focuses on one aspect of soldier performance: leading/supervising. Consequently, we still do not have a complete integrated structural model of job performance that includes situational judgment.

In sum, despite the enthusiasm regarding the validity and smaller subgroup mean score differences of SJTs, results vary widely. Differences in test development and

format, research methodology, and methodological weaknesses are partially responsible. One purpose of this research is to utilize the current best-practices information to develop an SJT for use in an applied setting. An integrated model of performance will be proposed to test the criterion related validity of the SJT. This model will include all the relevant predictors and criteria, and the analyses will be conducted using several techniques including SEM, a robust data analysis technique.

What do Situational Judgment Tests Measure?

Differences in the development and format of SJTs, research methodology, and methodological weaknesses appear to account for some of the variability in the validity of SJTs. However, at a more fundamental level, there exists little if any consensus regarding what situational judgment is and what SJTs measure (Hanson, Horgen, & Borman, 1998; McDaniel & Nguyen, 2001). If the nature and content of SJTs were better understood, then it would theoretically be possible to increase their validity and hence their usefulness. In addition to increasing their utility as a preemployment selection tool, an understanding of how and why they work may allow researchers and organizations to use SJTs as a development and training tool.

Various hypotheses have been made regarding the content of SJTs. Some researchers propose that SJTs are measuring a new construct. This proposition is usually based on evidence that suggests that SJTs can account for an additional amount of variance in the criterion over that accounted for by other well-researched predictors of performance such as general cognitive ability and experience. Others propose that SJTs are measuring little more than general cognitive ability or job knowledge or experience, and point to high correlations between SJT scores and measures of these constructs as

evidence. Still others have pointed to the correlations between SJT scores and personality as a clue to its content. Taking a more methodological stance, some have suggested that SJTs are simply a measurement method and that an SJT could be developed to tap any construct of interest. Following is a review of current theories and empirical evidence regarding the nature and content of SJTs. Subsequent to this discussion, it will be proposed that researchers to date have oversimplified the study of situational judgment by examining only the test taker's final response or some summary thereof. A hypothesis will be proposed that suggests that situational judgment is a cognitive information processing task composed of several skills and abilities.

A New Construct. Several researchers have proposed that SJTs are measuring a new construct. At the heart of these arguments is the supposition that "real life" is more complex and requires a different skill set than that required to answer questions on tests of general cognitive ability. In an early but eloquent argument, McClelland (1973) acknowledges that intelligence tests require a response to a very structured, specific and clear question, but that real life situations are much more ambiguous. This is likely why intelligence tests have only modestly predicted real world outcomes. If one wishes to predict real world outcomes, one must develop and administer tests that tap into the ambiguity found in these complex situations.

Researchers in the field of IO Psychology have traditionally referred to the construct being measured by SJTs simply as situational judgment. These researchers propose that this new construct is separate from other variables typically used to explain variability in performance (e.g., general cognitive ability and experience), and provide evidence of such by showing that SJT scores can explain an incremental amount of

variance in performance above what is explained by tests of general cognitive ability. However, the amount of incremental variability explained is often rather small and varies from study to study. For example, in Study 1, Weekley and Jones (1997) conducted regression analyses to examine the incremental validity of their video SJT. When entered into the regression equation after cognitive ability and experience, the video SJT was found to explain an additional 2.5% of the variance in the task performance for the cross-validation sample. Similarly, when the video SJT scores were entered into the regression equation first, experience accounted for an additional 1% and cognitive ability accounted for an additional 2.1% of the variability in task performance of 787 hourly discount retail employees. In Study 2, an empirically scored SJT explained 5.7% more of the variability in the task performance of 148 nursing home employees than cognitive ability and experience alone. The rationally scored video SJT explained an additional 9.6% of the variance in their task performance.

Weekley and Jones (1999) conducted similar analyses in two additional studies that examined the relationship between SJT scores and a composite of both task and contextual measures of performance. In Study One, scores on a paper-and-pencil SJT, when entered into a regression equation after cognitive ability and experience, were found to predict an additional 3.3% of the variability in performance for a sample of 844 retail employees. However, cognitive ability, when entered into the equation after SJT scores and experience, did not predict any additional variance in job performance. Smaller incremental validities were found for a different SJT in Study Two. When entered into the regression equation after cognitive ability and experience, scores on the SJT for 1040 hotel operators accounted for an additional 1.1% of the variance in the job

performance composite. Cognitive ability, when entered into the equation last accounted for a small but significant additional .7% of the variance in the job performance composite.

Stevens and Campion (1999) also found evidence of the incremental validity of SJT scores in predicting performance. In Study One, SJT scores, when entered into a stepwise regression equation after the aptitude test composite, explained an additional 8% of the variance in teamwork performance and an additional 6% in overall job performance. However, the SJT scores were unable to account for a significant increase in the variability explained for task work performance (incremental R^2 = 1%, *ns*). These results are of particular interest given the high observed correlation between the aptitude test composite and the SJT of .81.

Clevenger et al. (2001) examined the incremental validity of SJTs in three different samples with mixed results. The authors administered a battery of tests including tests of cognitive ability, conscientiousness, experience, situational judgment, and job knowledge to three samples: 412 investigative officers, 207 customer service representatives, and 107 engineers (note that the tests varied from sample to sample). Performance measures were then obtained from the participants in these three samples. When added into the regression equation last, the SJTs added incrementally to the prediction of performance for the sample of investigative officers. However, the increase in R^2 was .026, which while statistically significant, may not be practically significant. Furthermore, the SJTs did not add incremental predictive validity over and above the other variables for the other two samples.

Sternberg et al. (1995) propose the existence of a new construct called tacit

knowledge, which is tacit in that it is typically inferred rather than stated. Tacit knowledge is action-oriented, of practical use and relevance to the individual, and informally acquired. They propose that tacit knowledge differs from formal academic knowledge in that it is not trained, but rather attained via on-the-job experiences. Sternberg et al. (1995) provide anecdotal evidence of the difference between cognitive ability and more informal tacit knowledge by describing individuals who while possessing modest amounts of general cognitive ability are nevertheless able to generate and utilize sophisticated and novel techniques for job success. Conversely, they point to those individuals with high IQ scores who are unable to appropriately negotiate social situations. Despite the fact that tacit knowledge researchers use somewhat different terminology and have often not referred to the large body of situational judgment literature, tacit knowledge tests have been proposed to be so similar in content and strategy to SJTs that are routinely examined by IO psychologists as to be indistinguishable (see McDaniel et al., 1997, 2001).

Sternberg et al. (1995) provide indirect empirical evidence for the existence of a new construct by pointing out that while the population correlations are debatable, measures of general cognitive ability account for substantially less than 100% of the variance in real world measures of success. Consequently, a "g-ocentric" view of job performance that purports that general cognitive ability is the only or most important precursor to job or other real world success is inaccurate (Sternberg & Wagner, 1993). Rather, tacit knowledge is an important part of success on the job and in real life. Empirical support for the existence of a "new" construct that is unrelated to *g* comes from Wagner (1987, Experiment 2) who found that scores from a tacit knowledge test were not

significantly related to scores on a verbal reasoning ability test. In addition, Sternberg, Wagner, and Okagaki (1993) report in a series of experiments that scores on tacit knowledge tests "correlated poorly, if at all, with conventional ability-test scores" (p. 225). However, they and others (McDaniel et al., 2001) recognize the restriction of range in their samples, which consisted primarily of high performing undergraduate scholars, and the likely attenuation of the relationship between SJT scores and *g*.

Other researchers also acknowledge that IQ alone is unable to account for all the variance in either academic or on-the-job performance. These researchers, lead by Goleman (1995), propose the existence of a related construct called emotional intelligence. This construct is proposed to be important to the extent that it is measuring a unique non-cognitive skill or ability that allows for the successful negotiation of real world situations and contributes to success over and above the contribution of cognitive ability. In a recent summary article, Dulewicz and Higgs (2000) acknowledge that the construct is rather nebulous and has been referred to by researchers by an assortment of names including emotional quotient (EQ), emotional literacy, personal intelligence, interpersonal intelligence, and social intelligence. Their summary of the various elements of emotional intelligence identified in the literature makes it clear that researchers have defined and operationalized emotional intelligence in many different ways. However, the authors offer the following definition by the progenitor of Emotional Intelligence, Daniel Goleman, as an appropriate summation: "knowing what you are feeling and being able to handle those feelings without having them swamp you; being able to motivate yourself and get jobs done, be creative and perform at your peak; and sensing what others are feeling and handling relationships effectively" (p. 342). In sum,

emotional intelligence is a non-cognitive skill that allows one to successfully negotiate complex social situations.

Despite the importance of identifying constructs that predict on-the-job success over and above what is explained via cognitive ability tests and recent attempts to hone in on the essence of emotional intelligence, limited empirical evidence from organizations exists. In addition, no methodologically sound test of emotional intelligence currently exists. Dulewicz and Higgs (2000) address these shortcomings in a study of 58 managers. The authors identified competencies and personality traits presumed to be related to emotional intelligence. They used this information to develop an emotional intelligence scale (EQ), an intelligence scale (IQ), and a managerial intelligence scale (MQ) based on sub-sets of questions from the 16 PF, the Occupational Personality Questionnaire, and the Job Competencies Survey. They hypothesized that EQ scores would contribute uniquely to the prediction of performance but that IQ is also important in that some minimum level of IQ is required for success. They examined the relationship between IQ, EQ, and MQ and level of advancement in a study of 58 managers. When examined via multiple regression, the three variables together accounted for 71% of the variance in level of advancement over a 7-year period. In addition, EQ explained a greater proportion of the variance in level of advancement (R^2 = .36) than IQ (R^2 = .27). Furthermore, the combination of IQ and EQ accounts for more of the variability in level of advancement (R^2 = .52) than either of them alone. In sum, despite the difference in the EQ scale and more traditional SJTs, it appears that the EQ construct does have predictive validity for level of advancement in a sample of managers.

Legree and Busciglio (1992) discuss the nature of social intelligence or social

insight, which is proposed to be a sort of practical intelligence. Like Sternberg et al. (1995), they point to differences between academic intelligence and practical intelligence that are generally accepted among the general population. They provide anecdotal evidence of the distinction by pointing out the modest IQ scores of very successful gamblers. They propose that like tacit knowledge, social intelligence allows for the successful negotiation of complex social situations. However, they propose that social intelligence is based on social knowledge, which differs from tacit knowledge in that certain parts of it can be and are often taught. In addition, social knowledge differs from academic knowledge in that social knowledge is uncertain in that it requires making judgments about the likelihood of outcomes, whereas academic knowledge is generally presented as undisputed fact. Consequently, the manner in which individuals acquire social knowledge versus academic knowledge is quite different. For example, academic information is acquired under a fixed reinforcement schedule; that is, any academic question will be consistently rewarded by responding with the same "correct" answer, regardless of how many times it is presented. However, social knowledge is generally "acquired under complex and uncertain reinforcement schedules" (Legree & Busciglio, 1992, p. 5). In any given situation, several behavioral responses may be rewarded or no response may be rewarded. To negotiate these situations and to maximize the likelihood of being rewarded, individuals must rely heavily on probabilities and norms. In fact, it is this ability to use and manage uncertain information that distinguishes social intelligence from tacit knowledge or cognitive ability.

Legree (1995) provides empirical evidence of the existence of a social intelligence factor. The author administered the 49-item U.S. Army SJT of supervisory

skills using a Likert type response scale format, a newly developed dinner conversation scale, and an alcohol abuse scale to a sample of 200 U. S. Air Force recruits. 193 had complete data and were used in subsequent factor analyses. Cognitive ability test scores for all recruits were available from the previously administered Armed Services Vocational Aptitude Battery (ASVAB). The author hypothesized that factor analyzing the data would result in the three experimental scales loading on a separate social factor not captured by the ASVAB. The sample correlation matrix was corrected for restriction of range and then subjected to a confirmatory factor analysis. The hypothesis was confirmed. The CFA yielded five factors: verbal, speed, quantitative, social, and technical, with the three experimental scales loading significantly on the social factor. The author also examined the second order loadings of these five factors on a first order *g* factor. The social factor had a substantial 2nd order loading on *g* (.71 in the sample matrix and .89 in the corrected matrix), suggesting that social intelligence is highly correlated with general cognitive ability. Some caution should be used when interpreting these results as no dinner scale or alcohol abuse scale data were available from a sample who took the SJT under the more traditional forced choice method. Consequently, it is not possible to know whether the observed results were due to the SJT Likert-type response scale format.

Summary evidence of the incremental validity provided by SJTs is found in a meta-analysis by McDaniel et al. (2001). The authors used the average observed validity coefficient between SJTs and performance (r = .26), an estimate of the average validity coefficient between *g* and performance (r = .25), and the average observed correlation between SJTs and *g* tests (r = .36) to estimate the incremental validity of SJTs. When

combined, the SJT and the cognitive ability scores had a validity coefficient of .31, thus outperforming either individual measure. This suggests that SJTs may provide a small increase in criterion related validity over that provided by cognitive ability test scores. Note that the authors chose to exclude the results from a large study from their analyses on the relationship between SJT scores and cognitive ability test scores due to an unusually low observed correlation.

In sum, despite subtle differences in the constructs and the terminology surrounding them (e.g., situational judgment, tacit knowledge, emotional intelligence, social intelligence), these constructs appear to be very similar in some important ways. First, they are all presumed to be necessary for the successful negotiation of complex social situations that are encountered on the job. Second, they are purported to be measured by "situational tests." Finally, researchers have acknowledged both directly (McDaniel et al., 1997, 2001) and indirectly (Legree, 1995) that the constructs and ideas behind them are similar. Indeed, in some cases, the terms appear to be used interchangeably. Consequently, for purposes of this paper, it will be assumed that these constructs are all measuring situational judgment. However, the evidence related to this issue is inconclusive. Consequently, it is still not clear whether one or more new constructs are being captured.

Nothing More Than g. General cognitive ability (*g*) is a well-researched construct that is routinely accepted as one of the best single predictors of training proficiency and job performance (Hunter & Hunter, 1984; Schmidt & Hunter, 1998). Some researchers propose that SJTs are predictive of performance primarily to the extent that they tap *g*. Support for this notion is provided by researchers who conduct analyses to show that

scores on SJTs are highly correlated with scores on cognitive ability (McDaniel et al., 1997, 2001).

For example, Weekley and Jones (1999) examined the relationship between SJT scores and cognitive ability. In Study One, they developed an SJT and administered it to a development sample of 1973 hourly employees from retail stores. After developing a scoring strategy, they administered the revised version to a sample of 844 employees. They found that scores on the SJT correlated with cognitive ability at .42. 1040 hotel operations employees were tested in Study Two using a different SJT with similar results. The correlation between the hotel operator SJT and the same battery of cognitive ability tests was .48. The combined weighted average correlation of the two studies was .45.

Correlations between SJT scores and tests of general cognitive ability of even higher magnitude have also been reported. Krokos (1999) found a correlation of .71 between scores on the Correctional Officer Video Test and scores on a reading and vocabulary test. This result is surprising given that video tests are presumed to require less reading ability and thus less *g* loaded (Chan & Schmitt, 1997). Similarly, Stevens and Campion (1999) found a correlation of .81 between their teamwork SJT and scores on an aptitude test composite.

One group of researchers claims a negligible relationship between their measure of situational judgment and *g*. Sternberg et al. (1995) claim that their tacit knowledge tests are nominally related to *g*. However, others have called these results into question due to restriction of range in the samples of Yale undergraduates (McDaniel et al., 2001).

Summary evidence of the empirical relationship between SJTs and general cognitive ability is provided in a meta-analysis by McDaniel et al. (2001). These authors

reported an estimated population mean of .46. However, the 10th and 90th confidence interval percentiles were .17 and .75 respectively. Further analysis revealed that SJTs based on a job analysis and with less detailed questions had higher relationships with *g* tests. These results suggest that it may be possible to manipulate the size of the relationship between SJT scores and *g* tests, but it is unlikely that one could develop an SJT with a zero correlation with *g* tests.

In addition to the empirical evidence regarding the relationship between SJT scores and *g*, it has been argued that SJT scores and *g* are necessarily related. For example, even researchers who believe that SJTs are measuring a new construct have suggested that cognitive ability and situational judgment are theoretically related. Legree and Busciglio (1992) propose that the ability to use and manage uncertain information is a predictor of social knowledge, and that this skill is less important in structured academic settings where expectations are often explicitly expressed than in complex uncertain social situations. However, they also point out that this doesn't mean that the ability to use and manage uncertain information is unrelated to academic performance. Someone with well-developed social knowledge and intelligence will likely be better able to negotiate relationships with teachers, which could in turn improve academic performance.

Other researchers echo this general sentiment. For example, Northrop (1989, cited in McDaniel et al., 2001) suggests that *g* is an integral part of judgment, which suggests that it is not possible to separate them. Arvey (1986) suggests that *g* serves some executive manager function of other abilities. McDaniel and Nguyen (2001) also propose that any test measuring judgment will necessarily have some relationship with *g*.

In sum, scores on SJTs are almost always found to have a relationship with *g* tests. In addition, many have proposed that the two are necessarily related. However, the reported correlations vary widely.

Job Knowledge or Experience. Some researchers propose that SJTs are measuring job knowledge or experience. Schmidt and Hunter (1993) suggest that Sternberg and Wagner (1993) are proposing the existence of two new constructs: practical intelligence and tacit knowledge. Schmidt and Hunter (1993) disagree with this, and propose instead that tacit knowledge and practical intelligence are the same construct so there is only one new construct being introduced. Furthermore, they propose that regardless of how it is acquired, knowledge is knowledge, not ability. Hence, tacit knowledge as they have described it is not a new construct at all, but simply the demonstration of job knowledge. They support their proposal by providing evidence of the similarity in research findings between the job knowledge literature and the tacit knowledge literature. Finally, Schmidt and Hunter (1993) propose that knowledge and ability are different concepts and consequently, scores on tacit knowledge tests cannot appropriately be compared to *g*, but rather should be compared to performance on other measures of job knowledge.

Empirical evidence for the association between SJT scores and a 5-item self report measure of experience is provided in previously discussed research by Weekley and Jones (1999). The correlation between SJT scores and experience of 844 hourly retail employees in Study One was .26. The correlation between SJT scores and experience of 1040 hotel operators in Study Two was .16. The weighted average correlation across the two studies was .20. Previous research by Weekley and Jones

(1997) provided mixed results. In Study 1, the correlation between video SJT scores and the same 5-item self report measure of experience was .16 for the development sample and .13 for the cross validation sample. In Study 2 the authors compared scores from a rationally and empirically scored SJT. Correlations between the rationally scored SJT and experience were .14 and .20 for the development and cross validation samples respectively. However, no significant correlation was found between scores on the empirically scored SJT and experience for either the developmental or the cross validation sample.

Legree (1995) compared the mean scores of U. S. Army noncommissioned officers to mean scores of U. S. Air Force Recruits on an SJT developed by the Army to measure the supervisory skills. The author found "small to moderate, though significant differences" (p. 253) in mean scores, with officers scoring higher. This suggests that the ability to answer the Army's SJT questions correctly is at least partly a function of military experience.

McDaniel and Nguyen (2001) conducted what Hunter and Schmidt (1990) have termed a "bare-bones" meta-analysis (corrections made for sampling error only) on a group of studies reporting correlations between job experience and SJT scores. They report a mean observed correlation was .05. Included in the analysis was one study that found a negative relationship between job experience and SJT scores. After removing the results of that study from the analysis, the mean correlation increased to .07. Note, however, that the 95% confidence interval both including and excluding the study that resulted in a negative relationship included zero.

Others have examined both job knowledge and experience. Clevenger et al.

(2001) examined the relationship between SJT scores and both job knowledge (or job simulation) and experience across three samples. Results suggest that SJT scores are more highly correlated with job knowledge or a job simulation than with experience. Correlations between SJT scores and job knowledge or job simulation for the three samples were $r = .13$, $r = .19$, and $r = .37$ respectively. None of the correlations between SJT scores and experience for the three samples were significant.

Other researchers have chosen not to test the relationship between SJT scores and job knowledge or experience. Instead, they simply assume that the relationship exists and attempt to control for it. For example, Phillips (1992 and 1993) assumes that job knowledge and situational judgment test scores are related and attempts to control for this effect by providing test-takers with information regarding the job duties and responsibilities as part of the SJT.

In any case, to the extent that there is a relationship between job knowledge and scores on SJTs, this generates a somewhat philosophical debate regarding SJTs as a selection tool. That is, to the extent that SJTs are measuring job knowledge, they are measuring the likely job performance of the test-taker as opposed to the test-taker's potential to do well on the job (McDaniel et al., 2001; see Hanson, 1994). It could be argued that the impact of this dilemma on organizations is unimportant at least to some degree. That is, organizations benefit from selecting applicants who have higher job knowledge and who will likely have better performance than lower scoring applicants. However, to the extent that an applicant possesses the potential to perform work in new and innovative ways that have not previously been rewarded or measured, these tests are inadequate. In addition, to the extent that the SJTs are written, scored, or otherwise

approved by SMEs who are active participants in an organization, SJTs may perpetuate organizational dysfunction. In order to avoid potential misrepresentation of the predictive ability of SJTs and the qualifications (or lack thereof) of those applicants not selected on the basis of the SJT, organizations should be made aware of this effect.

In sum, some evidence suggests that there is a relationship between job knowledge or experience and scores on SJTs. However, there are significant differences in the sizes of the reported correlations. To the extent that experience and job knowledge are not perfectly correlated, these differences are understandable. That is, experience is proposed to be a precursor to job knowledge (Hanson, Horgen, & Borman, 1998). Consequently, experience alone may not increase job knowledge (and hence situational judgment). Rather, it is an employee's ability to profit from that experience that is important. Hanson, Horgen, and Borman (1998) remark that given the variable nature of experience and individual's ability to profit from it, it is surprising that experience alone has been shown to be predictive of SJT scores. Finally, McDaniel and Nguyen (2001) speculate that the relationship between job knowledge and SJT performance is likely to be more robust because the amount of knowledge that one gains on the job may be at least somewhat removed from the amount of time on the job. In sum, it may be useful to gather both tenure and information regarding the relevance of the experiences encountered in investigations of the nature and content of SJTs.

Personality. Only a few studies have examined the relationship between SJT scores and personality. McDaniel and Nguyen's (2001) "bare-bones" meta-analysis of the observed relationship between SJT scores and measures of the Big Five personality dimensions found that agreeableness (r = .25), conscientiousness (r = .26), and emotional

stability (r = .31) have statistically and practically significant correlations with scores on SJTs. However, great variability was found in the correlations, and most of the 95% confidence intervals contained zero. In addition, many of the studies examined were unpublished papers or technical reports rather than articles from peer reviewed journals. Consequently, the true relationship between SJT scores and personality is unknown.

A Measurement Method. One recently proposed theoretical alternative is that SJTs are a measurement method that can be used to test any construct of interest. For example, in a preliminary theoretical paper, Hanson, Horgen, and Borman (1998) acknowledge that SJTs are generally written to test judgment in complex social situations. However, they propose that an SJT could be developed to test any particular construct. Hence, SJTs are not measuring a construct that could be expected to vary from individual to individual, but rather are a measurement method that to date has generally been used to measure job knowledge. To support their hypotheses, the authors present empirical evidence that suggests that SJTs and job knowledge have similar relationships with performance.

McDaniel et al. (2001) and McDaniel and Nguyen (2001) also suggest that SJTs are a measurement method, and that SJTs can be developed to test any number of constructs related to performance. Again, however, it is proposed that although there will be variability in the observed correlations, that any test of situational judgment will likely have a non-zero relationship with general cognitive ability, some aspects of personality, and job knowledge (McDaniel & Nguyen, 2001).

Current Proposition: SJTs Measure a Cognitive Information-Processing Task. Researchers who have shown that SJTs provide incremental validity over other known

constructs provide convincing evidence that a new construct is being tapped. In addition, there is credible summary evidence that SJTs are correlated with *g*, job knowledge or experience, and personality. However, the demonstration that SJT scores are correlated with measures of other constructs does not demonstrate or illuminate the content of situational judgment, which these tests purport to measure. That is, an observed correlation between two variables could be the result of a spurious relationship (meaning that both variables are dependent on a third variable) or a conditional relationship (meaning that the relationship depends on the values of another variable). Consequently, the content of situational judgment and the tests that purport to measure it is still unclear.

What seems irrefutable is that SJTs present test-takers with complex social situations and negotiating them successfully likely requires multiple skills and abilities. Chan and Schmitt (1997) echo this sentiment regarding the complexity of real world situations. They report that situational judgment problems are "nearly always multidimensional in nature in the sense that an adequate solution or handling of the problem would involve several ability and skill dimensions" (p. 145). Similarly, McDaniel and Nguyen (2001) suggest that assumptions regarding the unidimensional nature of situational judgment are "misguided" (p. 107). In addition, they hint at the existence of multiple situational judgment constructs. They propose that asking the test-taker to identify the most or least effective response is measuring the test-taker's knowledge or judgment regarding the correct response. In contrast, they suggest that asking a test-taker to identify what she or he would most or least likely do is likely to result in the measurement of behavioral intentions from those who do not wish to fake, and knowledge or judgment from those who have succumbed to the need to provide

socially desirable answers. Finally, Vasilopoulos, Reilly, and Leaman (2000) contrast the typically unidimensional nature of personality and biodata items with SJT items that "often reflect multiple traits" (p. 56). Despite the intuitive appeal of the multidimensional nature of situational judgment, it has traditionally been conceptualized as a unidimensional construct and measured as a final outcome score. To the extent that job related situational problems are complex and require more than one ability or skill, this is likely an inappropriate approach.

Some researchers have attempted to empirically identify multiple constructs within SJT data. However, they have largely been unsuccessful (Hanson, Horgen, & Borman, 1998). For example, Hanson (1994) conducted a factor analysis of the SJT data collected on a sample of Army NCOs. However, the results did not yield multiple interpretable factors. I propose that the reason these attempts to verify the existence of multiple constructs have failed is primarily because it is the final answers to SJT items or summary scores based on final answers to those items that have typically been subject to analysis. I propose that these final answers represent only the outcome of situational judgment, not situational judgment itself. Furthermore, I propose that this narrow focus on the final answer has led to a lack of proper investigation on the various skills and abilities required for appropriate situational judgment. Northrop (1976) alluded to this discrepancy years ago, although SJT researchers have failed to investigate the implications.

A particularly interesting alternative that encompasses both the complexity of the task and the existence of multiple constructs is the hypothesis that answering an SJT question is a cognitive information-processing task (Robins, 1994), of which judgment is

only a part. This is a particularly intuitive alternative given the nomenclature; that is, an SJT is by definition a measure of judgment within a given situation. Finally, this approach seems reasonable in light of the often-cited notion that the most interesting questions about humans and their behavior require a cognitive approach (Ashcraft, 1994; Connolly, Arkes, & Hammond, 2000). In sum, it seems reasonable that one might gain insight into the nature of situational judgment by adopting an information processing perspective.

If one is to adopt an information processing perspective in an investigation of the nature of SJT performance and its subsequent impact on organizationally important outcomes, that one should look to the long and rich history of research by cognitive psychologists in the area of information processing. While it is not within the scope of this dissertation to review that literature in detail, a summary of current thinking in this area would be beneficial.

In its simplest form, cognitive information processing follows from some environmental stimulus and includes the following processes: perception, retrieval from memory, comprehension, judgment or decision-making, and finally, some response (Ashcraft, 1994; Flach, 1999; Radford, 1996). Note that despite the fact that these steps appear to be sequential, it is generally accepted that decision makers do not necessarily cycle through the stages sequentially; a decision maker may cycle through the stages in random order (Ashcraft, 1994). In addition to their nonsequential nature, the stages are also not presumed to be processed serially, but rather may be processed in parallel (Ashcraft, 1994), and may be influenced by feedback from previous processing (Flach, 1999). The steps, then, are more "intimately linked" (Flach, 1999, p. 122) than the stage-

model might make it appear. However, despite the somewhat artificial partitioning of the stage model, this model continues to be used by convention, recognized by researchers as valuable (Flach, 1999), and an appropriate way to study certain phenomena of interest (Ashcraft, 1994). In addition, it may be possible to manage the nonsequential and parallel issues via one's choice of statistical analysis. This issue will be addressed in subsequent sections.

An information processing approach to situational judgment would presume, then, that situational judgment requires some awareness and understanding of the situation, retrieval of relevant information from memory, generation of alternatives and various consequences, comparison of the alternatives and their consequences, selection of a preferred alternative, development of a behavioral intention, and finally, a choice to act or not. Again, the terminology provides an accurate reflection of the process. The "situation" in situational judgment is reflected in the perception, retrieval, and generation phases, and the "judgment" in situational judgment is reflected in the comparison of the choices, the selection of an alternative, and the choice of whether to act. Furthermore, an information processing perspective would suggest that SJT scores are a function not only of the test-taker's judgment in the final steps of the process, but also of his or her prowess in all the steps prior to judgment and of various situational characteristics. We turn now to a summary of a methodological issue that is relevant to decision making and the administration of SJTs.

Researchers have traditionally examined the performance of decision makers in the laboratory (Randel, Pugh, & Reed, 1996). These situations are typically rather sterile, often introducing simple well-defined dilemmas with a clear goal, in which a single

decision maker is required to make a judgment. This lack of fidelity in the decision-making tasks generally yields results that are overly simplistic and not generalizable to the "real world." More recently, a paradigm shift of sorts has occurred in decision-making research (Cannon-Bowers, Salas, & Pruitt, 1996). Researchers have begun to embrace the complexities inherent in the "real world," and have begun to examine decision-making and judgment under more naturalistic high fidelity settings. These situations contain more of the elements that are typical in daily life such as poorly defined problems, a changing context that alters goals and creates feedback loops, time constraints, important consequences, multiple constituencies and decision makers resulting in competing or changing goals, and possible conflict between organizational goals and the goals of the decision maker (Orasanu & Connolly, 1993).

Using the basic cognitive information processing steps as a guide and assuming a naturalistic decision making setting, it is proposed that situational judgment in the "real world" is the outcome of a cognitive information processing task composed of the following tasks (see Table 1).

While naturalistic situational judgment may be comprised of these basic tasks, researchers engaged in pre-employment selection and research generally assess situational judgment via a multiple-choice paper-and-pencil test. This method increases standardization and provides the ability to test large groups of applicants simultaneously. However, these tests are considerably different than "real life" tests of situational judgment. Specifically, situation awareness in a naturalistic setting may be affected by any number of personal and situational issues. In contrast, situation awareness in a test-taking environment is limited primarily by the amount of time available for scanning the

Table 1

Situational Judgment in a Naturalistic Setting

Situational Judgment Tasks
Perception – Situation awareness
Retrieval – Retrieval of relevant information from memory
Information-seeking – Decision maker may seek additional information
Thinking – Generation of possible alternatives and likely consequences
Decision-Making – Compare the alternatives
Situational judgment – Selection of a response
Development of behavioral intention – Decision to act (or not)
Action – Act on the choice (or not)

written situation and by the test-taker's motivation and ability to infer meaning from the situations presented. Second, while one may engage in additional information seeking activities in a naturalistic setting, this is not possible in a test-taking situation. Third, in naturalistic settings, one must generate the behavioral alternatives and if so desired, the likely consequences. Test-takers, on the other hand, are presented with a list of alternatives. They may or may not generate novel alternatives in addition to the ones provided and they may or may not consider the consequences of the alternatives provided. Their only task is to choose from among the alternatives provided. Fourth, because alternatives in a naturalistic situation are self-generated, the comparison process among those alternatives may be different than in a test-taking situation where the alternatives to be compared may not have been previously encountered or considered. Fifth, the selection of the best alternative in a pre-employment test-taking situation may

be disproportionately complicated by attempts to identify and select the socially desirable alternative(s). Finally, in naturalistic settings, one must decide on the appropriate behavioral response, develop a behavioral intention to act (or not), and then act on that intention (or not). Test-takers are required only to select a response alternative from the ones offered and to indicate that preference on paper; there is no test of whether she or he would develop a behavioral intention or would act on that intention. On the basis of these observations, it is proposed that situational judgment measured via a paper-and-pencil test is comprised of the following steps (see Table 2).

Table 2

Situational Judgment Measured via a Paper-and-pencil Multiple Choice Test

Situational Judgment Tasks
Perception – Situation awareness
Retrieval – Retrieval of relevant information from memory
Generation – Generation of consequences for alternatives provided and/or generation of novel alternatives
Decision-Making – Comparison Process
Situational Judgment – Selection of a response from those provided

Admittedly, there are significant differences in naturalistic situational judgment and test-taking situational judgment. Consequently, researchers who subscribe to the recent trend toward examining more naturalistic decision-making may consider paper-and-pencil SJTs to be a sub par method of assessing judgment. However, using the previously discussed elements of naturalistic decision making as a guide (Orasanu & Connolly, 1993), SJTs may have higher fidelity than it may first appear. First, when

SJTs are used as selection tools, the stakes are generally quite high. Decisions made in this environment often do have significant outcomes for the test-taker. Second, the situations presented in SJTs are ill defined to the extent that the test-taker must choose his or her response based solely on the information presented. Information seeking is often possible in naturalistic situations, but is not possible in paper-and-pencil administrations of SJTs. Third, the motivation of the test-taker to do well on the SJT in order to be selected may create substantial conflict as he or she attempts to reconcile differences between personal goals and preferences and perceptions of the organization's goals. Finally, it may be possible to create an SJT that contains some of the dynamic qualities of a real world situation. For example, as previously discussed, it may be possible to create a paper-and-pencil test that progresses in a temporally natural sequence. While paper-and-pencil administration of an SJT prohibits the situation from advancing on the basis of a test-taker's answer, advancing the series on the basis of choices that the test-taker might have made, or the introduction of new or contradictory information likely provides a more realistic situation than if the situations advanced solely on the basis of the test-taker's choices. Finally, time pressure could be introduced by timing a portion of the SJT.

In sum, an information processing approach that consists of perception, retrieval, generation, and judgment appears to be an appropriate way to conceptualize situational judgment. However, despite the intuitive nature of the proposal that SJTs are measuring a cognitive information-processing task, and despite the fact that judgment has been referred to as a process (Northrop, 1976), SJT researchers in the field of IO psychology have generally not considered this idea. Only one reference to this possibility was found

in a doctoral dissertation published in 1994. Robins (1994) speculates that situational judgment is an information-processing task likely composed of the following steps: perception, retrieval of similar past experiences and social norms from memory, and selection of an effective response. However, the speculation that situational judgment is an information-processing task is not the hypothesis of Robin's research (her hypotheses were discussed in a previous section), and no test of this idea is conducted. No SJT research could be found that has tested this notion of cognitive information processing directly. Given that no empirical evidence could be found, it is not known whether the information processing conceptualization of situational judgment is appropriate. This research will attempt to answer this question by proposing and testing a cognitive process model of situational judgment.

In addition, despite the appeal of developing a more naturalistic SJT by developing items that are temporally dependent and creating time pressure, the effect of these format issues is unknown. This research will attempt to answer these questions by developing a relatively naturalistic paper-and-pencil SJT that includes dependent situations that advance with temporal continuity and administering a portion of the SJT under timed conditions.

A Cognitive Process Model of Situational Judgment

A cognitive information processing perspective of situational judgment suggests that the tasks involved in answering a situational judgment question via paper-and-pencil test are situation awareness (SA), retrieval of similar experiences from memory (RETR), a generation or comprehension phase that consists of the generation of possible outcomes and novel alternatives (GEN), and a decision making process whereby the test-taker uses

various strategies to make the decision regarding which alternative to choose (DM). These steps will lead ultimately to the selection of an answer, which represents the outcome of situational judgment (SJ). To the extent that different skills and abilities may be required at each step, and consequently that individuals are expected to vary in their ability to complete these steps, they are presumed to be constructs that can be operationalized via questions on a paper-and-pencil test. This section proposes a cognitive process model of situational judgment that illustrates these constructs and their relationships (see Figure 1). Note that the constructs are not assumed to be discrete, and hence are allowed to correlate. Following is a summary of the literature regarding these constructs and plans to operationalize each.

Situation Awareness. Although cognitive psychologists often use the term *perception* to denote an individual's initial response to the stimulus, the term *situation awareness* is often used by judgment and decision making researchers to denote the knowledge that results from both perception and comprehension of the stimulus. Given the simplicity and flexibility that this approach offers in terms of operationalization in a paper-and-pencil administration, this research will conceptualize perception as situation awareness and will capitalize on existing research and models of situation awareness.

Situation awareness has been defined differently by various researchers and has been used in the study of applied problems in a variety of fields (Endsley, 1995a). Endsley (2000b) proposes that a basic definition of situation awareness is "knowing what is going on around you" (p. 5). A key notion is that the individual is not only aware of the current task or situation, but can identify the facets that are most important. Thus, situation awareness refers to knowledge about a dynamic situation, as opposed to

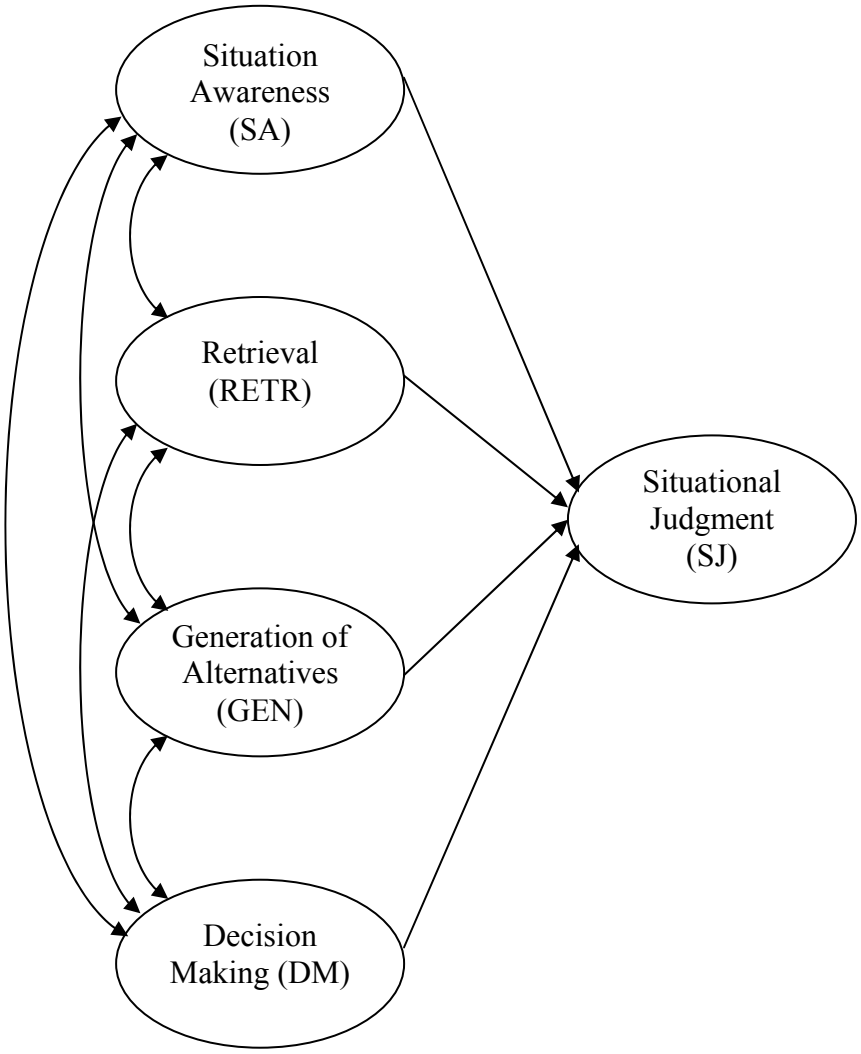

Figure 1. Cognitive Process Model of Situational Judgment

the processes involved in acquiring it or static knowledge. Endsley (2000b) reports a surge of interest in this construct beginning in the mid 1980's, likely as a result of increases in technology and information, which have complicated decision making tasks for workers in aviation, manufacturing, medicine, and other fields.

Several researchers have proposed theories regarding situation awareness. However, Endsley (1995a) proposes a general model of situation awareness within the decision-making context that, unlike some of the other models, is transportable across a variety of contexts. In addition, as proposed, the situation awareness construct can be operationalized, making it particularly appealing. Endsley's model incorporates various issues surrounding the task, the system, and the individual that influence decision-making. Specifically, the situation awareness construct is proposed to be a function of three hierarchical phases. Level 1 situation awareness is perception of the elements in the situation. This is the lowest level of awareness and consists primarily of knowing what elements are present. Level 2 situation awareness is comprehension of the stimuli. This phase requires the individual to identify the task or situation goals in order to understand the relative importance of the individual elements identified in the perception phase. Level 3 situation awareness is projection of future status. In this phase, the individual speculates regarding the likely future actions of the various elements present in the task or situation. The three phases of situation awareness are proposed to be influenced by characteristics of the person including attention, working memory, and long term memory, and characteristics of the task including system design, stress, workload, complexity, and automation. Finally, these three phases lead to decision-making, and

then ultimately to action or performance. Note that the link from situation awareness to performance is not always direct; that is, an individual could have sufficient and appropriate situation awareness and yet still perform poorly if his or her decision making process was flawed or if some other factor such as time pressure interfered. It is generally considered a necessary but insufficient condition for performance.

The situation awareness model is transportable and in fact, Endsley (1995a) acknowledges that the model cannot specify elements that are important across different domains. Rather the user must identify the elements specific to the task and operationalize the phases accordingly. However, Endsley (1995b) does provide a useful theoretical discussion of various methods of operationalizing situation awareness. Examples include relatively objective measures such as performance measures, subjective techniques such as self, peer, or supervisor ratings of situation awareness, and questionnaires that are designed to reveal the individual's situation awareness. Questionnaires can be administered posttest (after the task), online (during the task), or via a freeze technique, which entails stopping the task and asking individuals questions about the situation at that time. No technique is perfect: the posttest technique may be subject to memory issues, and the online technique may be influenced by current workload. The freeze technique, however, avoids these two pitfalls. Furthermore, empirical evidence using the freeze technique (Endsley, 1995b) suggests that subjects can accurately report situation awareness data for at least five or six minutes after the freeze. A second study (Endsley, 1995b) suggests that the performance of subjects is not affected by stops in the task, most likely because the questions used to elicit information regarding

the test-taker's situation awareness keep the pertinent issues in memory so that when the task resumes, she or he is able to continue the task with little decrement in performance.

Endsley (2000a) provides empirical evidence that this conceptualization of situation awareness can be operationalized and utilized in effective ways. She describes the Situation Awareness Global Assessment Technique, which is a generic tool and set of procedures that can be used to determine the situation awareness requirements of any particular job, which in turn can be used to develop the questions necessary to tap into those facets of situation awareness for that job. The resulting responses allow for the quantification of situation awareness. Endsley provides evidence that this technique has been shown to be effective in evaluating and comparing fighter aircraft systems, the potential impact of "free flight," and the automation of automobile navigation systems. The contribution of this technique is that these system issues did not always influence performance directly, but that they did influence one or more of the levels of situation awareness. The identification of the specific level of SA affected allowed for the determination of specific steps necessary to address the decrements.

In sum, Endsley's three-level model of situation awareness is based on precise definitions and processes and has been operationalized successfully. It will be used as a guide to develop the current research. Specifically, the current research will employ Endsley's three-level model of situation awareness using a modified freeze technique. Because the SJT will be administered via paper-and-pencil rather than on a computer, it is not possible to freeze the "screen" and hence prevent participants from referring back to the item stem while answering the situation awareness questions. Consequently, no

attempt will be made to query participants regarding Level 1 issues. However, in my opinion, it is appropriate to assess Level 2 (comprehension) and Level 3 (projection of future status) phases in the paper-and-pencil environment. Level 2 situation awareness (comprehension of the stimuli) will be operationalized two ways: being able to identify the dilemma and to identify the most important goal presented in the item stem. Level 3 (projection of future status) will be operationalized by asking participants to identify the correct projection of future circumstances of the elements. Given that time is an important consideration in each of the situations presented in the SJT, this operationalization makes intuitive sense. It is hypothesized that variability in these measures will be realized even though participants will be able to look back to the item stem. Furthermore, given that the task is standardized, characteristics of the task are primarily controlled for, and consequently, differences in levels of situation awareness will be due primarily to individual differences.

Retrieval. Retrieval (RETR) is an important facet of cognitive information processing. The retrieval process suggests that when faced with a situation for which she or he must make a decision, an individual will attempt to retrieve relevant information and experiences from memory. If the situation(s) drawn from memory (gained through experience or training) are similar to the situation at hand, then the individual is said to have a mental model or schema of the situation (Endsley, 1995a). It is presumed that to the extent that these models are accurate, they will allow the individual to make better judgments.

Note that Endsley's model (1995a) incorporates retrieval from memory in the

situation awareness construct. However, RETR is being hypothesized in this dissertation as a separate construct for several reasons. First, in order to determine if retrieval from memory is being employed as part of the process, it is necessary to measure it separately from the outcomes such as comprehension of the stimuli and projection of future status. More simply, a four factor model more closely resembles the conventional conceptualization of information processing. Second, a separate RETR construct allows for the possibility that one might have adequate comprehension of the stimulus and projection of future status without having memories from which to draw. Third, the idea that RETR might be a separate construct is in keeping with Endsley's (1995a) comments regarding the nature of situation awareness: that alone it is not a guarantee of good performance.

In sum, retrieval of experiences from memory is an important part of the information-processing task. It will be operationalized in two ways in the present research: first, by asking the test-taker to indicate how many similar personally-experienced situations she or he happened to recall from memory, and second, by asking the test-taker to rate the similarity of their personal experience to the details presented in the SJT item.

Generation. General information processing models contain a "thinking" or overall comprehension stage. Ashcraft (1994) suggests that this phase consists of development of alternatives and their meanings, and overall evaluation. It is proposed that in the multiple-choice test-taking environment where the test taker is not required to generate alternatives but rather only to select from among those offered, this construct

refers to processes by which the test-taker may generate novel alternatives that are not listed as a response alternative, or may generate potential consequences or outcomes for the various alternatives that are offered. In keeping with Ashcraft's (1994) conceptualization, the construct is called generation (GEN) to denote this idea that something "new" is being created and considered by the test-taker. Like retrieval, this idea is generally incorporated into the situation awareness piece of Endsley's process model. However, for similar reasons, it will be treated as a separate construct in the present research.

The GEN phase will be operationalized in two ways in the present research. First, by asking the test-taker to report how many novel alternatives she or he generated and second, by asking test-takers for which of the alternatives she or he generated a possible consequence or outcome. This will reveal both for which alternatives she or he generated consequences and by extension, how many of the alternatives she or he considered, hence providing both qualitative and quantitative information regarding this process.

Decision-Making: There are many judgment and decision-making theories that attempt to describe how a decision maker might make a final judgment or choice of response. An important distinction among the theories and research methodologies is whether the goal is to investigate how individuals should make decisions (the normative case) or whether the goal is to examine how individuals actually make decisions (the descriptive case). Normative models assume that the decision maker is a rational being who will estimate the probability of the various outcomes and the subjective utility or value of the outcomes, and then select the alternative that maximizes the possibility of the

largest subjective reward (Loke, 1996). Specifically, the class of Expected Utility theories proposes that the decision maker utilizes any number of numeric calculations of probability and utility to compare the options. Proponents of this category of theories suggest that their superiority lies in the fact that individuals generally want to make good and rational decisions (see Dawes, 1998 for a review).

Criticisms of the utility theories are based in the fact that probability and utility are often insufficient to account for decision making (see Loke, 1996 for a review). That is, in a variety of circumstances, individuals do not make rational decisions based on numeric calculations of subjective reward. Instead, decision makers often use a variety of alternative strategies such as heuristics that are generally designed to reduce the amount of effort required to reach a decision or other affective strategies, both of which are subject to biases and errors. For example, Rivero, Holtgrave, Bontempo, and Bottom (1996) provide empirical evidence of the inadequacy of the Expected Utility theories to explain decision-making. The authors present participants with a cognitively complex game, the St. Petersburg gamble, for which participants have to pay to play. Results suggest that participants do not calculate expected utility as a way of determining how much money they would pay to play, presumably because the calculation requires a lot of mental effort. Instead, participants make their decision using an expectation heuristic that simplifies the decision rule by breaking it down into parts. Unfortunately, the use of heuristics can lead to errors.

The realization that decisions-makers do not always make rational decisions has led some to suggest that utility evaluation is just one of many decision making strategies

that may be used, and that it is the context that will ultimately determine which strategy will be used (Rettinger & Hastie, 2001). In support of this idea, research from a variety of disciplines including psychology, medicine, and law has identified a plethora of factors that play a role in decision-making. For example, characteristics such as expertise (Randel, Pugh, & Reed, 1996), culture (Radford, 1996), fatigue and stress (see Thomas, Doyle, & Browning, 1996 for a review), attractiveness (see Sim & Loke, 1996 for a review), gender (Miller, 1996), and reference point (Dawes, 1998) have been shown to influence decision making strategies and choices. Characteristics of the task such as framing (Levin & Reicks, 1996), content, complexity, and alternatives presented (Dawes, 1998), and subject matter (Rettinger & Hastie, 2001) may also have an impact on how decisions are made. Finally, aspects of the situation such as the existence of time pressure can have an influence on judgment and decision-making (Radford, 1996), especially if one is relatively inexperienced (Klein, 1997).

This veritable laundry list of factors that have been shown to have an influence on a decision maker's choice have left researchers in a bit of a quandary. How might one study a phenomena that is so complex and situation specific? In 1996, Cannon-Bowers, Salas, and Pruitt lamented the fact that there is no comprehensive model or taxonomy for determining how people make decisions. These authors call for such a model and suggest that the strategies could be organized based on whether they are a function of the situation, task, or the decision maker. Indeed, two researchers have recently proposed integrated models that attempt to identify all the factors that can influence judgment and decision making (Endsley, 2000b; Radford, 1996). These models include characteristics

of person, situation and task. They also include process issues such as strategies, and responses and criteria for good decision-making. However, the models are so complex as to be at best difficult to test empirically.

So how might the present research address the decision-making construct? While both the normative and the descriptive approaches seem reasonable, it is the opinion of this researcher that the decision regarding which approach is more appropriate should be determined by the situation and the research objectives. For example, a predictive validation study is in many ways conducive to the normative approach. That is, job applicants are highly motivated to perform well on preemployment tests, and are proposed to be more likely to engage in rational decision-making. Furthermore, many of the characteristics that influence how individuals make decisions are held constant in this testing environment. Indeed test-takers in the present context are relatively homogenous and will all be performing the same task under the same conditions. Given these facts, it may seem appropriate to assume that individuals will process rationally, i.e., that individuals could be expected to scan all the alternatives and to select an alternative based on rational calculations of probability and subjective utility. A multiple-choice investigation of the strategies used by test-takers would only need to contain rational decision-making strategies.

However, there are several convincing arguments for utilizing the descriptive approach in the current study. First, this study examines the concurrent validity of an SJT. The motivation of the test-taker for whom the SJT administration has no real consequence may be somewhat more variable than for the test-taker who may be denied a

job on the basis of his or her scores, suggesting a somewhat weaker inclination to process rationally.

Second, recent research in the field of naturalistic decision making suggests that experts differ from novices in the decision making process. One theory of naturalistic decision-making, Recognition Primed Decision model (RPD), asserts that individuals can use their experience to make sound decisions without having to perform extensive analysis on the alternatives (Klein, 1997). Generally, this model posits that experts focus on assessing the situation rather than assessing different courses of action and judging one option to be superior to others. Individuals are proposed to use their knowledge and experience to recognize problems that they have previously encountered and for which they already know solutions. Given the complexity of decision-making in a naturalistic setting, no one model is likely able to address all the issues involved. In fact, the RPD model does not address attentional or metacognitive processes (Klein, 1997). However, the RPD model does provide some insight into how experts can capitalize on their experience to arrive at sound decisions "without having to compare the strengths and weaknesses of alternative courses of action" (p. 287). Given that the current study will include the scores of participants with varying degrees of experience, it seems reasonable that there may be differences in the processes they utilize in the decision making process.

Finally, there is a lack of empirical evidence regarding the use of decision-making strategies in the situational judgment test-taking environment. No research could be found to date that investigates strategies being used during an SJT administration.

Given these three factors, it is proposed that a hybrid approach is the best possible approach. That is, it is proposed that in the decision-making phase (DM), test-takers will strive to be rational in their decisions, but that various nonrational strategies may also be utilized. More specifically, the expectation is that the preferred strategy will depend on the situation. This facet of DM will be operationalized by presenting participants with a list of possible comparison strategies including both rational (numeric or probabilistic) and emotive (less objective or affective) strategies. Participants will be asked to rate the importance of each of these strategies in making their decision for that particular question. Data analysis will consist of comparing the strategies used across situations presented in the SJT.

To address the issue of expertise, effort will be made to identify the extent to which the individual is using an expert strategy (using a mental model to identify a preferred course of action and then scanning the alternatives to find that or a similar action) versus using a novice strategy (scanning all the alternatives and comparing them). Unfortunately, a paper-and-pencil administration does not allow for a calculation of the amount of time spent engaging in these tasks. Consequently, the expertise issue will be addressed by asking the participant to report whether she or he knew what she or he would likely do after reading the question and then only went to the list of alternatives to locate that action, or whether scanning and comparing all the alternatives was necessary. The participant will be presented with a list of strategies that lie on a continuum of most "recognition" based to most "comparative." Participants will be asked to identify which method they used while making their decision.

In sum, individuals often use rational strategies when making decisions, but they do not always do so. It is unknown what decision making strategies will be used by participants in the present research, although there is reason to believe that experts will engage in a slightly different process. Furthermore, there is reason to believe that the strategy used will depend on the situation. Consequently, the present research will simply attempt to identify which decision-making strategies participants prefer differ by situation and whether experts engage in different process than novices.

Situational Judgment. The criterion of interest in the cognitive process model is situational judgment. This construct will be operationalized using a situational judgment inventory developed specifically for this research. In keeping with previously identified format issues, the SJT presents three situations. Each situation has its own main character and is composed of a series of dependent multiple choice questions that unfold in a temporally natural sequence. The participants will be required to select the answer that they would most likely do and the answer they would least likely do. It was not possible in the current research to gather effectiveness ratings for each of the response alternatives from the SMEs. Rather the SMEs simply indicated the alternative that they believed was most effective and the alternative they believed was least effective. Consequently, despite the appeal of generating several SJT scores based on the SME ratings, the present research will generate two SJT scores: an M-correct score, which is the percentage of questions to which the participant selected the SMEs most effective alternative as the alternative she or he would most likely do and an L-correct score, which is the percentage of questions to which the participant selected the SMEs least effective

alternative as the alternative she or he would least likely do. The development of the SJT questions will be discussed in greater detail in the Methods Section.

Summary. In sum, there is sufficient theoretical reason to propose that answering an SJT item is a cognitive information-processing task composed of several specific tasks that are not necessarily sequential or serial in nature, and that situational judgment depends to some extent on the situation. Testing the cognitive processing model will assist in determining the efficacy of the information processing approach. If it can be shown that SJT test-takers engage in information processing tasks in answering an SJT question, this would assist in illuminating the content of situational judgment. In addition, such a result may assist in reconciling current differences in opinion regarding the content of situational judgment and the tests that measure them. That is, to the extent that the information processing stages require various levels of general cognitive ability, experience and personality, the researchers whose ideas are presented here may all be at least partially correct. SJTs may measure *g*, job knowledge, and personality to varying degrees. For example, situation awareness has been shown to be a function of job knowledge, with experts generally scoring higher (Randel, Pugh, and Reed, 1996). Second, the retrieval process is by definition related to the number and nature of previous experiences. Third, the efficiency and speed with which one can perform various cognitive information processing tasks have been shown to be a function of *g* (Vernon, Nador, & Kantor, 1985). Fourth, it seems reasonable that personality will have an influence on which decision-making strategy is employed. For example, those high in conscientiousness may pay particular attention to organizational goals in their decision

making, while others who are more flexible may attend more to their own goals. Finally, if situational judgment is a function of several skills or abilities, then this may explain why it often accounts for more of the variance in performance than using any of the measures alone. Support for this notion that differing opinions regarding situational judgment may all be correct is provided in a previously discussed study by Legree (1995). After finding evidence of a separate social intelligence factor that also loads highly on a first order *g* factor, Legree concludes that Sternberg and Wagner (1993) who believe that practical intelligence is a new construct separate from *g*, and Jensen (1993) who predicted that tacit knowledge tests would be *g* loaded, are perhaps both correct. Examining SJTs in a cognitive processing light may shed some light on these issues.

An Integrated Model of Performance

Situational Judgment as a Partial Mediator. Not only is there debate regarding the nature and content of situational judgment, there continues to be differing opinions regarding the role of situational judgment in predicting performance. A great deal of the research to date regarding SJTs examines situational judgment simply as one of several predictors of job performance. Typically, cognitive ability, job knowledge or experience, personality, and situational judgment (or some subset of these variables) are proposed and tested as predictors of job performance. However, if one is to understand the nature and content of situational judgment and how to influence it, one must consider both its predictors and its role in influencing subsequent job performance. This marks a shift in the conceptualization of SJTs from being an independent variable to being a dependent variable. Several researchers have embraced this perspective. For example, in one of the

few efforts to combine multiple predictors of job performance in one theoretical model, Campbell (1990) proposes that differences in declarative knowledge (DK), procedural knowledge and skill (PKS), and motivation (M) predict subsequent differences in job performance. Furthermore, individual differences in ability, personality, learning and experience, and motivation are responsible for individual differences in DK, PKS, and M. To the extent that SJTs measure PKS (see Krokos, 1999), Campbell's model suggests that *g*, experience, and personality are predictors of situational judgment, and that situational judgment is a mediator of the relationship between these predictors and job performance.

Additional theoretical evidence is provided by Motowidlo, Borman, and Schmit (1997) who suggest that cognitive ability and personality influence contextual habits, skills, and knowledge, and task habits, skills, and knowledge, which in turn influence job performance. They build this proposition on the work of others (e.g., Campbell, 1990) who have also suggested the existence of mediators between individual differences and job performance. To the extent that SJTs are a measure of task or contextual knowledge, this suggests that situational judgment may be a partial mediator of the relationship between these predictors and job performance.

Borman et al. (1993) provide empirical evidence of the utility of an SJT as a mediator in the previously discussed study of supervisory experience that included a job knowledge construct that was operationalized as scores on a SJT. The results suggest that SJT scores are a reflection of one's job knowledge, which in turn has an influence on one's performance. Taken together, the theoretical and empirical work of these authors

suggests that situational judgment can best be conceptualized as a partial mediator between various predictors and performance. However, the current work does not include a test of a complete integrated model of job performance that illuminates the contribution of situational judgment. This section proposes an integrated model of performance that examines the direct and indirect relationship between the job performance predictors and job performance criteria, highlighting the role that situational judgment plays as a partial mediator.

Predictors of Situational Judgment. There is both theoretical and empirical support for *g* as a predictor of situational judgment. Legree & Busciglio (1992) suggest that the "ability to learn and combine uncertain contingencies may be an important predictor of the ability to function in many ambiguous social situations" (p. 7). Empirical evidence that *g* may be a predictor of situational judgment is provided by Weekley and Jones (1999). In Study One, job performance was regressed onto a model of cognitive ability, experience, and SJT scores. Cognitive ability, when entered into the equation last was unable to account for any more of the variability in performance than SJT and experience. However, SJT scores, when entered after cognitive ability and experience scores, did add significantly to the prediction of performance. This suggests that SJT scores are more strongly related to job performance than cognitive ability, and that cognitive ability perhaps influences performance through its effect on situational judgment. However, results from Study Two did not support this hypothesis.

Theoretical evidence for experience as a predictor of situational judgment is provided by Flach (1999), who proposes that feedback from a response in an information

processing task likely in turn influences future perception and decision making. To the extent that those with greater experience have solicited or been exposed to greater or more accurate feedback, they are more likely to have more accurate situational judgment. Further theoretical evidence is provided by Hanson, Horgen, and Borman (1998) who propose that SJTs measure job knowledge, and consequently two of the predictors of SJT scores are experience and the ability to profit from that experience. Training is proposed to be a type of experience. The authors acknowledge that experience has demonstrated significant correlations with SJT scores, and as such is likely a relevant predictor. However, they caution that some of the most obvious measures of experience (e.g., tenure) may be inadequate in that the mere passage of time does not ensure that the employee was exposed to important job-related situations or that she or he profited from them.

Additional theoretical support for experience as a predictor of situational judgment is provided by Legree and Busciglio (1992). These authors frame the situational judgment problem in terms of reinforcement theory. They propose that individuals in social situations want to maximize the likelihood of reinforcement (e.g., the probability of B given A). However, it "may take years of experience" (p. 4) to learn to estimate the possible consequences of actions under the variable reinforcement schedule that exists for most complex social situations. Furthermore, given that reinforcement in many situations may actually be random, and not a direct result of any particular behavior or action, the ability to estimate the probability of B given not_A is even more complex and could theoretically take longer to acquire.

Researchers have also proposed a theoretical relationship between personality and situational judgment. For example, Hanson and Ramos (1996) suggest that SJTs are measuring something more than *g*, and that perhaps this "something" is personality. Vasilopoulos, Reilly, and Leaman (2000) suggest that an SJT item may tap several personality traits. They propose that a sales clerk who is faced with a customer wishing to return an item without the original sales receipt might display conscientiousness (follow store policy) or display emotional stability (remain calm).

Empirical evidence of a relationship between personality and situational judgment is provided by McDaniel and Nguyen (2001) in their previously discussed meta-analysis. Recall that the authors report that three of the Big Five personality dimensions have a statistically and practically significant relationship with SJT scores: agreeableness (r = .25), conscientiousness (r = .26), and emotional stability (r = .31). However, these relationships are reported entirely at the measurement level, which only suggests that the scores are correlated. Measurement level analyses are insufficient to illuminate the latent structure among constructs.

Only one empirical study could be found that examined personality as a predictor of situational judgment at the structural level. In a previously discussed dissertation, Hanson (1994) tested the relationship between various traits and situational judgment. Specifically, she examined Dominance, Dependability, and Work Orientation as predictors of situational judgment. Results suggest that Dependability has a direct effect on SJT scores and that Dominance and Work Orientation have direct effects on SJT scores and indirect effects on SJT scores via supervisory experience and training. The

robust nature of these analyses lends very credible support to the idea of personality as a predictor of SJT performance. However, the most widely accepted taxonomy of personality to date is the Big Five taxonomy consisting of Openness, Conscientiousness, Extraversion, Agreeableness, and Need for Stability. Given the general acceptance of this taxonomy as an adequate summarization of the major components of personality, it is important to consider which if any of the Big Five factors might be adequate predictors of situational judgment.

In sum, there is some evidence that suggests that *g,* experience, and personality may all contribute to successful and appropriate situational judgment, which in turn may predict job performance. However, some of this work is theoretical. The empirical work that exists did not incorporate all the relevant constructs (e.g., *g,* experience, and personality), the most widely accepted taxonomies (e.g., the Big Five), or the best data analysis techniques (e.g., SEM). Consequently, the extent to which *g,* experience, and personality can be operationalized adequately, and the extent to which these constructs make independent contributions to situational judgment and direct contributions to performance is unknown. This study will attempt to answer these questions empirically.

Predictors of Performance. In addition to being proposed as possible predictors of situational judgment, research suggests that *g,* personality, and experience also contribute directly to job performance. The most robust predictor is *g*, which has been found to be highly and consistently predictive of performance across a variety of jobs. Hunter and Hunter (1984) found that *g* as measured by the General Aptitude Test Battery (GATB) had a mean validity of .54 for training success criterion and .45 for job

performance criterion. However, because mean scores on *g* tests vary significantly among subgroups, their use as a selection tool may create unacceptable levels of adverse impact (Hunter & Hunter, 1984). Consequently, researchers and organizations have been motivated to identify variables that could predict job performance independent of cognitive ability.

Although personality is a logical choice as a potential predictor of job performance, for many years the search for valid personality predictors was largely unfruitful. This was likely due to the lack of consensus regarding the nature of personality and how it should be classified and operationalized (Barrick & Mount, 1991). More recent analyses employing the Big Five personality taxonomy have been more successful. Barrick and Mount (1991) found in their meta-analysis that conscientiousness is consistently and significantly related to job performance across several occupations, with estimated true validities ranging in size from .20 to .23. Extraversion and agreeableness were also found to be significant albeit less robust predictors of performance for some job categories.

Performance Criteria. Austin and Villanova (1992) suggest that compared to the predictors of job performance, the criteria have been "the orphans of the validation process" (p. 860). This is not to say that job performance has not been examined. Rather, the focus has been primarily on the measurement of performance at the expense of understanding its latent structure. These researchers and others (Campbell, 1990; Hanson, Horgen, and Borman, 1998) have called for reduced reliance on job performance measures and an increase in attention to the underlying latent structure of job

performance.

Some progress is being made in identifying the number and nature of the constructs that comprise the criterion space. In one of the first attempts to develop a model of job performance, Campbell (1990) proposed that a theoretical 8-factor taxonomy adequately represents the criterion domain of virtually all the jobs listed in the Dictionary of Occupational Titles: Job-specific Task Proficiency, Nonjob-specific Task Proficiency, Written and Oral Communication Tasks, Demonstrating Effort, Maintaining Personal Discipline, Facilitating Peer and Team Performance, Supervision, and Management/Administration. Not every job is proposed to contain all eight constructs, but every job will require some combination of them.

Borman and Motowidlo (1993) propose that job performance has both a task and a contextual dimension. Task performance behaviors are those that are performed in direct service to the task and hence to the organization. In contrast, contextual performance behaviors are volitional behaviors that indirectly serve the organization and the people in it. Borman and Motowidlo suggest that there are five categories of contextual behaviors: volunteering for extra assignments, being persistent in accomplishing tasks, assisting coworkers, following organizational policies regardless of their individual impact, and formally and informally supporting organizational goals. Motowidlo and Van Scotter (1994) provide empirical evidence of the independent contributions of task and contextual performance using a sample of Air Force mechanics. When entered into a hierarchical regression equation last, supervisory ratings of task performance explained 13% more of the variance in ratings of overall job performance

than contextual performance alone. Conversely, contextual performance, when entered into the regression equation last, explained 11% more of the variance in ratings of overall job performance than using task performance alone.

Grant (1997) hypothesized that the criterion space for State Troopers could be modeled by combining Campbell's eight-factor model and Borman and Motowidlo's two-factor model. Although support was not found for the hypothesized models, an exploratory factor analysis of the criterion space found three interpretable factors: Know In-Role, Do In-Role, and Extra-Role. Grant and her colleagues (Wilson & Grant, 1997; Krokos, 1999) renamed these factors Knowing the Job, Doing the Job, and Citizenship in subsequent analyses.

In sum, there is significant evidence that suggests that job performance is multidimensional. Several situational judgment researchers have embraced these recent job performance models. For example, Weekley and Jones (1999) and Stevens and Campion (1999) employed Borman and Motowidlo's multidimensional model and examined the effects of various predictors on both task and contextual performance criteria. However, the majority of SJT researchers have not employed the most recent constructs or model of job performance. Rather they have focused on the measurement aspect of performance by simply collecting supervisory ratings of various job performance behaviors and then collapsing the individual ratings into an overall or composite performance score by averaging across those behaviors (Weekley & Jones, 1997). Other SJT researchers have recognized that job performance is multidimensional but have then collapsed the performance measures into a composite. For example,

Clevenger et al. (2001) had supervisors rate the 9 or 10 performance dimensions identified in job analyses for that job category. However, they collapsed these dimensions into a single performance measure. This is likely an oversimplification of the relationship between SJT and performance. This research will investigate the relationship between the previously identified predictors, and multiple dimensions of performance to include both "knowing" and "doing" facets of the task dimension and a contextual (i.e., citizenship) dimension.

An Integrated Model of Performance. Although great strides have been made, there currently exists no integrated model of job performance that incorporates all the relevant predictors and their direct and indirect effects on performance. Fortunately, however, recent work on job performance modeling has incorporated latent constructs that are transportable across jobs and situations. The purpose of this research is to incorporate what is known about integrated models of job performance and to apply this knowledge to the prediction of performance in a sample of undergraduate university scholarship recipients. It is proposed that *g,* experience, and personality are direct predictors of performance and that situational judgment is a partial mediator of the relationship between these predictors and performance. That is, *g,* experience, and personality are proposed to influence the multiple facets of performance directly and also to influence performance indirectly through situational judgment. Specifically, it is proposed that *g,* experience, conscientiousness, agreeableness, and need for stability are all necessary and significant predictors of situational judgment.

In keeping with the best available research regarding the multidimensional nature

of performance, it is proposed that the criterion space for scholarship recipients is multidimensional. Specifically, the relevant performance dimensions are Academic Performance (AP), Program Specific Performance (PSP), and Contextual Performance (CP). In keeping with the evidence that suggests that the predictors influence performance directly, *g* and experience influence AP and PSP and the three measures of personality (conscientiousness, agreeableness, and need for stability) will influence the CP dimension of performance. The conceptual integrated model of performance is presented in Figure 2.

Summary. This literature review has shown that despite the fact that SJTs demonstrate criterion related validity, some SJTs appear to "work" better than others. In addition, although mean subgroup scores on SJTs have been shown to be smaller than their *g* test counterparts, some SJTs seem to have higher differences than others. Furthermore, there exists little agreement regarding what situational judgment is or what SJTs are actually measuring. This research proposes that situational judgment is an information processing task. The content of the situational judgment will be examined by testing the information processing tasks proposed to be required. Second, this research proposes that situational judgment is a partial mediator of performance. This idea will be tested by using situational judgment as a partial mediator in an integrated model of performance that contains *g,* experience, and personality as predictors, and academic performance, program specific performance, and contextual performance as the performance criteria. The following section details the specific hypotheses to be tested.

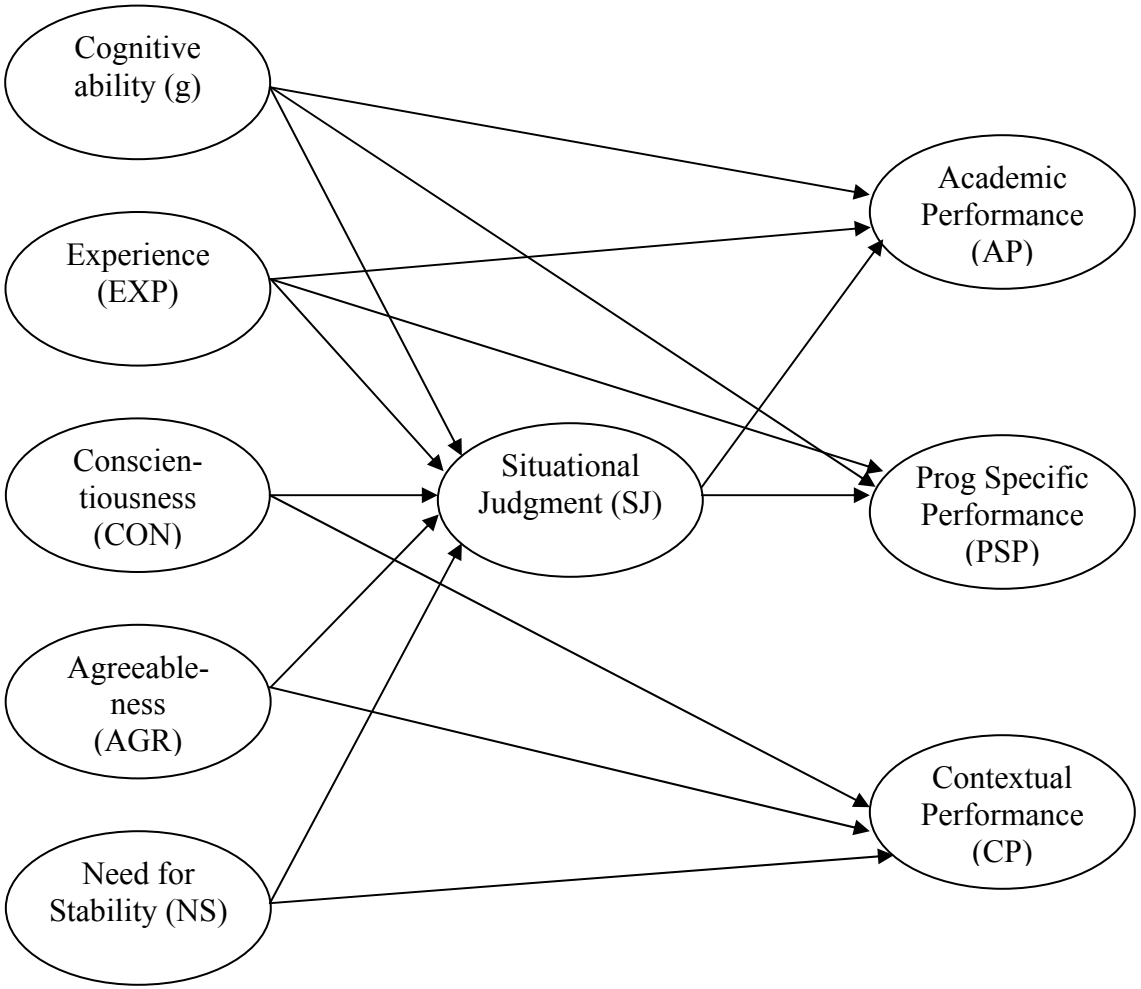

Figure 2. Integrated Model of Performance

Research Question #1

What is the relationship between information processing variables and scores on a situational judgment test? This research will develop an SJT using the best methods possible. It is hypothesized that situational judgment as measured by this SJT is a cognitive information-processing task that requires several skills and abilities. This research will employ regression and correlational techniques to examine the individual level hypotheses. Structural equation modeling will be employed to examine whether the latent constructs can be operationalized adequately using data from undergraduate scholars. The specific hypotheses regarding the cognitive process constructs are as follows:

H1 = Participants who have high situation awareness will have higher overall SJT scores.

H2 = Older participants and those who report retrieving a greater number of more similar past experience(s), will have higher overall SJT scores.

H3 = No hypothesis is made regarding the generation of alternatives (GEN). That is, it is unknown whether participants will report having generated alternatives in addition to the ones provided or whether they will report having generated consequences for all or part of the alternatives, or whether either of these generation tasks will be associated with higher, lower, or similar SJT scores.

H4 = Older participants and those who report retrieving a greater number of more similar past experiences will engage in more recognition based processing and less comparison based processing than younger participants or those with fewer or less similar past experiences. Higher levels of recognition based processing will be associated with higher

SJT scores.

H5 = Overall fit of the cognitive process model of situational judgment will vary by situation and by whether the M-correct or L-correct SJT scores were used.

Research Question #2

Does situational judgment mediate the relationship between important individual attributes (g, personality, and experience) and performance? This research will employ regression and correlational analyses to examine individual level hypotheses. Structural equation modeling will be used to determine whether the latent constructs can be operationalized adequately using data from undergraduate scholars. The hypotheses are as follows:

H6 = SJT scores will be positively related to the academic (AP), program specific (PSP), and contextual dimensions (CP) of performance. However, given the additional social component involved in the PSP and CP, SJT scores will be more highly related to the PSP and CP dimensions of performance.

H7 = *g,* experience (EXP), agreeableness (AGR), and conscientiousness (CON) will be positively related to overall scores on the SJT.

H8 = Need for stability (NS) will be negatively related to overall scores on the SJT.

H9 = *g* will be associated with SJT scores, but will be more highly associated with SJT scores for situations that were administered under timed conditions than under untimed conditions.

H10 = *g* and experience (EXP) will make direct positive contributions to both the academic (AP) and program specific (PSP) dimensions of performance.

H11: Personality will be related to the contextual dimension (CP) of performance. That is, agreeableness (AGR) and conscientiousness (CON) will be positively related to contextual performance (CP) and Need for Stability (NS) will be negatively associated with contextual performance (CP). Evidence regarding the content of situational judgment may be used in further analyses. For example, if situational judgment is found to be composed of the hypothesized information processing constructs, then it would be prudent to consider which predictors have a significant relationship with those constructs. As this model has not yet been tested, no hypotheses regarding this question are being proposed.

Method

Participants

Participants are 233 undergraduate students (scholars) from a large university who are recipients of a highly competitive four-year academic scholarship. Complete data are available for 211 scholars for the cognitive processing model. Of these 211 scholars, 114 or 54% are female and 97 or 46% are male. The scholars are divided almost equally by year: 62 are first year scholars, 50 are second year scholars, 49 are third year scholars, and 50 are fourth year scholars. Complete data for the integrated model of performance are available from 176 scholars. Of these 176 scholars, 97 or 55% are female and 79 or 45% male. The year division is as follows: 63 are first year scholars, 46 are second year scholars, 39 are third year scholars, and 28 are fourth year scholars.

A second sample of 315 undergraduate students enrolled in an introductory

psychology course at the same university is used as a comparison sample.

Measures

Cognitive Process Model Measures. The cognitive process model of situational judgment proposes that situational judgment is the result of four cognitive information-processing tasks: Situation Awareness (SA), Experience (EXP), Generation (GEN), and Decision Making (DM). These factors are operationalized via a series of nine questions embedded in the SJT. SA is operationalized by asking the test-taker to identify the dilemma, the most important goal, and the most important future projection. RETR is operationalized by asking the test-taker to report the number of similar situations she or he retrieved from memory and by rating the most similar memory on a Likert-type scale. GEN is operationalized by asking the test-taker to report the number of novel alternatives generated and the number of response alternatives for which the consequences were considered. Finally, DM is operationalized two ways. First, the test-taker is asked to report how they arrived at their final selection. This score is calculated on a continuum from expert strategy (the test-taker had an answer in mind and used primarily a recognition strategy to locate that answer in the list provided) to novice strategy (the test-taker had no ideas after reading the item stem and had to read and compare all of the alternatives in order to select a response). Second, decision making was operationalized by asking the test-taker to rate a list of nine decision-making strategies on Likert-type scale of importance in their decision.

This series of nine cognitive processing questions is asked following an SJT item; the test-taker is instructed to answer the cognitive processing questions with regard to

how they answered that previous SJT item. Note however, that given the time required to answer the information processing questions, it is not possible for every SJT item to be followed by this series of questions. Instead, the series of information processing questions follows only one question in each of the three situations in the SJT.

Finally, SJ is operationalized six ways: using the sum of the scores of the individual SJT items for each of the three situations using either the M-correct or L-correct scoring strategy. The cognitive process model constructs and manifest measures are summarized in Table 3.

Figure 3 illustrates the operationalized cognitive processing model. The structural model will be tested six times: once for each of the three situations presented in the SJT using either the M-correct SJT score or the L-correct SJT score as the criterion.

Integrated Model of Performance Predictor Measures: The integrated model of performance proposes that *g,* experience, and personality predict performance both directly and indirectly via situational judgment. Following is a description of the manifest variables used to operationalize each factor in this model.

General cognitive ability (g) is operationalized for this sample of scholars using high school grade point average, SAT scores (verbal and math), and a standardized high school rank score. These archival data are available from the scholarship program office.

Experience (EXP) is operationalized four ways. First, graduating year or "Class" serves as a general measure of experience. For example, a scholar admitted in the school year beginning 2001 should graduate in 2005; this is the value of the Class variable for that scholar. The other three manifest variables for the EXP factor come from items on

Table 3

Cognitive Process Model of Situational Judgment Constructs and Manifest Measures

Predictor Constructs (n=4)	Manifest Measures (n=17 per situation)
Situation awareness (SA)	• Identification of the goal • Identification of the dilemma • Projection of future status
Retrieval of past experiences (RETR)	• Number of similar situations retrieved from memory • Similarity of most similar situation retrieved from memory
Generation of Novel Information (GEN)	• Number of novel alternatives generated • Number of consequences considered
Decision Making (DM)	• Expert vs. Notice Strategy (Continuum from Recognition to Scanning) • Decision Making Strategies: nine strategies rated on a Likert scale of importance in answering SJT item
Criterion Construct (n=1)	**Manifest Measure (n=1 per situation)**
Situational Judgment (SJ)	• SJT M-correct summary score for Scholarship (sum of n=5 SJT items) • SJT M-correct summary score for Leadership (sum of n=8 SJT items) • SJT M-correct summary score for Service (sum of n=4 SJT items) • SJT L-correct summary score for Scholarship (sum of n=5 SJT items) • SJT L-correct summary score for Leadership (sum of n=7 SJT items) • SJT L-correct summary score for Service (sum of n=3 SJT items)

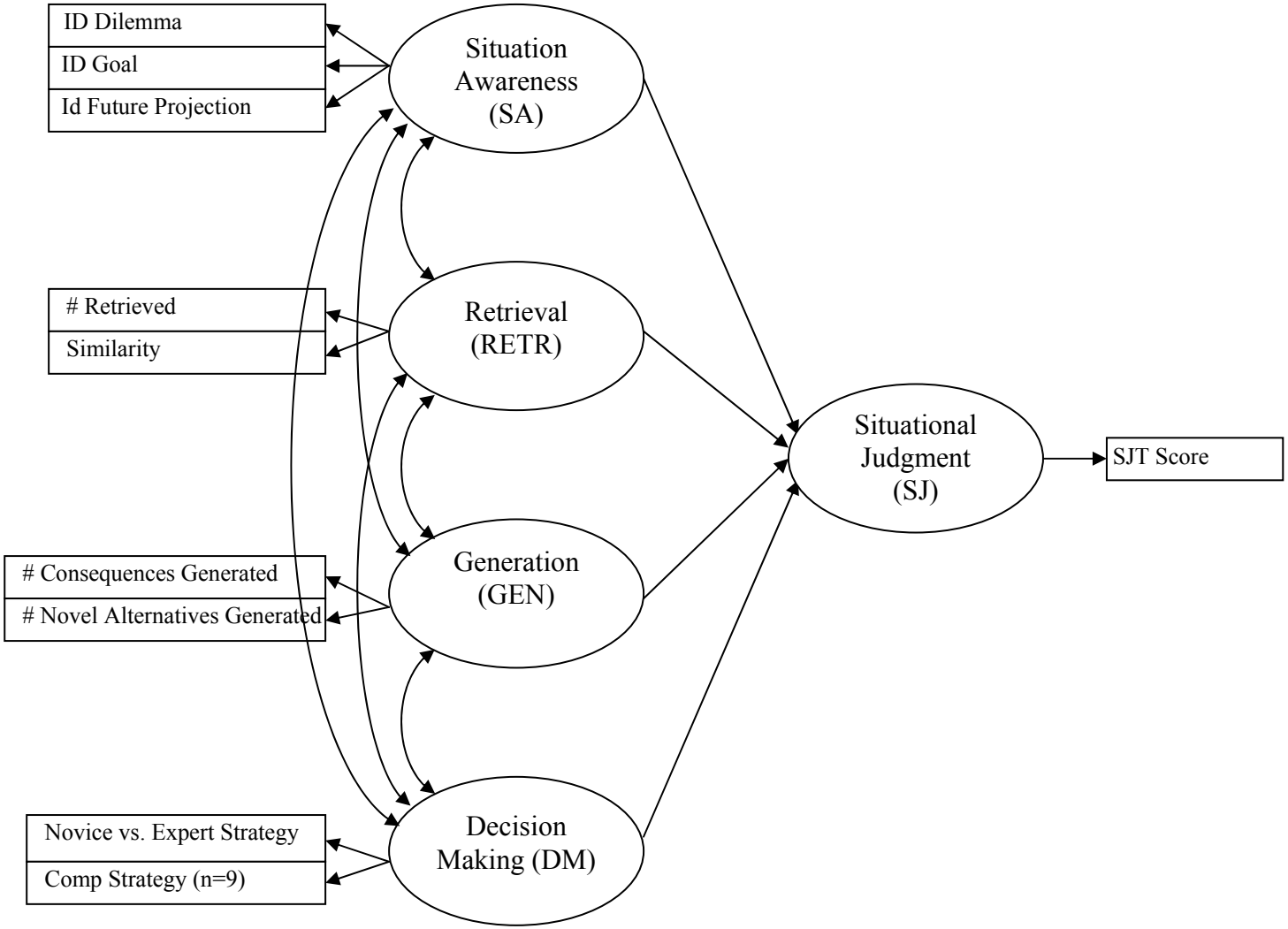

Figure 3. Operationalized Cognitive Process Model of Situational Judgment
Note: Model will be run once for each of the three situations using M-correct and L-correct SJT Scores

the SJT. As part of the cognitive processing series of questions, test takers are asked to report the number of similar situations retrieved from memory and to rate the similarity of that memory to the situation presented on a Likert type scale. These items are the second and third operationalizations of EXP. Finally, the SJT asks the test-taker to estimate the amount of experience she or he has had with dilemmas like the ones presented in each of the three situations. The final manifest variable for the EXP factor is this self rated experience item from the SJT.

The three personality factors (Conscientiousness, Agreeableness, and Need for Stability) are operationalized using the six factor subscores from the Revised NEO Personality Inventory (NEO PI-R™), which was administered to the scholars in the spring of 2002. The predictor constructs and their respective manifest measures are summarized in Table 4.

Integrated Model of Performance Criterion Measures: The integrated model of performance proposes that performance is composed of three factors: Academic Performance (AP), Program Specific Performance (PSP), and Contextual Performance (CP). AP is operationalized using academic performance measures from the spring of 2002 including semester grade point average, total grade point average, and rank in class. These data are available from the scholarship program office.

The PSP and CP factors are both operationalized by ratings of performance. A performance appraisal rating instrument was developed on the basis of the criterion dimensions and behaviors list and was administered specifically for research purposes. The ratings were completed in the summer of 2002 by the scholarship program director

Table 4

Integrated Model of Performance Predictor Constructs and Manifest Measures

Predictor Construct	Manifest Measure
Cognitive Ability (g)	• SAT Score – Verbal (SATV) • SAT Score – Math (SATM) • High school GPA (HSGPA) • High school class rank (HS RANK)
Experience (EXP)	• Graduating year (CLASS) • Number of similar situations retrieved from memory • Similarity of most similar situation retrieved from memory • Amount of experience with similar situations
Conscientiousness (CON)	• NEO PI-R™ C1: Competence • NEO PI-R™ C2: Order • NEO PI-R™ C3: Dutifulness • NEO PI-R™ C4: Achievement Striving • NEO PI-R™ C5: Self-Discipline • NEO PI-R™ C6: Deliberation
Agreeableness (AGR)	• NEO PI-R™ A1: Trust • NEO PI-R™ A2: Straightforwardness • NEO PI-R™ A3: Altruism • NEO PI-R™ A4: Compliance • NEO PI-R™ A5: Modesty • NEO PI-R™ A6: Tender-Mindedness
Need for Stability (NS)	• NEO PI-R™ N1: Anxiety • NEO PI-R™ N2: Angry Hostility • NEO PI-R™ N3: Depression • NEO PI-R™ N4: Self-Consciousness • NEO PI-R™ N5: Impulsiveness • NEO PI-R™ N6: Vulnerability

(the supervisor), who consulted with others on an as needed basis to make the ratings on all the scholars. Peers also completed the ratings in the spring of 2002 using the same instrument. PSP is operationalized using supervisor and peer ratings of two facets of

Scholarship, two facets of Leadership, and one facet of Service. The final operationalization for PSP is the mean number of community service hours (self-reported number of hours per semester averaged over the scholar's tenure at the university). Finally, CP is operationalized using supervisor and peer ratings of four facets of Character. The criterion constructs and their manifest measures are summarized in Table 5. The operationalized integrated model of performance is illustrated in Figure 4.

Table 5

Integrated Model of Performance Criterion Constructs and Manifest Measures

Criterion Construct	Manifest Measure
Academic Performance (AP)	• Semester GPA (SGPA) • Total GPA (TGPA) • Rank in Class (Class Rank) • Program Rank (PG Rank)
Program Specific Performance (PSP)	• Ratings of Scholarship by faculty/staff panel and peers • Ratings of Leadership by faculty/staff panel and peers • Ratings of Service by faculty/staff panel and peers • Mean number of community service hours
Contextual Performance (CP)	• Ratings of Character by faculty/staff panel and peers

Procedure

Overview. The research process consists of five stages: criterion development, SJT development, performance rating instrument development, data collection, and data analysis and reporting. See Table 6 for a general description and timeline.

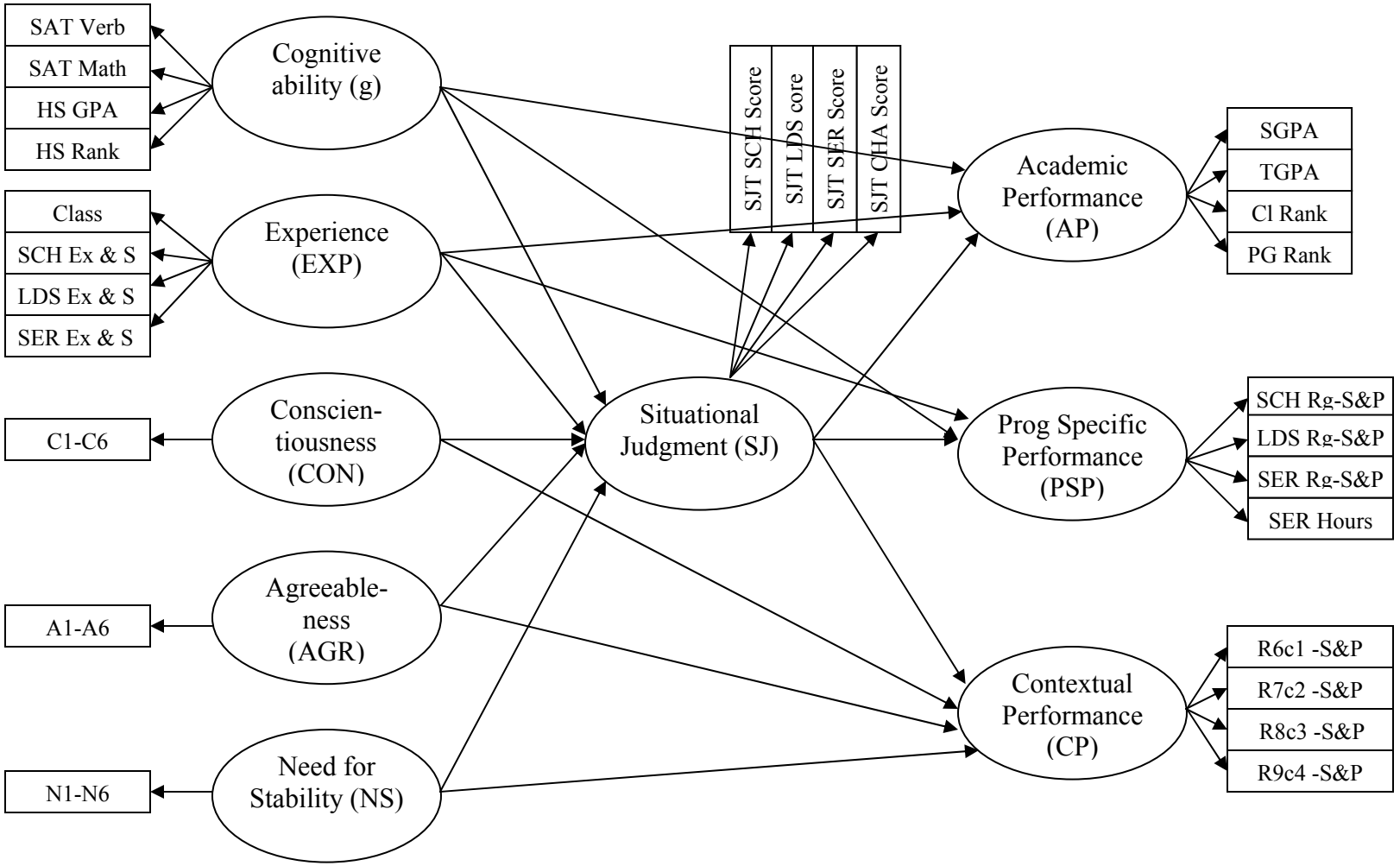

Figure 4. Operationalized Integrated Model of Performance (Note: each personality predictor has six associated scores. The six boxes have been collapsed into one for ease of use.)

Table 6

Project Overview

Task	Date
Criterion Development	Oct 2001 – Nov 2001
SJT Development	Nov 2001 – Feb 2002
Performance Rating Instrument Development	March 2002 – April 2002
Data Collection	Feb 2002 – July 2002
Data Analysis and Reporting	August 2002- March 2003

Criterion Development. In order to build a valid situational judgment test, one must have a clear understanding regarding what situations are likely to be encountered by incumbents, and what behavioral responses (i.e., performance) are expected. At its inception in 1995, the scholarships program identified four critical performance dimensions for its Scholars: Scholarship, Leadership, Service, and Character. However, these dimensions were never defined at the level of specificity required to develop an SJT. As late as the fall of 2001 when this research project began, there was little consensus, particularly outside the program directors, regarding what behaviors appropriately reflected these dimensions. This lack of consensus was evidenced by low reliabilities in the interviewer ratings (Pond & Cantwell, 2002). Consequently, a criterion development phase was required before proceeding with the development of the SJT.

The first step in the criterion development process was to gather and examine published scholarship program materials. Specifically, program publications, news releases, application materials, interviewer training materials and rating forms, the scholar acceptance agreement, and scholar instructions were examined. Next, several

informal interviews with Program officials were conducted. Finally, the author and the project team drew from experience. From these materials, interviews, and experience, a list of criterion behaviors was developed to be used as a guide in subsequent discussions.

A group of eleven (11) people who have had routine and sustained contact with the scholars were identified as SMEs and were invited to join the project team and to participate in the SJT development. Their first task was to meet for the purpose of defining the criterion space. The pre-generated list of behaviors was not distributed to the SMEs, but rather was used by the author as a guide to facilitate discussion. The SMEs verified that the four established performance dimensions (Scholarship, Leadership, Service, and Character) were adequate and were to be retained. The next step was to develop a list of specific behaviors to appropriately reflect these four dimensions. The resulting list of behaviors was similar in content but more succinct than the researcher's pre-generated list. The list was then sent to each member of the SME team for final comment and revision. Resulting comments were discussed and incorporated as appropriate, resulting in a final criterion behavior list (see Appendix A). This list of behaviors remained intact through the course of the project, although minor modifications were made for clarification purposes.

Development of the Situational Judgment Test. There were a number of issues and concerns to be addressed in the design and development of the SJT. The issues and steps are outlined in Table 7, which is followed by a more detailed discussion. Before developing the situations, there were several basic format issues that needed to be addressed. First, it was decided that paper-and-pencil multiple-choice format would be

Table 7

Development of the Situational Judgment Test

1. Basic Format Issues
    - Administration method
    - Instructions for respondents
    - Length of administration
    - How to assess the information processing steps
    - Appropriate use of SJTs (selection versus development)
    - Appropriateness of situations developed
2. SJT Development
    - Write items and possible response alternatives based on behaviors identified in the Criterion Development Phase
    - Gather SME input regarding items and possible alternatives for 1 of the 3 situations
    - Rework items/alternatives based on SME input
    - SMEs rank order the alternatives for all 3 situations
    - Assess agreement
    - Rework items/alternatives based on SME comments/agreement
    - SMEs rate the most and least effective for all 3 situations
    - Assess agreement
    - Select keyed (most and least effective) items
3. Choose Scoring Method

an appropriate administration method in this application because the cognitive ability of participants is extremely high; they possess more than adequate ability to read and understand complex situations presented in paper form. There was some concern that test takers may be influenced by perceptions of effectiveness based on the gender or race of the main character(s) in the situations. Consequently, gender and race neutral names were used and all gender specific pronouns were excluded from the SJT. In order to reduce the possibility that a test-taker would be unfamiliar with a situation or might base

their answers on individual differences in their personal situations, no reference was made to parents or guardians.

As previously stated, given the lack of physical risk in these situations, it is unlikely that there would be a large discrepancy between what the test-taker would report that she or he "should do" versus what she or he "would do." In addition, given that no administrative decisions were to be made on the basis of the SJT scores and that no one in the scholarship office would have access to individual scores, concerns about impression management were small. Consequently, it was decided that test-takers should be asked, "What would you do if you were in (insert the name of the main character)'s situation?"

There were several administrative concerns regarding the SJT including the length of time required for administration. The goal was to develop a test that could be completed in approximately two hours or less. Consequently, it was decided that the test should contain approximately 18-20 questions with a list of five response alternatives from which the test-taker could choose.

Based on this desire to keep the time required to take the SJT to a minimum and the previous discussion on fidelity in paper and pencil SJTs, it was decided that the questions within each of the three situations should build upon one another in a temporally natural sequence. For example, the main character may be faced with a particular decision in question 1, and the following question would build on that scenario by progressing it forward in time, much as it would happen in "real life." Care was taken not to progress the questions in any discernible pattern. That is, a question might build

on a previous question based on the correct or keyed answer to the previous question, the incorrect response to the previous question, or on the basis of new information. It was felt that this approach of temporal sequencing, although not typical of SJTs, would allow for the presentation of greater detail without significantly increasing the length of the test or the time required to administer it. In addition, it would allow for greater fidelity with real world situations.

In order to address the cognitive process model of situational judgment, a series of questions was included to quantify the cognitive processing steps taken by the test-taker. However, it was decided that following each question in the SJT with the series of nine information processing questions would make the SJT too lengthy and time consuming. Consequently, it was decided that the nine cognitive processing questions would follow only one question in each of the three situations. The decision regarding which question should be followed by the information processing series questions was determined based on which items had the most SME agreement regarding the most and least effective response alternatives.

Another issue regarding the development of the SJT stemmed from the debate regarding the best use of SJTs. Some argue that SJTs are best used for selection purposes, while others argue that such tests are best suited for development and training. It was decided that the SJT developed in the current research could be used for both purposes because it provides both an understanding of the content of situational judgment and the process of situational judgment. That is, the SJT items have one answer that is agreed upon by the majority of SMEs to be the correct or most effective one.

Consequently, the SJT scores can be used as an objective means of comparison and selection decisions could be made based in whole or in part on these comparisons. In addition to its use as a selection tool, answers to the embedded cognitive processing questions could be used to determine how and why the test-taker arrived at his or her answer. For example, it would be clear whether the test-taker chose the correct answer but for the wrong reasons or vice versa. Weaknesses in processing or in logic would be revealed and could then be used for developmental purposes. Confidentiality of the items could be protected by using the original SJT as a template to create a parallel version to be used solely for developmental purposes.

Of critical concern were that the situations be valid and appropriate for the application. It was decided that the situations should be written such that the dilemmas would test the criterion behaviors identified in the criterion development phase. In addition, there was concern that the situations be realistic, reflect the current scholarship environment, and that there should be one clearly correct answer. It was decided that the best way to address these concerns would be to involve the previously identified SMEs in virtually every stage of test development.

Once the basic format issues had been decided, then the task of developing the SJT began in earnest. The first task was to generate scenarios and a list of possible response alternatives based on the criterion behavior list. The project team then evaluated the response alternatives and worked toward consensus regarding which alternative was the "best." The next step was to meet with the SMEs individually. During this one-hour interview, the SMEs were shown one of the three situations without

the pre-generated response alternatives and asked to provide possible alternatives. They were asked to provide alternatives that the "best" scholar would likely choose and behaviors that the "worst" scholar would likely choose. They were then asked to identify appropriate distracters that would likely look like a good choice to an applicant but that clearly did not represent a behavior that the scholarship program would want. On the basis of these data, the items and response options were modified, resulting in a revised SJT with each question now having five response alternatives.

In the next phase, the SMEs were called together to review the new version, rank order the response alternatives, and indicate why they chose the answer they choose. These results were examined to identify possible weaknesses in the item stems. In addition, written comments from the SMEs identified several other problem areas in the situations that were easily remedied.

The SME rankings were also examined to identify possible weaknesses in the response alternatives. The rankings were entered into a database. Based on the SME comments, rankings, and level of agreement, the items and response alternatives were reworked. Note that typical SJT development at this phase would involve simply dropping items for which acceptable agreement was not reached. Given the temporal sequencing of the items in this SJT and the concomitant dependence of the items, this option was not appropriate. Consequently, the author and the project team discussed each item in detail taking into consideration the information from the SMEs. Changes were made specifically to induce agreement.

Once these changes were made, the SMEs were once again asked to review the

entire SJT. They were instructed to indicate the most and least effective alternative for each of the items. The agreement among the SMEs was significantly better with this revised version and those responses were subsequently used to determine the keyed (most and least effective) responses.

The final SJT consists of three situations, followed by between six and eight multiple choice questions, for a total of 21 questions. Each multiple choice question has five response alternatives. One question in each of the three situations is followed by the series of nine information processing questions. The directions for administering the SJT are provided in Appendix B. A sample SJT item is provided in Appendix C.

The final step in the SJT development process was to decide on a scoring method. There is evidence that asking test-takers to rate the appropriateness of on a Likert-type scale results in increased reliability. In addition, if SME ratings of effectiveness can be gathered on each of the response alternatives, this creates even more scoring options (see previous discussion). However, a forced choice method that only requires the test-taker to choose the most and least likely choice also yields a great deal of information about the test-taker and requires less agreement regarding the appropriateness of each alternative. In addition, asking the SMEs to generate a Likert rating on each alternative (as opposed to simply gathering a most and least effective choice) would have in this circumstance been a significant burden. Consequently, it was decided that the respondents would be asked to indicate the alternative that she or he would most likely do and the alternative that she or he would least likely do. Scoring consists of simply identifying the number of "most effective" responses the test-taker identified as the "most likely do" option (for an

M-correct score) and the number of "least effective" responses the test-taker identified as the "least likely do" option (for an L-correct score).

Development of the Performance Rating Instrument. A performance appraisal rating instrument was developed on the basis of the criterion dimensions and behaviors list. The rating instrument presents the following dimensions: managing academic potential, critical thinking skills, leadership skills, seeking/accepting leadership roles, quality of service behavior, self awareness, integrity, adaptability/resiliency, presence, and overall performance. Lists of example behaviors are provided as anchors for each of three levels: needs improvement, effective, and extremely effective. The supervisor rated each scholar on a scale of 1 (*Needs Improvement)* to 7 (E*xtremely Effective*). Peer raters also rated the scholar on a scale of 1-7, but were additionally given the opportunity to rate their global familiarity with the ratee on a scale of 1 (*Hardly at all)* to 5 (E*xtremely well*), and their opportunity to observe the ratee on each performance dimension on a dichotomous scale (0 = L*ittle or no opportunity to observe*, 1= *Opportunity to observe*).

Data Collection: The scholars were administered the SJT in groups in March 2002. The scholars completed the peer ratings of performance and the NEO PI-R™ in groups in April 2002. Supervisory ratings were generated in the summer of 2002.

Results

Overview

The mean SJT score for the Scholarship, Leadership, and Service situations combined is 50.2 (*SD* = 10.9). The mean SJT score for the Scholarship, Leadership,

Service, and Character questions combined is 47.3 (*SD* = 9.9). There is no statistically significant difference between the mean SJT score for Whites and Non-Whites. Similarly, there is no significant difference between mean SJT score for females and males.

The cognitive processing model and the integrated model of performance were analyzed following the two-phase process outlined in Hatcher (1994). The objective in phase one was to identify the best fitting measurement model. While many of the manifest variables used in the current research are established operationalizations, many are new operationalizations developed specifically for this research. Given the incipient nature of several of these variables and the possibility that they may not adequately load on the proposed factor, phase one analyses were used to make decisions regarding modifications that would result in the best fitting measurement model. This required several steps. First, TETRAD II© (version 3.1) was used to conduct a tetrad analysis on factors with more than 4 manifest variables. A tetrad analysis purifies the underlying factor by identifying the variables for which the tetrads will vanish, thus allowing one to eliminate extraneous or poor performing measures. Variables indicated by the tetrad analysis were removed from subsequent analyses. Next, CFAs were conducted on the predictor and criterion side of the models separately to determine the fit of the measurement model. If necessary, modifications were made to improve fit. Finally, the overall measurement models were tested via CFA. These results were reviewed to determine if model fit could be improved. Phase two investigated the structural portion of the models. All models were tested using PROC CALIS in the SAS® System for

Windows (version 8.02). The specifics of the data analyses of the two models follow.

Cognitive Processing Model

Overview. The cognitive processing model is analyzed in two phases. Tetrad analyses and CFAs are conducted on the model in phase one. The structural paths are examined in phase two. Item descriptions and descriptive statistics for the cognitive processing model variables can be found in Appendix D. Correlation matrices for the cognitive processing models are presented by situation in Appendix E – Appendix G.

Phase One. Research Question #1 hypothesized that answering an SJT item is a cognitive information processing task. Specifically, it was proposed that four cognitive processing tasks would be predictive of SJT scores: Situation Awareness (SA), Retrieval (RETR), Generation (GEN), and Decision Making (DM). It was further hypothesized that the fit of the cognitive processing model would vary by situation and by scoring method. Testing the cognitive processing model by the three situations (Scholarship, Leadership, or Service) and by the two criterion scoring methods (M-correct or L-correct SJT score) yields six models, each with four predictor factors predicting the SJT criterion score. The first step in the analysis was to conduct a tetrad analysis on the factors with more than four manifest variables to determine whether any variables should be eliminated. A tetrad analysis on the DM factor suggests keeping all ten of the manifest variables; all the tetrad equations in the measurement model that includes all ten variables pass the Bonferroni test.

Next, a CFA was conducted on the predictor side of the models by situation, allowing the four processing factors to correlate. However, before the analyses could be

conducted, a decision had to be made regarding the SA factor. The manifest variables for the SA factor are binary; that is, the SA factor manifest variables take on a value of 0 if the scholar got the item incorrect or 1 if the scholar got the item correct. Technically this violates the assumptions necessary for use in a CFA and SEM. Various alternatives were considered in an effort to remedy the situation. Although a tetrachoric correlation matrix is more suited to binary data, the use of such matrices is also problematic in terms of the assumptions; that is, using a tetrachoric matrix for these variables does not appropriately allow one to assume an underlying interval level structure for the individual variables. In addition, some of the SA questions were consistently answered correctly or incorrectly, thus making their inclusion in any matrix problematic. Finally, while the SAS® software program does have an external macro available to generate tetrachoric correlation matrices, the macro contains a warning that the data can only be used for descriptive purposes in the SEM program; the hypothesis tests using the Maximum Likelihood Method will be inaccurate. Given these difficulties, it was decided to collapse the three individual manifest variables for the SA construct into a single situation awareness summary score and use that summary score as the single manifest measure of situation awareness. The predictor side CFAs were then run using this summary SA score.

Table 8 contains the results of the predictor side measurement model CFAs for the six cognitive processing models as follows: the Scholarship situation predictor measurement model (ScM_{mp}), the revised Scholarship situation predictor measurement model (ScM_{mp}-Rev), the Leadership situation predictor measurement model (LdM_{mp}), the revised Leadership situation predictor measurement model (LdM_{mp}-Rev), the Service

situation predictor measurement model (SvM_{mp}), and the revised Service situation predictor measurement model (SvM_{mp}-Rev).

Table 8

Goodness of Fit Indices for the Predictor Measurement Models for the Cognitive Processing Model by Situation

	FF	GFI	χ^2	*df*	*p*	RMSEA	NNFI	NFI
ScM_{mp}	.87	.89	183.56	85	<.0001	.07	.57	.52
ScM_{mp}-Rev	.36	.94	76.78	39	.0003	.07	.76	.72
LdM_{mp}	.77	.90	161.49	85	<.0001	.07	.76	.67
LdM_{mp}-Rev	.60	.91	125.19	60	<.0001	.07	.78	.73
SvM_{mp}	.89	.89	186.65	85	<.0001	.08	.71	.65
SvM_{mp}-Rev	.39	.94	81.69	40	<.0001	.07	.85	.81

ScM_{mp} = Scholarship situation predictor measurement model;
ScM_{mp}-Rev = Scholarship situation predictor measurement model – Revised;
LdM_{mp} = Leadership situation predictor measurement model;
LdM_{mp}-Rev = Leadership situation predictor measurement model – Revised;
SvM_{mp} = Service situation predictor measurement model;
SvM_{mp}-Rev = Service situation predictor measurement model – Revised;
FF = Fit Function; GFI = Goodness of Fit Index; χ^2 = Chi Square; *df* = degrees of freedom, *p* = probability; RMSEA = Root Mean Square Error of Estimation; NNFI = Non-Normed Fit Index; NFI = Normed Fit Index;

The results suggest that model fit was generally poor (see Table 8). Specifically, the GFI indices are all .90 or smaller, the NFI and NNFI indices are all less than .85, and the RMSEA values are greater than .05. After reviewing the model fit indices, factor loadings, and the modification indices, modifications were made to the predictor measurement models. Specifically, Q11ba, Q11bf, Q11bh, and Q11bi were eliminated

from the Scholarship situation model, Q26bd and Q26c were eliminated from the Leadership situation model, and Q33bd, Q33bi, and Q33c were eliminated from the Service situation model. The model fit indices for the revised measurement models are also presented in Table 8. After making these modifications, all three models showed marked improvement but still only resulted in moderate fit. The revised model GFIs increased, but are all .94 or smaller, the RMSEAs are still greater than .05, and the NNFI and NFI indices are .85 or smaller.

As the criterion side of the model contains only a single summary SJT score, it was not necessary to conduct a CFA on the criterion side of the model. However, a full model CFA was conducted by situation and by scoring method, allowing the four predictor side factors and the criterion side SJT score to correlate. Table 9 presents the goodness of fit indices for the six full measurement models by situation (Scholarship, Leadership, and Service) and SJT scoring method (M-correct or L-correct score). They are as follows: the Scholarship situation measurement model using M-correct SJT score ($ScMM_m$), the Leadership situation measurement model using M-correct SJT score ($LdMM_m$), the Service situation measurement model using M-correct SJT score ($SvMM_m$), the Scholarship situation measurement model using L-correct SJT score ($ScLM_m$), the Leadership situation measurement model using L-correct SJT score ($LdLM_m$), the Service situation measurement model using L-correct SJT score ($SvLM_m$).

Five of the full measurement model CFAs resulted in only moderate fit as evidenced by FF greater than .38, GFI .94 or smaller, RMSEAs all greater than .05, NNFI .85 or smaller, and NFI .80 or smaller. The final model for the Leadership situation using

Table 9

Goodness of Fit Indices for the Full Measurement Models for the Cognitive Processing Model by Situation and SJT Scoring Method

	FF	GFI	χ^2	*df*	*p*	RMSEA	NNFI	NFI
$ScMM_m$	.40	.94	84.78	46	.0004	.06	.76	.72
$LdMM_m$	One improper solution: one negative eigenvalue.							
$SvMM_m$	.42	.94	89.10	47	.0002	.07	.85	.80
$ScLM_m$	.38	.94	79.59	46	.0015	.06	.79	.73
$LdLM_m$	.62	.91	130.73	69	<.0001	.07	.79	.72
$SvLM_m$*	.46	.93	96.76	47	<.0001	.07	.82	.79

*Results accepted because the factor loading for Q33i is within one standard error.
$ScMM_m$ = Scholarship situation measurement model using M-correct SJT score;
$LdMM_m$ = Leadership situation measurement model using M-correct SJT score;
$SvMM_m$= Service situation measurement model using M-correct SJT score;
$ScLM_m$ = Scholarship situation measurement model using L-correct SJT score;
$LdLM_m$= Leadership situation measurement model using L-correct SJT score;
$SvLM_m$= Service situation measurement model using L-correct SJT score;
FF = Fit Function; GFI = Goodness of Fit Index; χ^2 = Chi Square; *df* = degrees of freedom, *p* = probability; RMSEA = Root Mean Square Error of Estimation; NNFI = Non-Normed Fit Index; NFI = Normed Fit Index

the M-correct SJT score resulted in an improper solution: a factor loading for Q26h of 1.10 and hence was rejected. The model for the Service situation using the L-correct SJT summary score also resulted in an improper solution. However, this model was accepted because the factor loading of 1.004 for Q33i is within one standard error. None of the modifications suggested by PROC CALIS were appropriate, so no further modifications were made to the measurement models.

In sum, tetrad analyses were conducted in phase one, although no reductions were

required. Modifications made to the predictor side measurement models were generally successful, resulting in improved model fit. The full model CFAs yielded moderately acceptable fit, with the exception of the Leadership situation model with M-correct SJT scores, which was rejected. Phase two tests the structural portion of the cognitive processing models by situation and SJT scoring method.

Phase Two. Evaluation of the structural models consisted of several steps: verifying model identification, determining whether the model fit is acceptable, and finally, examining the modification indices for potential improvements to the model. Model identification in the six structural models was verified by determining that the number of data points is greater than the number of parameters to be estimated. The number of data points is calculated using the formula: $(p (p + 1)) / 2$, where p is the number of manifest variables. Using this criterion, all six of the structural models are overidentified.

Next, model fit indices were examined to determine model fit. Table 10 provides a summary of the model fit indices for the original and revised cognitive processing models tested by situation and by SJT scoring method. They are as follows: the Scholarship situation structural model using M-correct SJT score ($ScMM_m$), the revised Scholarship situation structural model using M-correct SJT score ($ScMM_m$ – Rev), the Leadership situation structural model using M-correct SJT score ($LdMM_m$), the revised Leadership situation structural model using M-correct SJT score ($LdMM_m$ – Rev), the Service situation structural model using M-correct SJT score ($SvMM_m$), the revised Service situation structural model using M-correct SJT score ($SvMM_m$- Rev), the

Scholarship situation structural model using L-correct SJT score ($ScLM_m$), the revised Scholarship situation structural model using L-correct SJT score ($ScLM_m$-Rev), the Leadership situation structural model using L-correct SJT score ($LdLM_m$), the revised Leadership situation structural model using L-correct SJT score ($LdLM_m$-Rev), the Service situation structural model using L-correct SJT score ($SvLM_m$), and the revised Service situation structural model using L-correct SJT score ($SvLM_m$-Rev).

When summarized, the indices suggest only moderate fit. While the Chi-Square to *df* ratio is less than 2.0 in every model, Fit Functions range from .38 to .62 and the GFI indices are .94 or smaller. In addition, the RMSEA values are all .06 or greater. The modifications suggested by PROC CALIS were inappropriate in all of the models except the Leadership situation using the M-correct score. The fit of this model was improved by eliminating Q26h. The fit indices for this revised model are also shown in Table 10.

In sum, the model fit indices shown in Table 10 provide evidence of only moderate fit, suggesting that the cognitive processing model of situational judgment may not be appropriate. The factor loadings and path coefficients demonstrated similar or increasing mediocrity as discussed below.

M-correct SJT Scores. Figure 5 shows the factor loadings and path coefficients for the Scholarship situation structural model using M-correct scores. All factor loadings are significant at $p < .05$, indicated by a single asterisk. However, only two of the factor loadings are greater than .70, and the expert decision making strategy question (Q11c) factor loading is unexpectedly in the negative direction. All the path coefficients are large enough to be of practical importance. However, due to high standard errors,

Table 10

Goodness of Fit Indices for Structural Models by Situation and Scoring Method

	FF	GFI	χ^2	*df*	*p*	RMSEA	NNFI	NFI
$ScMM_m$	.40	.94	84.78	46	.0004	.06	.76	.72
$ScMM_m$-Rev	None of the modifications suggested appropriate							
$LdMM_m$	Improper solution: one negative eigenvalue.							
$LdMM_m$-Rev	.58	.91	122.39	58	<.0001	.07	.75	.72
$SvMM_m$	.42	.94	89.10	47	.0002	.07	.85	.80
$SvMM_m$-Rev	None of the modifications suggested appropriate							
$ScLM_m$	.38	.94	79.59	46	.0015	.06	.79	.73
$ScLM_m$-Rev	None of the modifications suggested appropriate							
$LdLM_m$	.62	.91	130.73	69	<.0001	.07	.79	.72
$LdLM_m$-Rev	None of the modifications suggested appropriate							
$SvLM_m$	.46	.93	96.76	47	<.0001	.07	.82	.79
$SvLM_m$-Rev	None of the modifications suggested appropriate							

$ScMM_m$ = Scholarship situation structural model using M-correct SJT score;
$ScMM_m$ – Rev = Scholarship situation structural model using M-correct SJT score-Revised;
$LdMM_m$ = Leadership situation structural model using M-correct SJT score;
$LdMM_m$ – Rev = Leadership situation structural model using M-correct SJT score-Revised;
$SvMM_m$= Service situation structural model using M-correct SJT score;
$SvMM_m$- Rev = Service situation structural model using M-correct SJT score-Revised;
$ScLM_m$ = Scholarship situation structural model using L-correct SJT score
$ScLM_m$-Rev = Scholarship situation structural model using L-correct SJT score-Revised;
$LdLM_m$ = Leadership situation structural model using L-correct SJT score;
$LdLM_m$-Rev = Leadership situation structural model using L-correct SJT score-Revised;
$SvLM_m$ = Service situation structural model using L-correct SJT score;
$SvLM_m$-Rev = Service situation structural model using L-correct SJT score-Revised;
FF = Fit Function; GFI = Goodness of Fit Index; χ^2 = Chi Square; *df* = degrees of freedom, *p* = probability; RMSEA = Root Mean Square Error of Estimation; NNFI = Non-Normed Fit Index; NFI = Normed Fit Index

only the SA factor is statistically significant in predicting the summary SJT score. The loadings and path coefficients for the Leadership situation using the M-correct SJT score are presented in Figure 6. A similar pattern emerged. While all the factor loadings are statistically significant, all of the DM factor loadings are .50 or smaller. The path coefficients are smaller for this model, but even the largest coefficient of .15 for RETR fails to reach significance due to high standard errors. Finally, the model for the Service situation using M-correct scores yielded similar results (see Figure 7). All of the factor loadings are significant, but the factor loadings for the DM factor are .59 or smaller. Two of the path coefficients are significant in this model; SA and GEN have modest but significant predictive relationships with situational judgment (.18 and .17, respectively).

L-correct SJT score. The structural cognitive processing models using L-correct scores as the criterion variable had difficulties similar to the models using the M-correct scores. First, consider the model for the Scholarship situation using the L-correct SJT score (see Figure 8). All of the factor loadings are significant, but seven of the ten are .50 or smaller. In addition, the expert decision making strategy question (Q11c), is opposite in direction than proposed. The path coefficients showed a rather bizarre relationship with the L-correct SJT score; the coefficients for RETR and GEN are practically large at .70 and -.36. However, the standard errors are extremely high for these two factors preventing the coefficients from being statistically significant. The factor loadings and path coefficients for the structural model for the Leadership situation using L-correct scores are shown in Figure 9. As before, the factor loadings are all significant, but eight of the twelve are less than .50. None of the path coefficients are significant. Figure 10

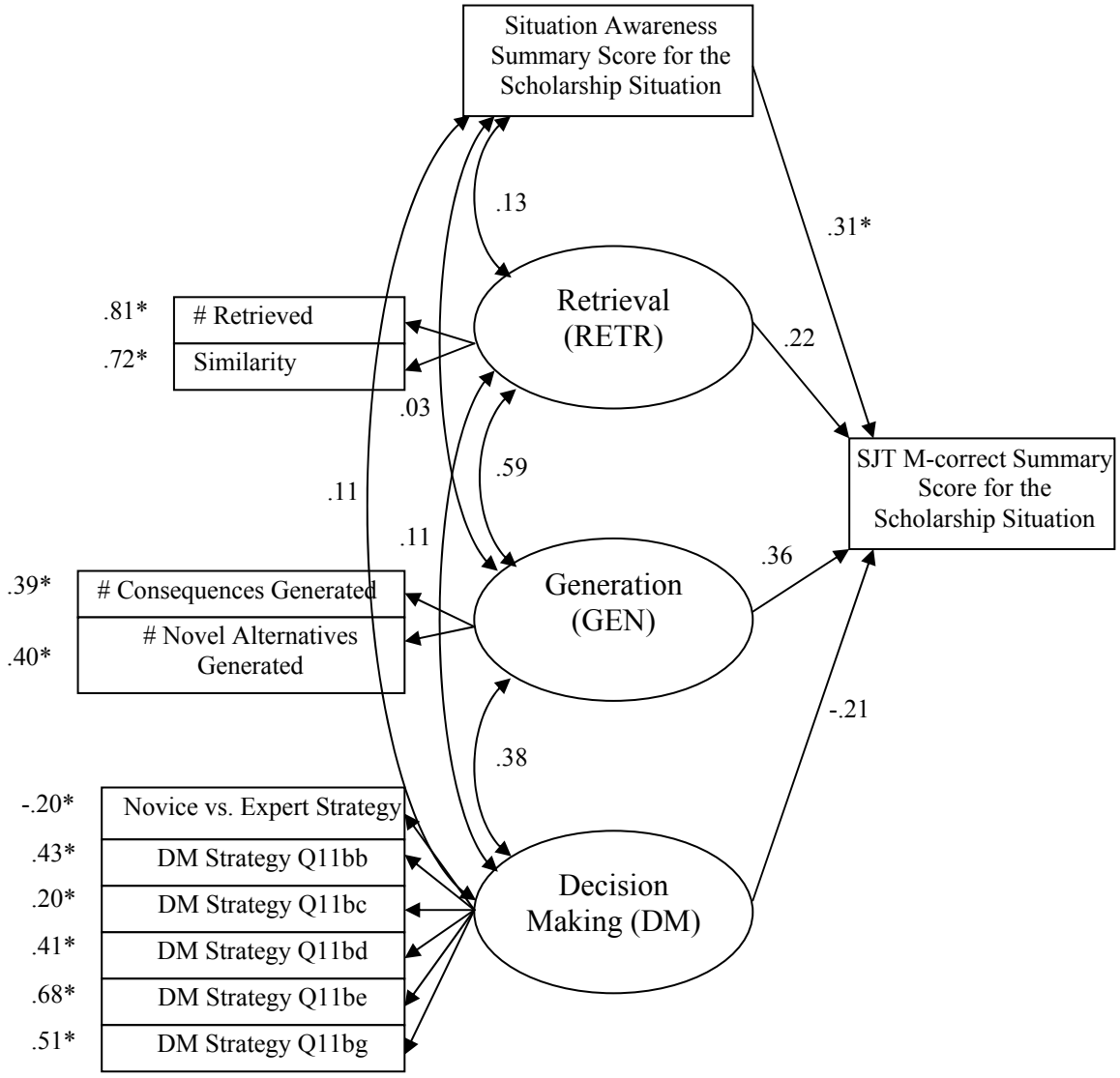

Figure 5. $ScMM_m$: Scholarship Situation Structural Model Using M-correct SJT Score
Note: Data points listed directly beside the manifest variable boxes are standardized factor loadings.

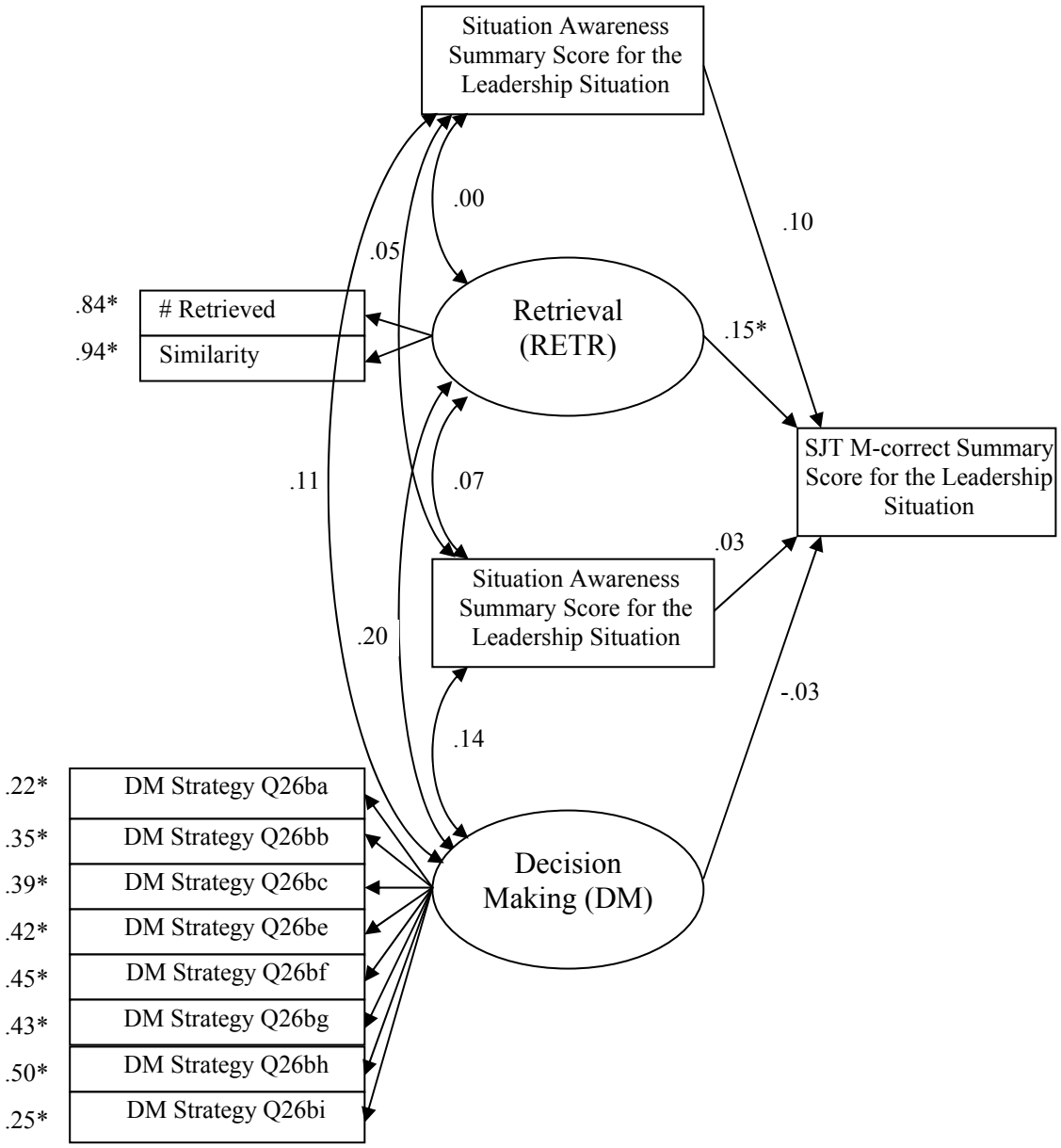

Figure 6. $LdMM_m$ – Rev: Revised Leadership Situation Structural Model using M-correct SJT Score

Note: Data points listed directly beside the manifest variable boxes are standardized factor loadings

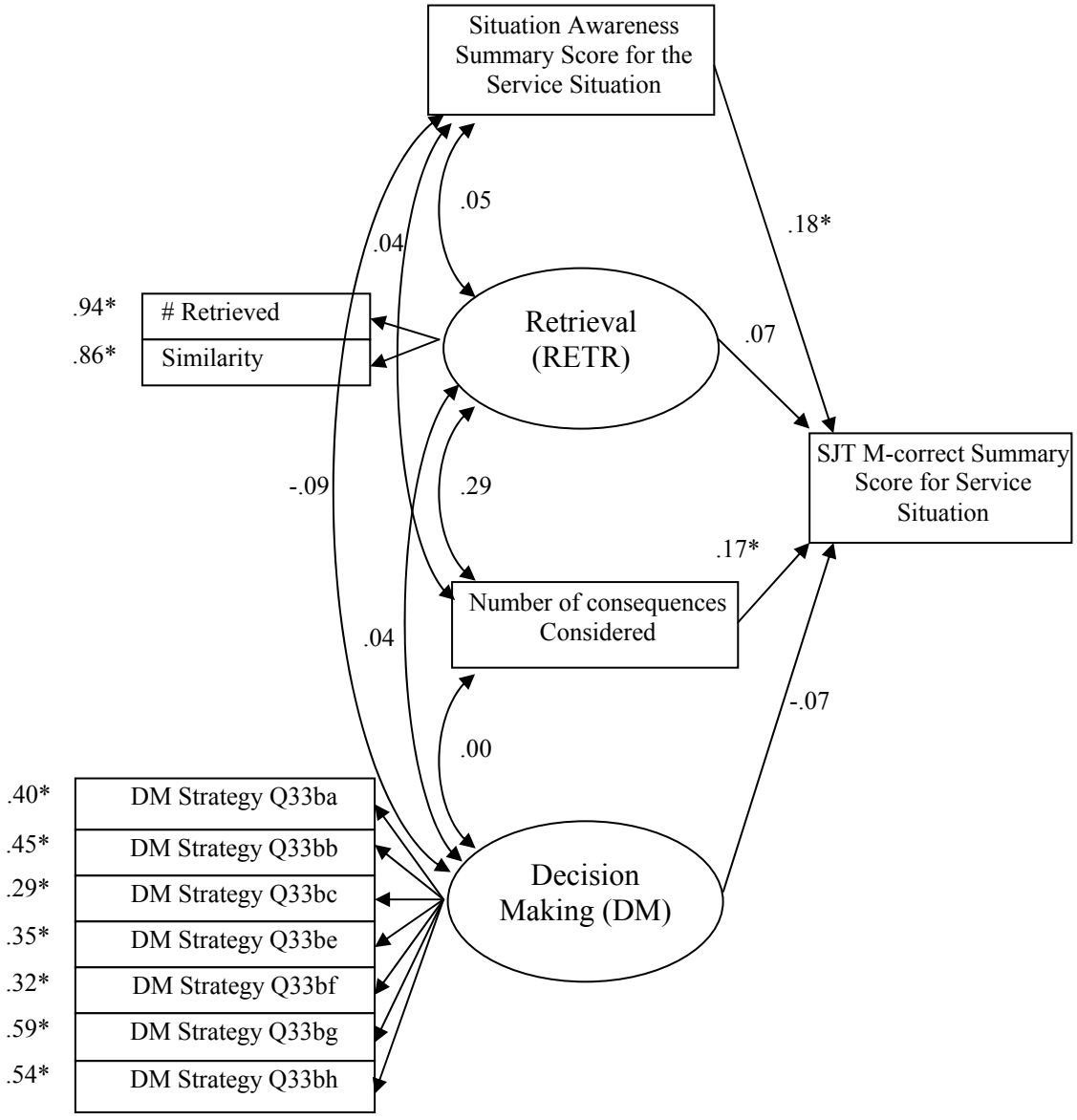

Figure 7. $SvMM_m$: Service Situation Structural Model Using M-correct SJT Score
Note: Data points listed directly beside the manifest variable boxes are standardized factor loadings.

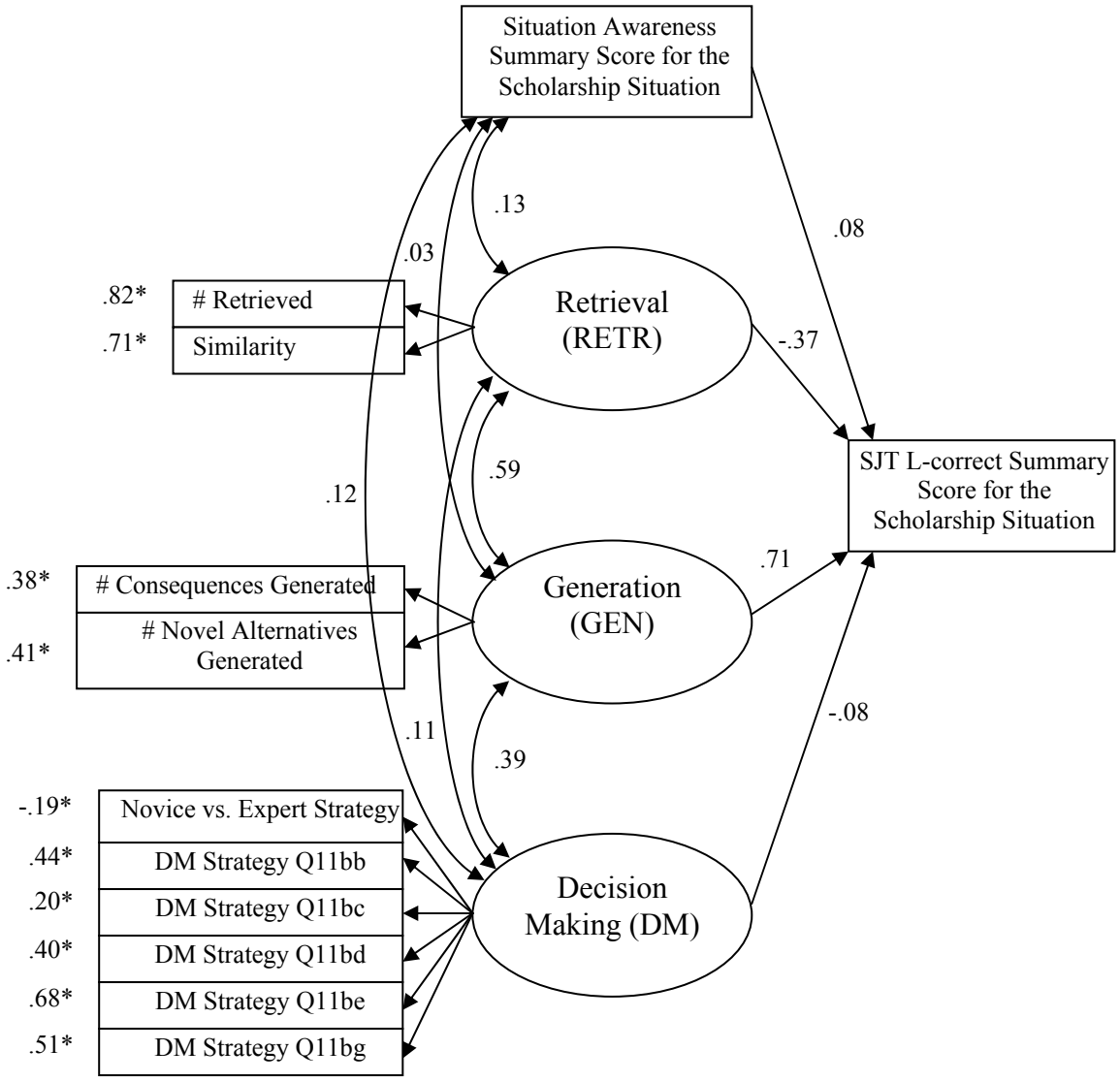

Figure 8. $ScLM_m$: Scholarship Situation Structural Model Using L-correct SJT Score
Note: Data points listed directly beside the manifest variable boxes are standardized factor loadings.

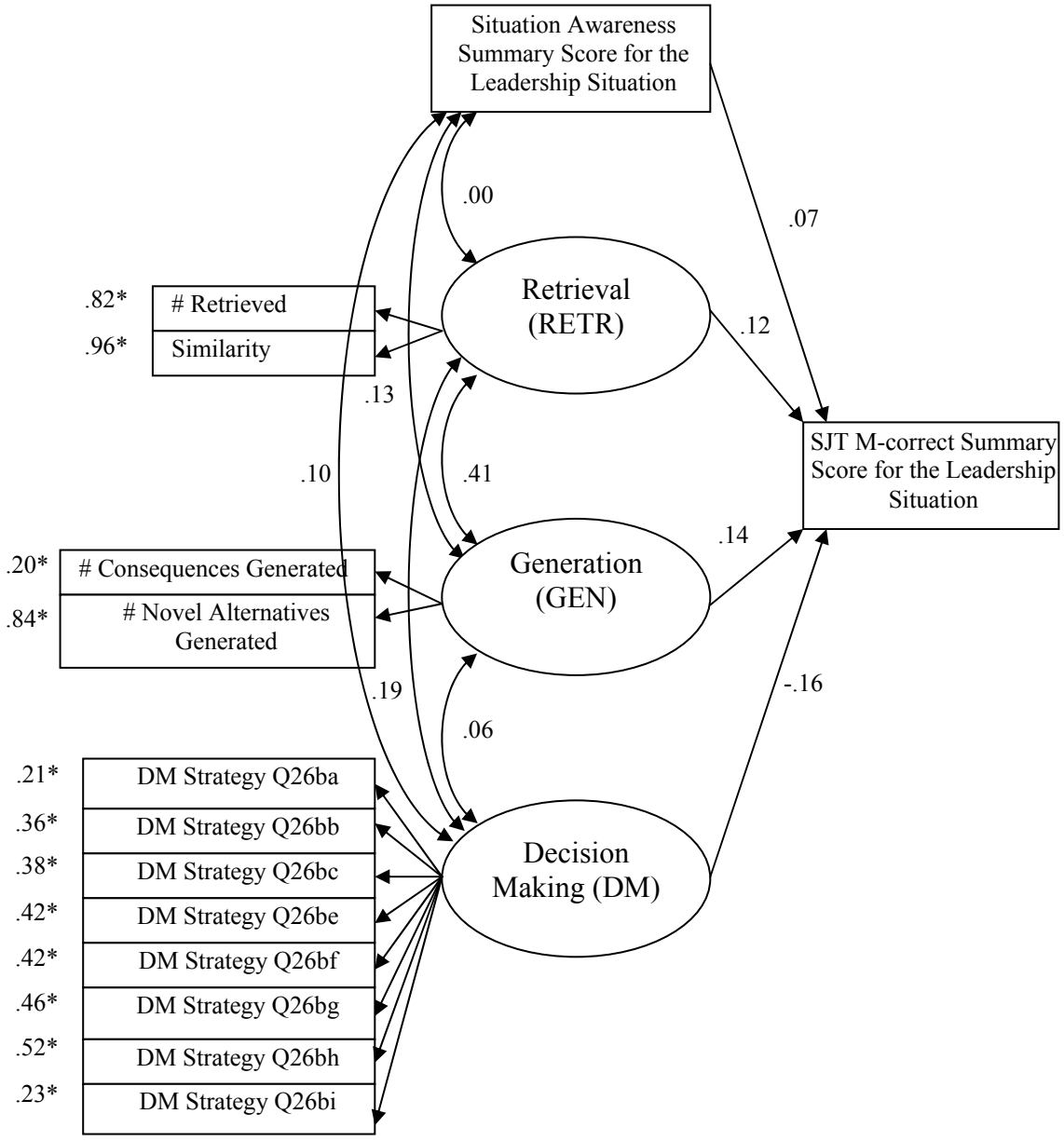

Figure 9. $LdLM_m$: Leadership Situation Structural Model Using L-correct SJT Score
Note: Data points listed directly beside the manifest variable boxes are standardized factor loadings.

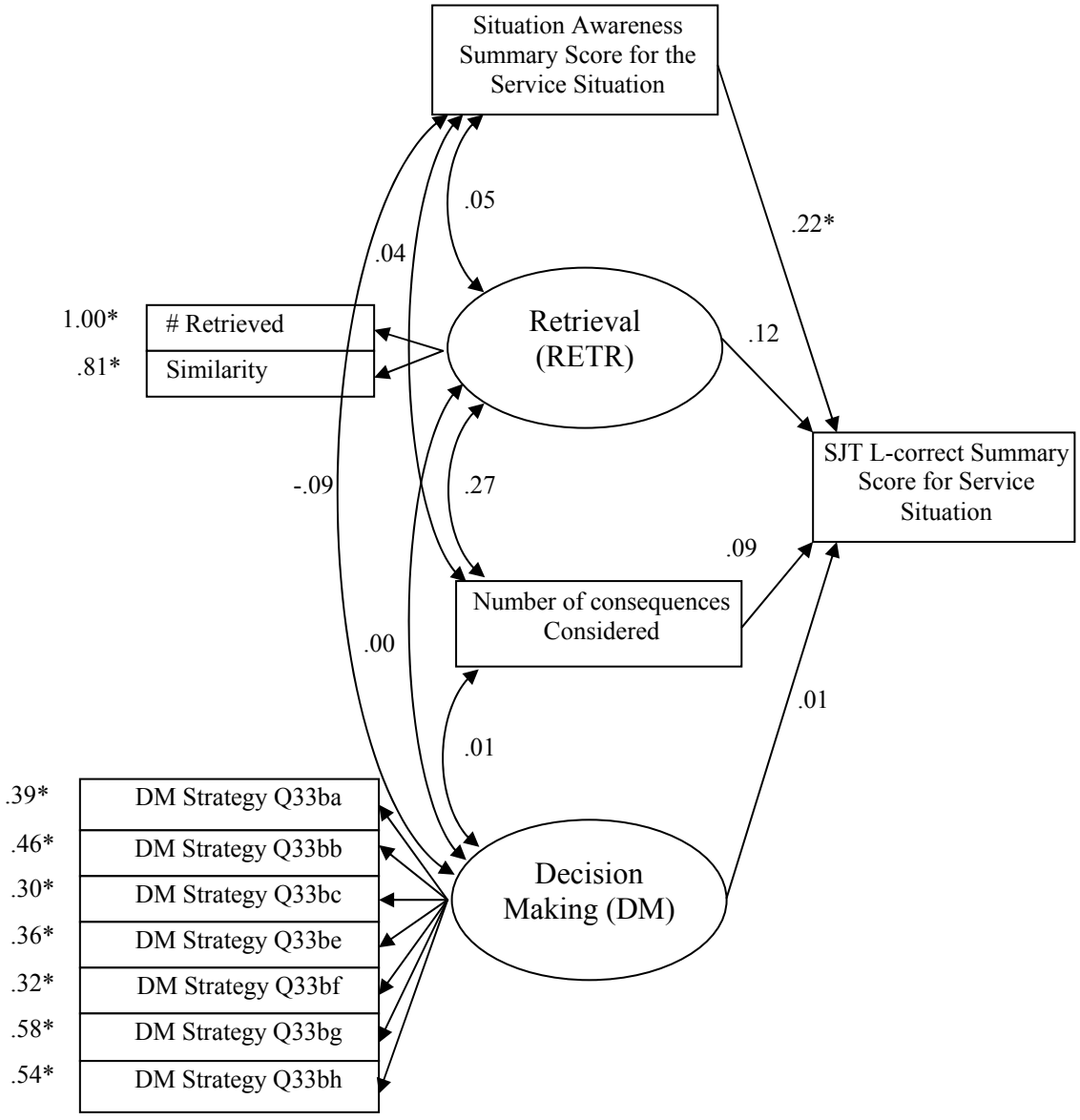

Figure 10. $SvLM_m$: Service Situation Structural Model Using L-correct SJT score
Note: Data points listed directly beside the manifest variable boxes are standardized factor loadings.

shows the final cognitive processing structural model to be addressed: the model for the Service situation using the L-correct SJT score. All of the factor loadings are significant, but five out of the nine are less than .50. The path coefficient for SA of .22 is the only predictor with a significant relationship with situational judgment.

In sum, the model fit indices for the cognitive processing models vary considerably across situation and scoring method, but are only moderate. While three of the cognitive processing factors were significant predictors of SJ in at least one model (SA, RETR, and GEN), the pattern of factor loadings and path coefficients for the structural models fail to provide consistent support for the cognitive process model of situational judgment.

Integrated Model of Performance

Overview. The integrated model of performance is analyzed in two phases. Tetrad analyses and CFAs are conducted in phase one in an effort to purify the constructs. The structural paths are examined in phase two. Several post hoc analyses are also conducted and reported here. Item descriptions and descriptive statistics for the integrated model of performance variables can be found in Appendix H. The correlation matrix for the integrated model of performance variables is presented in Appendix I.

Phase One. Phase one began with a tetrad analysis of all the factors in the model with more than four manifest variables. TETRAD II© suggested the following reductions: eliminate Q26j, Q33j, and Q29 from Experience (EXP), eliminate C5SelfD from the Conscientiousness (CON) factor, eliminate A2Stra and A5Mod from the Agreeableness (AGR) factor, eliminate N4SelfC and N5Imp from the Need for Stability (NS) factor, and

finally, eliminate R4l2 and R5sv1 from the Program Specific Performance (PSP) factor. After eliminating these manifest variables, all the tetrad equations implied by the resulting models passed the Bonferroni test. The reduced set of variables was used in subsequent analyses. Note that although the model as proposed included rank variables, clean data were unavailable for these variables and hence they were not included in the analyses of the integrated model. In addition, shrinkage for the peer ratings was sufficient to require it to be dropped from the analyses.

The next step was to conduct the measurement model analyses via CFA. Summary goodness of fit indices are reported in Table 11 for the predictor side measurement model (M_{mp}), the revised predictor side measurement model (M_{mp}-Rev), the criterion side measurement model (M_{mc}), the revised criterion measurement model (M_{mc}-Rev), the full measurement model (M_m), and the revised full measurement model (M_m-Rev).

First, a CFA was conducted on the predictor side measurement model to test how well the remaining manifest variables represent the factors they are purported to measure. The results suggest a poor fitting model (see Table 11). The GFI, NNFI, and NFI are particularly small at .82, .67, and .76 respectively. A review of the modification indices suggests that the personality inventory manifest variables are problematic. Eight of the ten largest Lagrange multipliers involve adding paths from various personality manifest variables to other personality factors. Based on the recommendations of Hatcher (1994), the eight identified cross loaders were eliminated. The CFA on this further reduced model resulted in an improper solution. Furthermore, the Lagrange multiplier suggested

Table 11

Goodness of Fit Indices for the Measurement Models for the Integrated Model of Performance

	FF	GFI	χ^2	*df*	*p*	RMSEA	NNFI	NFI
M_{mp}	2.60	.82	451.77	220	<.0001	.08	.67	.76
M_{mp}-Rev	.55	.93	96.6	74	.0400	.04	.95	.85
M_{mc}*	.51	.91	88.61	32	< .0001	.10	.93	.92
M_{mc}-Rev	.13	.96	23.23	13	.0390	.07	.98	.97
M_m	2.49	.86	436.29	362	.0044	.03	.94	.76
M_m-Rev	1.30	.89	227.97	175	.0044	.04	.95	.85

M_{mp} = Predictor measurement model;
M_{mp}-Rev = Revised predictor measurement model;
M_{mc} = Criterion measurement model, *model had one improper solution;
M_{mc}-Rev = Revised criterion measurement model;
M_m = Full measurement model;
M_m-Rev = Revised full measurement model;
FF = Fit Function; GFI = Goodness of Fit Index; χ^2 = Chi Square; *df* = degrees of freedom, *p* = probability; RMSEA = Root Mean Square Error of Estimation; NNFI = Non-Normed Fit Index; NFI = Normed Fit Index;
* = This model demonstrated one improper solution: TGPA had a factor loading of 1.05. The model is not rejected because this value is within one standard error of 1.0.
** = This model also demonstrated one improper solution: TGPA has a factor loading of 1.01, but the model is not rejected because this value is within one standard error of 1.0.

adding still more paths from the personality manifest variables to other factors. Based on the inability to clean up the personality factors and based on the fact that the research literature offers the most consistent support for the use of Conscientiousness in predicting performance, the AGR and NS factors were eliminated from the predictor side of the model. Running the CFA without these two personality factors yielded an acceptable model, and after dropping the Class variable that had a very low factor loading of .09, the

model fit improved yet again to produce a good fitting model. This model is the final predictor side measurement model (see Table 11).

The next step in phase one was to run a CFA on the criterion side of the model using the TETRAD II© purified factors. The model fit was acceptable (see Table 11). However, this analysis yielded an improper solution: a factor loading greater than 1.0 for TGPA. In addition, the correlation between PSP and CP factors was .92, and the largest Lagrange multipliers were added paths for the supervisor ratings between the PSP and CP factors. Given the cross loaders and the high correlation between the factors, it was decided to merge the two factors together into a new factor. This new factor was then purified via TETRAD II© and on the basis of that analysis, supervisor ratings R1s1, R3l1, R5sv1, and R6c1 were eliminated. A rational analysis of the remaining manifest variables for this factor (supervisor ratings of critical thinking skills, seeking or accepting leadership roles, integrity, adaptability/resiliency, and presence) suggest that it can best be identified as nonacademic performance (NAP). The reduced set of variables for this new factor was used in the subsequent CFA analysis of the criterion side measurement model. Model fit improved, and after dropping CSHr, which had a low factor loading (.20), the model fit was improved yet again. This model was accepted as the final measurement model on the criterion side (see Table 11).

The final step in phase one was to conduct a CFA on the full measurement model. The revised model consists of three predictors (g, EXP, and CON), one partial mediator (SJ), operationalized by the summary scores by situation and by scoring method, and two criterion factors (AP and NAP) that are all allowed to covary. The resulting model fit

was acceptable according to most indices. For example, the NNFI = .94 and the RMSEA = .03. However, although the value of the Fit Function for Maximum Likelihood Estimation technically has no upper bound, the value of 2.49 is large enough in comparison to the other model estimates to indicate a problem with model fit. The pattern of factor loadings for the SJ factor also indicated problems; all of the standardized factor loadings for the SJ factor were less than .50, and two of the eight scores failed to reach significance. Based on these issues and on the general failure of SJT items to support a clear factor structure (Hanson, Horgen, & Borman, 1998), it was decided to run the full measurement model using the overall SJT score as a single manifest variable for situational judgment. This modification resulted in an improvement in model fit, however, the Q38 variable had a low factor loading of .18. After dropping this variable, the model fit was good and it was accepted as the revised and final measurement model (see Table 11).

Phase Two. Phase two of the analysis involved analyzing the revised integrated model of performance to examine the structural relationship among the factors. This model is presented in Figure 11. Table 12 summarizes the goodness of fit indices for the initial integrated model of performance (M_i) and the revised integrated model of performance (M_{rev}). The data do not fit the model as hypothesized; the SEM results in two improper solutions and diagonal residuals too high to compute test statistics. However, the initial model as depicted in Figure 11 is quite stringent in that none of the predictors are allowed to covary. Allowing the predictors to covary did improve model fit. The improper solutions and diagonal residual problems were eliminated allowing the

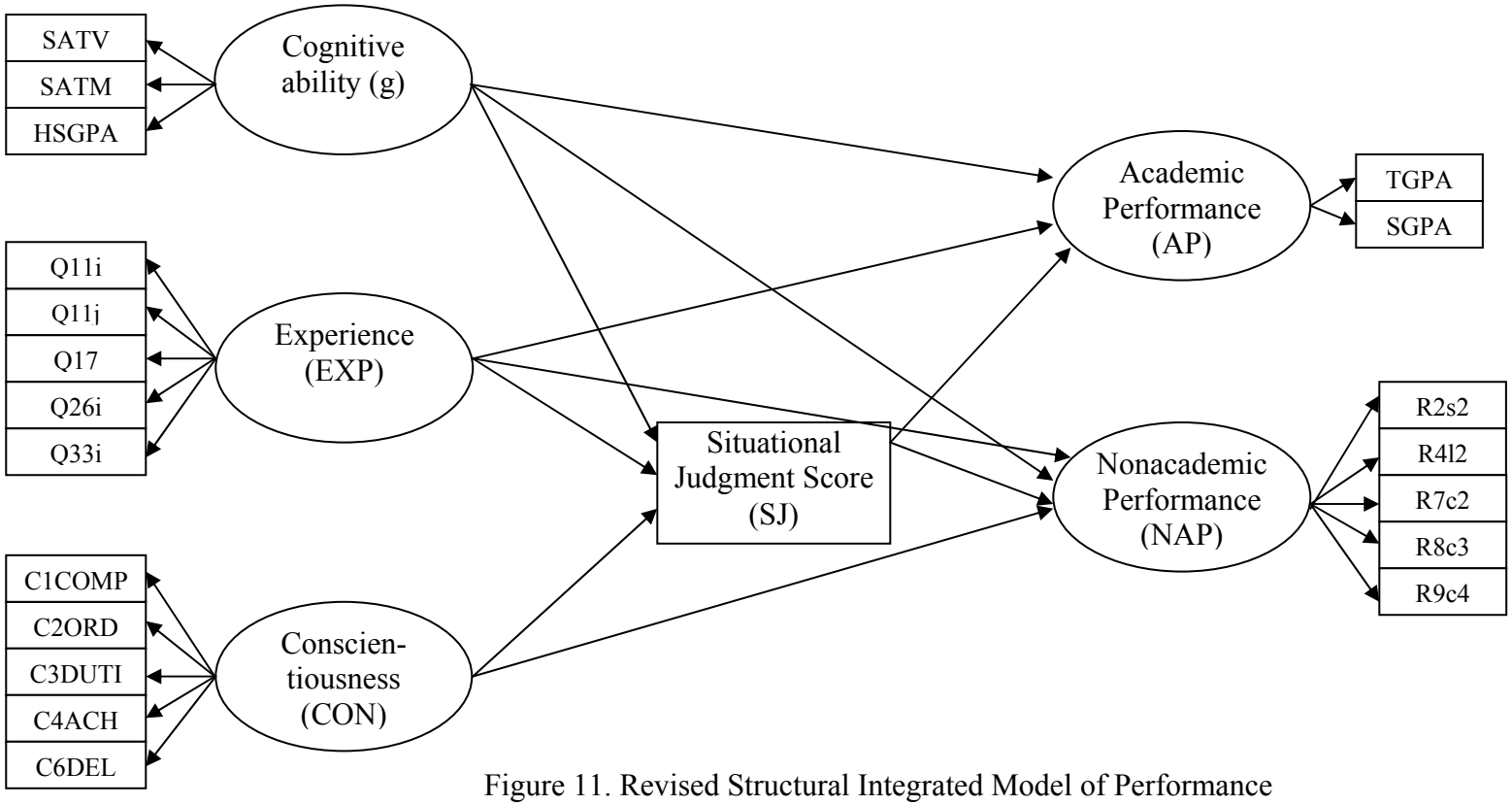

Figure 11. Revised Structural Integrated Model of Performance

Table 12

Goodness of Fit Indices for the Structural Integrated Model of Performance

	FF	GFI	χ^2	*df*	*p*	RMSEA	NNFI	NFI
M_i	Two improper solutions, diagonal residuals too high.							
M_{rev}	1.44	.88	252.38	177	.0002	.05	.93	.84

M_i = Initial Integrated Model of Performance;
M_{rev} = Revised Integrated Model of Performance; identical to initial model except F1-F3 allowed to covary;
FF = Fit Function; GFI = Goodness of Fit Index; χ^2 = Chi Square; *df* = degrees of freedom, *p* = probability; RMSEA = Root Mean Square Error of Estimation; NNFI = Non-Normed Fit Index; NFI = Normed Fit Index;

various model fit indices to be calculated. The result was a moderately acceptable model. The modification indices suggested allowing the disturbance terms on the criterion side to correlate. However, allowing this modification resulted in one negative eigenvalue and diagonal residuals too high to compute statistics. Several other models were run based on suggestions from the modification indices. However, no modification that was appropriate significantly or consistently improved model fit. Consequently, this was accepted as the final structural model. Factor loadings and path coefficients will be discussed.

The standardized factor loadings and path coefficients for the final structural model are shown in Figure 12. All the factor loadings are significant at $p < .01$, as evidenced by double asterisks. However, only four of the causal paths are significant. Specifically, the paths from *g* to AP and from *g* to NAP were both significant at $p < .01$

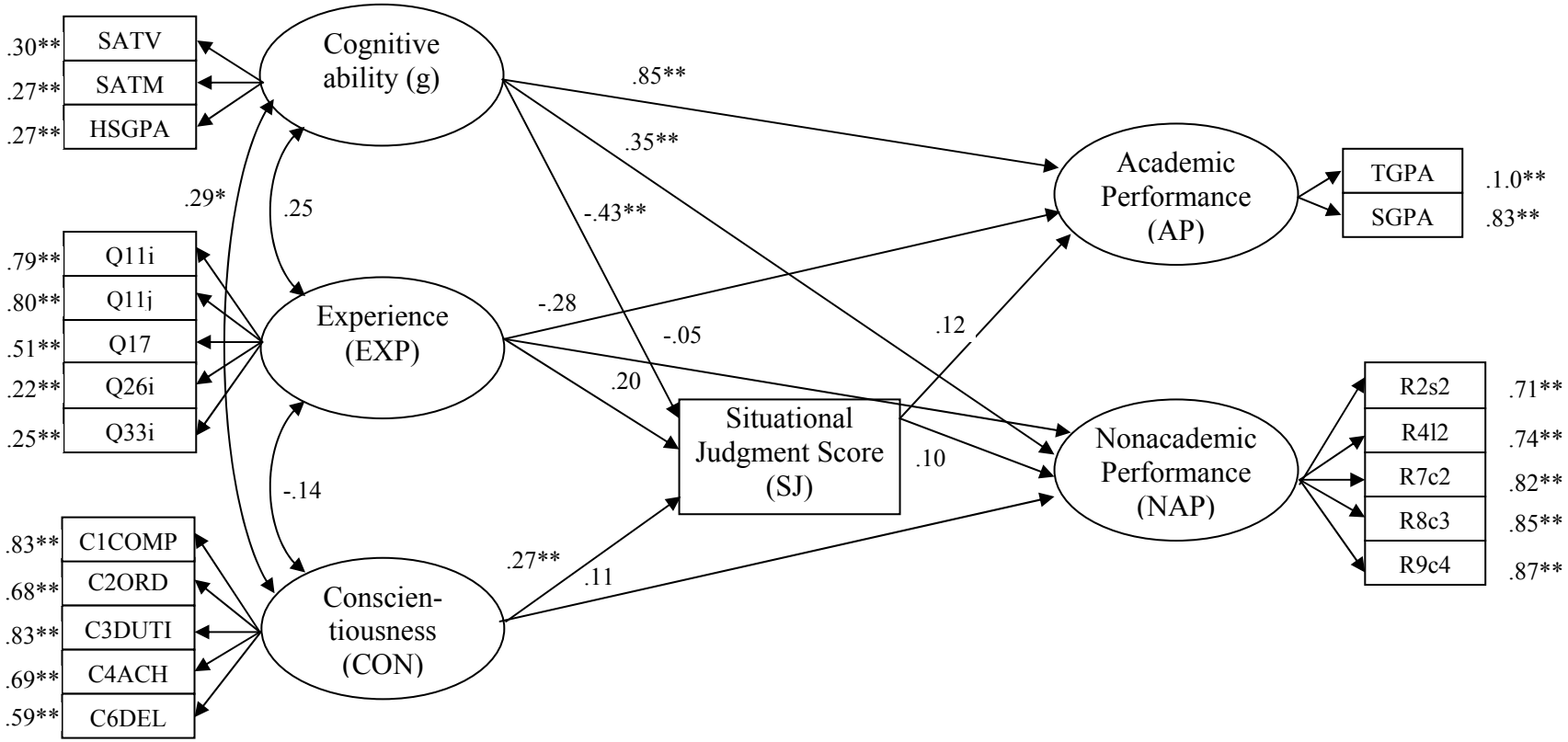

Figure 12. Final Structural Integrated Model of Performance with Standardized Factor Loadings and Path Coefficients
Note: Data points listed directly beside the manifest variable boxes are standardized factor loadings.

and practically large at .85 and .35. The path from *g* to SJ was significant at $p < .01$ and large at -.43. The path from CON to SJ of .27 is practically and statistically significant at $p < .01$. The overall SJT score was not predictive of either of the criterion factors. However, dropping the SJT score altogether did not significantly improve model fit and consequently it was retained.

Post Hoc Analyses on the Integrated Model of Performance. Given the difficulties experienced with the integrated model of performance, several post hoc analyses were conducted. The problems with the personality factors during the measurement model CFAs were unexpected. In order to gain a more thorough understanding of the results, an EFA was conducted on the 18 manifest personality variables that were purported to measure Conscientiousness, Agreeableness, and Need for Stability. The results support a four factor structure instead of the three-factor structure as hypothesized (see Appendix L).

The data also did not support a three factor structure on the criterion side (AP, PSP, and CP). One possible reason for this is that the supervisor ratings of performance do not represent the factors they were proposed to represent. To examine this possibility further, an EFA was conducted on the supervisor ratings of Scholarship, Leadership, Service, and Character to see if a four factor structure would be supported. The data do not support a four factor structure; the EFA retains only two factors (see Appendix M).

The systems data (predictors and criteria) that were collected were unable to withstand the rigors of SEM. Consequently, a third post hoc analysis involved running a series of regression analyses modeling the various performance criteria on the overall SJT

scores. Results of these analyses suggest that the overall SJT score is not predictive of any of the relevant criterion variables including TGPA, SGPA, or the ten supervisor ratings of performance.

Finally, in an effort to further determine the nature of the SJT, a post hoc analysis comparing the performance of the scholars for whom the SJT was developed to the performance of a group of 315 undergraduate students was conducted. The undergraduate students were administered the SJT[2] under the same standardized conditions. Evidence of the SJTs construct validity would be established if the 315 undergraduate students performed more poorly than the scholarship recipients. In support of this notion, the 211 scholarship recipients scored higher on the SJT (M = 47.32, SD = 9.86) than the 319 undergraduate students (M = 36.76, SD = 9.17). A t test was conducted to determine if these means are significantly different. Examination of the pooled test for equal variances suggests that the difference in means is significant (t = 12.59, $p < .0001$).

Discussion

The data from the present research fail to support the cognitive processing model or the integrated model of performance as originally hypothesized. Significant modification of the hypothesized models did result in moderate fitting models. However, the SEMs yielded generally unimpressive results for the causal paths, particularly for the cognitive processing models. Many of the proposed structural paths in the cognitive processing models were either not significant or were in the opposite direction than hypothesized. The most promising variable is the Situation Awareness score, which is

predictive of SJT score in three of the six models. The specific processing model hypotheses will now be addressed.

H1 proposed that higher SA scores would be associated with higher SJT scores. This hypothesis is supported in three of the six models: the Scholarship situation using the M-correct SJT summary score, the Service situation using the M-correct SJT summary score, and the Service situation using the L-correct SJT summary score.

H2 proposed that retrieval and experience would be associated with higher SJT scores. The RETR factor did have a significant path coefficient in the Leadership situation using the M-correct score. However, the path coefficients for RETR varied wildly in magnitude across the other five models, and in all five models failed to reach significance due to high standard errors.

H3 stated that no hypothesis could be made regarding the generation of alternatives and consequences. The GEN factor suffered significant modification in several of the models, but was a significant predictor of the M-correct score for the Service situation. The path coefficients for GEN for the other models, while practically significant in some cases, were not statistically significant due to high standard errors.

H4 proposed that more experienced participants and those who retrieved a greater number and more similar situations from memory would engage in more recognition based or expert processing. The correlation matrix between the variables (see Appendix J) shows that the correlations vary from situation to situation. However, in opposition to H4, none of the correlations between experience (Class, Q17, Q29, Q38, Q11i, Q11j, Q26i, Q26j, Q33i, Q33j) and expert processing (Q11c, Q26c, and Q33c) were found to be

significant. H4 also proposes that higher levels of the recognition based processing questions (Q11c, Q26c, and Q33c) will be associated with higher SJT scores. The correlations between the recognition based processing variables and SJT summary and overall scores were again found to vary from situation to situation (see Appendix K). However, only two of the correlations in the matrix are significant, and they are in the negative direction. In sum, both facets of H4 fail to be supported.

Finally, H5 proposes that the overall fit of the cognitive process model of situational judgment will vary by situation and by scoring method. This hypothesis is supported in that each of the six models did yield to CFA in different ways with different sets of manifest variables being retained. In addition, the models do have different fit, factor loading, and path coefficient patterns. However, given the mediocrity of the indices and the lack of consistent significance in the path coefficients, it is inappropriate at this juncture to suggest that H5 is supported.

In sum, the data provide only partial and weak support for the cognitive processing model. The model fit indices are not particularly high. The factor loadings, while significant, are of modest magnitude. Finally, most of the path coefficients are not significant or are of small magnitude. The data generally fail to support the cognitive processing model of situational judgment.

The integrated model of performance faired a bit better and did produce four significant path coefficients. The magnitudes of the path coefficients for *g* confirm what logic would propose; that is, the largest coefficient is from g to academic performance, and the lower but still significant relationship is with nonacademic performance. In

addition, the practically and statistically significant relationship between CON and SJT of .27 suggests that as proposed, SJ is in this environment to some degree a function of conscientiousness. However, many of the proposed structural paths were either not significant or were in the opposite direction than hypothesized. These results will now be discussed in relation to the originally proposed hypotheses.

This research proposed several specific hypotheses regarding the causal paths in the integrated model of performance. However, the modifications made to both the measurement and structural portions of this model require that various hypotheses be ignored. Specifically, although six hypotheses (H6 – H11) were proposed for the integrated model of performance, removal of the AGR and NS factors, and the collapsing of the PSP and CP factors into a new NAP factor prevent H8, H10, and H11 from being examined. H6 proposed that SJT scores would be positively related to AP. The data do not support this hypothesis. In fact, SJT scores were not predictive of either of the criterion factors.

H7 proposed that *g,* experience (EXP), agreeableness (AGR), and conscientiousness (CON) will be positively related to overall scores on the SJT. AGR was dropped from the model; however, the path coefficient of -.43 from g to SJT indicates a relationship between g and SJT, but in the direction opposite to what was proposed. This relationship may be merely an artifact of the way *g* and SJ were operationalized in the current research. That is, *g* is operationalized as high school grades and SAT scores. However, SJ is operationalized to capture four dimensions of scholar performance (Scholarship, Leadership, Service, and Character). In fact, for several SJT

items, the keyed answer is not necessarily the option that will lead to the highest grades. For example, one SJT item presents a situation where a scholar has a great deal of interest in a particular class but is performing poorly. The keyed answer according to the SMEs is to continue pursuing the class as this option demonstrates intellectual curiosity and intellectual perseverance. However, this option would not necessarily result in the highest GPA. In sum, perhaps scholars who have high grades and high SAT scores focus on choosing SJT response alternatives that will most likely result in high academic performance, but not necessarily on choosing the alternatives that will make them good scholars overall. To the extent that this is true, the scholarship program should use caution in weighting high school academic performance too highly in the selection process. Finally, as previously stated, the data do support a significant relationship between CON and overall SJT score, but they do not support a relationship between EXP and SJT score.

Finally, H9 proposed that *g* will be associated with SJT scores, but will be more highly associated with SJT scores for situations that were administered under timed conditions than under untimed conditions. This hypothesis cannot be tested as stated via SEM, as modifications and other analyses suggest the use of a single overall SJT score as opposed to situation summary scores. However, it is possible to compare the current SEM results using the overall SJT summary score that includes all the items on the SJT with part of situation one timed (SJT4) with an SEM that uses a different SJT summary score (SJT47) that includes only the items from situation one that the participant completed in seven minutes plus all the items from situations two and three. This allows

for a more direct comparison of the path coefficients. However, this SEM fails to reach optimization. H9 cannot be confirmed or rejected.

In sum, both the measurement model and the structural model for the integrated model of performance required substantial modification before finding acceptable fit. The data do not generally support the causal paths proposed in the cognitive processing model and only partially support the paths proposed for the integrated model of performance.

Post hoc analyses shed some light on why the results were not as expected. Specifically, phase one analyses for the integrated model of performance revealed that the factor structures of both the predictor and criterion sides were not supported. Specifically, the three-factor NEO PI-R™ structure (Conscientiousness, Agreeableness, and Need for Stability) on the predictor side was not supported, and the three factor structure (Academic Performance, Program Specific Performance, and Contextual Performance) on the criterion side was not supported. Possible reasons for these difficulties include the nature of situational judgment, the unique qualities of the sample, and specific limitations of the current research. These issues will now be addressed.

The Nature of Situational Judgment

Situational judgment is admittedly a complex concept. The introduction of this dissertation highlights the considerably different opinions of accomplished situational judgment researchers regarding the nature of situational judgment. It has been proposed that situational judgment is a function of general cognitive ability, experience, personality, or is a new judgment construct, or some combination. Still others consider

SJTs to be a measurement method. Finally, the decision making literature suggests that decision making is influenced by so many situational variables as to make predicting it extremely difficult. In fact, it was this lack of agreement among the experts that served as the impetus for the current research. However, this lack of agreement also speaks to a central notion: the development of a reliable and valid SJT is not an easy task. The present research confirms this difficulty.

The SJT developed for this research was found to have nonsignificant correlations with both the criterion factors. This suggests several possibilities. First, it is possible that the SJT developed for this research simply does not capture the germane incidents and behaviors for this population. However, given the concerted effort to prevent this via a lengthy criterion development phase and SME involvement at every step makes this seem unlikely.

Second, it is possible that the manifest variables for the criterion factors, and most importantly the ratings of performance, do not adequately capture the underlying performance dimensions. Given that the supervisor ratings of performance do not load on the four performance dimensions they were designed to capture suggests possible problems with the ratings. This is a common difficulty in performance research.

Third, and most importantly in my opinion, it is possible that the situations experienced in this context are complex and vague to the point of making it incapable of being captured. In particular, the consequence of error in the scholar environment is substantially less than for other populations for whom SJTs have been developed (e.g., correctional officers). In addition, while the scholarship program does have specific

performance goals, there are no set policies or procedures regarding how the scholars are to arrive at these performance goals; there is no Standard Operating Procedures guide for being a scholar in this environment. This again, is quite unlike jobs where the performance goals are clearly specified but so are the policies and procedures necessary to arrive at those goals. Perhaps the SJT failed in this environment because there are too few direct consequences of error and too many possible paths to arrive at the desired performance.

Finally, related to this notion is the idea that this sample of scholars is capable of considering situations with a great deal of intellectual "horsepower." It is possible that the inability of this SJT to predict or be predicted is due to their ability to discern fine shades of meaning and to consider more possible solutions to arriving at the same goals.

The Nature of the Sample

The participants in the current study are recipients of a competitive university scholarship, and hence do not represent the average population. Furthermore, they do not represent the average undergraduate student population. These scholars are unique individuals who routinely demonstrate extremely high levels of ability and performance (e.g., very high SAT scores and grades) when compared to the general population. They are likely considerably different in terms of their experiences, and perhaps even their personality. This departure from the "average" and the concomitant restriction of range is evident in tests of univariate and multivariate normality. The majority of the variables when considered alone in a univariate analysis are not normally distributed. When grouped together by factor, none of the groups demonstrate multivariate normality. A

notable exception is the SJT scores, which do demonstrate univariate and multivariate normality.

The lack of normality in the variables causes several difficulties. First, many statistical techniques technically require normally distributed data. Although it can be argued that minor departures from normality are inconsequential, this is still a consideration for the data analyses. A second difficulty is that the proposed factor structures of various concepts included in the models are based on the general population. For example, the NEO PI-R™ was developed and normed using adults and college students from the general population (Costa & McCrae, 1992). While meta-analytic studies support the notion of a five factor structure of personality, data from this population fail to support this structure. Perhaps the five factor structure of personality holds only for a more "average" population. To the extent that this is true, researchers working with unique populations should exercise caution when proposing or assuming factorial structures. In sum, it is proposed that it is the unique nature of this sample that is responsible for at least some of the unexpected results.

Limitations of the Current Study

Finally, the current study has specific conceptual and methodological limitations. From a conceptual standpoint, the current research sought to identify and operationalize the cognitive processing steps involved in answering an SJT item. While there is a significant literature associated with cognitive processing and decision making, situational judgment was not found to have been examined in this way previously. Limitations therefore possibly include the selection of the processing factors as well as

their operationalization. Due to the incipience of this conceptualization and the concomitant paucity of research on the subject, it is not surprising that problems were encountered. It is this incipience, however, that makes it inappropriate to conclude that a processing approach is incapable of capturing the essence of situational judgment.

Length of time required for administration of the SJT was a considerable concern during its development. This concern, coupled with the inclusion of the 27 cognitive processing questions to the SJT, meant that only a small number of actual SJT items could be written and included. For example, while there were five items written to capture the Scholarship dimension, only two items were written to capture critical thinking skills and two items were written to capture effective management of potential. It is possible that there were too few SJT items to capture the complexity of the underlying concepts.

From a methodological standpoint, of concern are the ratings. The post hoc analysis on the supervisor ratings suggests that the ratings were not factorially pure and did not capture the underlying performance dimension they were designed to measure. The ubiquity of this problem in personnel research suggests that raters generally have difficulty discerning among multiple components of performance. However, in this particular research, it is also possible that the problem is being caused by the very recent development of the criterion behaviors list. Although the scholarship program was founded in 1995 on the four performance dimensions used in this research (Scholarship, Leadership, Service, and Character), the list of criterion behaviors that comprise those dimensions was not defined until just before the SJT was developed. Perhaps not enough

time had passed for the SMEs to incorporate the full list of criterion behaviors into their memory of critical incidents and responses for these scholars. This could lead to inaccurate ratings, inaccurate judgments regarding the appropriateness of the SJT items and response alternatives, or both.

Another methodological weakness is the result of the nature of the SME group itself. Given the nature of their work (many of the SMEs are professors or university administrators), it was extremely difficult to get them together for meetings. This created two difficulties. First, much of their contribution was made at the individual level in one-on-one meetings with this researcher. It is possible that the development of valid SJT items is more appropriately the result of the iterative process by which SMEs contemplate and develop the items and response alternatives through discussion. To the extent that this is true, the items would have been improved if the SMEs had collaborated in a group setting. Second, due to time constraints and the university schedule, the full group of SMEs was not called upon to generate critical incidents, but instead was only asked to critique the SJT item stems after they had been written by the author and reviewed by a smaller group of SMEs. Again, to the extent that the involvement of the SMEs was limited, it is likely that the quality of the SJT suffered.

Finally, in an effort to increase the fidelity of the situations, the SJT employs natural temporal sequencing of items. While this technique likely had the desired effect, it also created two significant difficulties. First, items in a sequence can be written to either proceed based on the correct answer to the previous question, the incorrect answer to the previous question, or based on new information. For example, if the correct

answer for a question was to seek advice from one's mentor and the incorrect answer was to seek advice from a roommate, then the following question can either proceed based on the main character having sought advice from the mentor, having sought advice from the roommate, or on the basis of new information presented. A difficulty in writing a temporally sequenced SJT is that the items cannot be written with a consistency that reveals the keyed answer to the previous question. That is, a following question cannot always be a function of the correct or incorrect answer from the previous question. While it is felt that the present research was successful in avoiding this pitfall, it is nonetheless a consideration as it required significant time and effort to avoid.

A second difficulty was that the temporal sequencing meant that the items were tied together in story fashion, therefore making it impossible to eliminate single items without having to rewrite subsequent items. This is of particular concern given that it is the usual case that only one out of five SJT items generated will be retained for use (W. C. Borman, December, 2002, personal communication). The current research resolved this dilemma by simply editing items until adequate SME agreement was reached. However, this required a great deal of time and effort that may have been better spent generating more appropriate items.

Future Research

While the data do not support the hypothesized cognitive information processing model of situational judgment, it is difficult to draw any real conclusions from this as the SJT itself performed so contrary to what was expected. It is possible that a cognitive

processing approach is appropriate for studying situational judgment and that the particulars of the current SJT and the current sample made it difficult to see a clear pattern in those relationships. It is hoped that the failure of the current research not be accompanied by an abandonment of the underlying concept. It is the opinion of this researcher that great progress can be made in the area of situational judgment by considering it as an information processing task. One way that future research may improve on the current research would be to examine the cognitive processes with an SJT that has been validated for use with a particular population.

Another technique that could assist researchers in their examination of a cognitive processing approach to situational judgment would be to test participants individually and ask them to verbally report their thought processes as they read and answer an SJT item. This technique would perhaps allow for a more complete or accurate conceptualization of the factors involved in the process and how best to operationalize them. Finally, the time line of the current project did not allow for computer administration of the SJT. However, a great deal of relevant data could be gathered this way. For example, computer administration would allow one to accurately measure the amount of time each participate engaged in reading the item stem, reading each of the response alternatives, and how many times (if any) she or he returned to the item stem after having viewed the response alternatives. These questions would assist one in operationalizing the factors more accurately.

With regard to the integrated model of performance, several possibilities should be explored. First, future research in this area should continue to strive for excellence in

the area of performance ratings. In particular, the criterion development phase should be completed as far in advance of any SJT development as possible. In addition, raters should be trained more fully.

Second, the SJT was found not to be predictive of the criterion factors in the current research. As previously discussed, this is perhaps due to two things: the lack of clear and immediate consequence of error for this population, and due to the lack of a clear and accepted methods or procedures for arriving at maximum performance. The previously discussed meta-analysis by McDaniel at al. (2001) suggests that the mean population validity coefficient for SJTs is .34. However, there is great variability; the introduction of this paper found validity coefficients ranging from 0 to .56. While the meta-analysis did consider several factors that might be responsible for the variability in the coefficients, the type of job for which the SJT was developed was not identified or considered in the analysis. Future research should examine whether the type of job, and specifically the consequence for error and the amount of standardization in procedures and policies, has an influence on the upper limit of the validity of the SJT.

Conclusion

The data generally do not support cognitive processing model or the integrated model of performance. However, the results contribute to the existing research on situational judgment in several ways. First, this research provides an example of the temporal sequencing technique in SJT development. Second, it provides support for the idea that SJ is to some degree a function of personality. Third, it provides the first known

empirical attempt at conceptualizing answering an SJT item as a cognitive information processing task. Finally, it generates further questions that can be used to explore important issues. In particular and perhaps most importantly, is the possibility that the difference in the validities of SJTs is a function of the consequence of error and whether policies and procedures for achieving maximum performance are present and well understood in the context under investigation. It is the hope of this researcher that these questions will be addressed in future research.

References

American Educational Research Association, America Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Arvey, R. D. (1986). General ability in employment: A discussion. *Journal of Vocational Behavior, 29*, 415-420.

Ashcraft, M. H. (1994). *Human memory and cognition* (Second Ed.). Harper Collins College Publishers.

Austin, J. T. & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77(6)*, 836-874.

Barrick, M. R. & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Borman, W. C. (December, 2002). Personal communication.

Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel selection. *Annual Review of Psychology, 48*, 299-337.

Borman, W. C., Hanson, M. A., Oppler, S. H., Pulakos, E. D., & White, L. A. (1993). Role of early supervisory experience in supervisory performance. *Journal of Applied Psychology, 78(3)*, 443-449.

Borman, W. C. & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt and W. Borman (Eds.), *Personnel selection in organizations* (pp. 71-98). San Francisco: Jossey-Bass

Publishers.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology: Vol. 1* (2nd ed., pp. 687-732). Palo Alto: Consulting Psychologists Press, Inc.

Cannon-Bowers, J. A., Salas, E., & Pruitt, J. S. (1996). Establishing the boundaries of a paradigm for decision-making research. *Human Factors, 38(2)*, 193-205.

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(1)*, 143-159.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S., (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86(3)*, 410-417.

Connolly, T., Arkes, H. R., & Hammond, K. R. (Eds.). (2000). Judgment and decision making: An interdisciplinary reader (2nd Ed). Cambridge: University Press.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Professional Manual – Revised NEO Personality Inventory (NEO PI-R) and NEO Five Factor Inventory (NEO-FFI)*. Lutz, FL: Psychological Assessment Resources, Inc.

Dalessio, A. T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology*, 9(1), 23-32.

Dawes, R. M. (1998). Behavioral decision making and judgment. In D. T. Gilbert & S. T.

Fiske (Eds.), *The handbook of social psychology*: Vol. 1. (4th ed., pp. 497-548). New York: McGraw-Hill.

Dulewicz, V. & Higgs, M. (2000). Emotional intelligence: A review and evaluation study. *Journal of Managerial Psychology, 15(4)*, 341-372.

Endsley, M. R. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37(1)*, 32-64.

Endsley, M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors, 37(1)*, 65-84.

Endsley, M. R. (2000a). Direct measurement of situation awareness: Validity and use of the SAGAT. In M. R. Endsley & D. J. Garland (Eds.) *Situation awareness analysis and measurement*. Mahwah, NJ: Lawrence Erlbaum Associations, Publishers.

Endsley, M. R. (2000b). Theoretical underpinnings of situation awareness: A critical review. In M. R. Endsley & D. J. Garland (Eds.) *Situation awareness analysis and measurement* (pp. 5-32). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Flach, J. M. (1999). Beyond error: The language of coordination and stability. In P. A. Hancock (Ed.) *Human performance and ergonomics* (pp. 109-128). San Diego: Academic Press.

Grant, L. D. (1997). A comprehensive examination of the latent structure of job performance. (Doctoral dissertation, North Carolina State University, 1996). *Dissertation Abstracts International, 57*, 6629.

Goleman, D. (1995). *Emotional Intelligence: Why it can matter more than IQ*. New York: Bantam Books.

Hanson, M. A. (1994). Development and construct validation of a situational judgment test of supervisory effectiveness for first-line supervisors in the U.S. Army. (Doctoral dissertation, The University of Minnesota, 1994). *Dissertation Abstracts International*.

Hanson, M. A. & Borman, W. C. (1993). Development and construct validation of the situational judgment test (SJT). ARI report #230.

Hanson, M. A., Horgen, K. E., & Borman, W. C. (1998). Situational judgment tests as measures of knowledge/expertise. Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology. Dallas, Texas.

Hanson, M. A. & Ramos, R. A. (1996). Situational judgment tests. In Richard S. Barrett (Ed.), *Fair employment strategies in human resource management* (pp. 118-124). Westport, CT: Quorum Books.

Hatcher, L. (1994). *A step-by-step approach to using the SAS® system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Hunter, J. E. & Schmidt, F. L. (1990). *Methods of meta-analysis*. Newbury Park: Sage Publications.

Jensen, A. R. (1993). Test validity: *g* versus "tacit knowledge." *Current Directions in Psychological Science 2(1)*, 9-10.

Klein, G. (1997). The recognition-primed decision (RPD) model: Looking back, looking forward. In C. E. Zsambok and G. Klein (Eds.) *Naturalistic decision making* (pp. 285-292). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Krokos, K. J. (1999). A test of an integrated model of correctional officer job performance. Unpublished master's thesis, North Carolina State University, Raleigh, NC.

Landy, F. J. & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87(1)*, 72-107.

Landy, F. J., Shankster, L. J., & Kohler, S. S. (1994). Personnel selection and placement. *Annual Review of Psychology, 45*, 261-296.

Legree, P. J. (1994). The effect of response format on reliability estimates for tacit knowledge scales. ARI Research Note 94-25.

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

Legree, P. J. & Busciglio, H. H. (1992). Tacit social knowledge and acquisition as a function of general intelligence and the ability to learn and utilize uncertain social feedback and contingencies. ARI Research Note 93-10.

Levin, I. P. & Reicks, C. J. (1996) Contemporary applications of research on judgment and decision making. In W. H. Loke (Ed.), *Perspectives on judgment and decision making* (pp. 19-30). Lanham, MD: The Scarecrow Press, Inc.

Loke, W. H. (1996). Models of judgment and decision making: An overview. In W. H. Loke (Ed.), *Perspectives on judgment and decision making* (pp. 3-18). Lanham,

MD: The Scarecrow Press, Inc.

McDaniel, M. A., Finnegan, E. B., Morgeson, F. P., Campion, M. A., & Braverman, E. P. (1997, April). *Predicting job performance from common sense*. Paper presented at the 12th annual Society of Industrial Organizational Psychology, St. Louis, MO.

McDaniel, M. A., Finnegan, E. B., Morgeson, F. P., 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(4)*, 730-740.

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

McClelland, D. C. (1973). Testing for competence rather than for "intelligence." *American Psychologist, 28*, 1-14.

Miller, M. L. (1996). Gender bias in mock jurors' decisions about damage awards. In W. H. Loke (Ed.), *Perspectives on judgment and decision making* (pp. 207-218). Lanham, MD: The Scarecrow Press, Inc.

Motowidlo, S. J., Borman, W., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10(2)*, 71-83.

Motowidlo, S. J. & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology, 66*, 337-344.

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. & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79(4)*, 475-480.

Northrop, L. C. (1976). The definition and measurement of judgment. Technical Report: 76-17. United States Civil Service Commission. Washington, DC.

Orasanu, J. & Connolly, T. (1993). The reinvention of decision making, In G. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 188-201). Norwood, NJ: Ablex.

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)*, 403-411.

Ployhart, R. E. & Ryan, A. M. (2000). A construct-oriented approach for developing situational judgment tests in a service context. Unpublished manuscript.

Pond, S. B. & Cantwell, A. (2002). Park Scholars Selection Process: Reliability Analyses Technical Report (Technical Report 2001.01 in partial fulfillment of requirements specified in Park Scholars Research Proposal 2001-2002. North Carolina State University).

Radford, M. H. B. (1996). Culture and its effects on decision making. In W. H. Loke (Ed.), *Perspectives on judgment and decision making* (pp. 49-69). Lanham, MD: The Scarecrow Press, Inc.

Randel, J. M, Pugh, H. L., & Reed, S. K. (1996). Differences in expert and novice situation awareness in naturalistic decision making. *Int. J. Human-Computer Studies, 45*, 579-597.

Rettinger, D. A. & Hastie, R. (2001). Content effects on decision making. *Organizational Behavior and Human Decision Processes*, 85(2), 336-359.

Ree, M. J. & Earles, J. A. (1993). *g* is to psychology what carbon is to chemistry[1]: A reply to Sternberg and Wagner, McClelland, and Calfee. *Current Directions in Psychological Science, 2(1)*, 11-12.

Rivero, J. C., Holtgrave, D. R., Bontempo, R. N. & Bottom, W. P. (1996). The ST. Petersburg Paradox: Data at last. In W. H. Loke (Ed.) *Perspectives on judgment and decision making* (pp. 263-272). Lanham, MD: The Scarecrow Press, Inc.

Robins, K. W. (1994). Effects of personality and situational judgment on job performance. (Doctoral dissertation, The University of California at Berkeley, 1994). *Dissertation Abstracts International*.

Schmidt, F. L. (1994). The future of personnel selection in the U.S. Army. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 333-350). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Schmidt, F. L. & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge. *Current Directions in Psychological Science, 2(1)*, 8-9.

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.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85(6)*, 956-970.

Sim, B. & Loke, W. H. (1996). A review of the influence of extralegal factors on legal decision making. In W. H. Loke (Ed.), *Perspectives on judgment and decision making* (pp. 201-206). Lanham, MD: The Scarecrow Press, Inc.

Smiderle, D., Perry, B. A., & Cronshaw, S. F. (1994). Evaluation of a video-based assessment in transit operator selection. *Journal of Business and Psychology, 9(1)*, 3-22.

Smith, P. C. (1976). In M. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 745-775). Chicago: Rand McNally College Publishing Company.

Sternberg, R. J. & Wagner, R. K. (1993). The g-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science, 2(1)*, 1-5.

Sternberg, R. J., Wagner, R. K., & Okagaki, L. (1993). Practical intelligence: The nature and role of tacit knowledge in work and at school. In J. E. Puckett & H. W. Reese (Eds.), *Mechanisms of everyday cognition* (pp. 205-227). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

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(2), 207-228.

Thomas, S. A., Doyle, J., & Browning, C. (1996). Clinical decision making: What do we know about real-world performance? In W. H. Loke (Ed.) *Perspectives on judgment and decision making* (pp. 175-186). Lanham, MD: The Scarecrow Press, Inc.

Vasilopoulos, N. L., Reilly, R. R., & Leaman, J. A. (2000). The influence of job familiarity and impression management on self-report measure scale scores and response latencies. *Journal of Applied Psychology, 85(1)*, 50-64.

Vernon, P. A., Nador, S., & Kantor, L. (1985). Reaction times and speed-of-processing: Their relationship to timed and untimed measures of intelligence. *Intelligence, 9*, 357-374.

Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology, 52(6)*, 1236-1247.

Weekley, J. A. & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*, 25-49

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

Wilson, M. A. & Grant, L. D. (1997). Validation of a trooper selection system: Project technical report (Research Proposal Number: 96-1147 NCSU).

Woehr, D. J. & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Appendixes

Appendix A
Scholar Performance Criteria

The performance dimensions for scholars are Scholarship, Leadership, Service, and Character. Each dimension is defined below.

Scholarship

1. Academic performance
    - Good grades: official PF rule is that scholars must meet the following minimums: 2.5 for first year, 2.75 for 2nd year, 2.9 for 3rd year, and 3.0 for 4th year. However, the expectation is that scholars have 3.0 GPA. The best scholars will excel well beyond this minimum.
    - Making appropriate progress towards degree: must take the required 15 hours per semester for a total of 120 semester hours over 8 semesters.

2. Effective management of potential
    - Explores outside learning experiences: seeks, applies for, and participates in research, special programs, projects, or conferences outside of class that will enhance learning and goal achievement
    - Broadly trained in area of interest: takes a variety of classes that support long term goals
    - Career objectives: has a well written and well thought out personal plan of development (PPD) that outlines specific plans to meet career goals or plans to clarify what goals are.
    - Develops resources: finds and works with faculty mentors who can assist in reaching and/or defining goals
    - Displays and cultivates special interests and aptitudes

3. Critical thinking skills
    - Intellectual curiosity: burning desire to learn evidenced by asking questions and exploring new interests, talents, and ideas
    - Ability to see/make connections: integrates existing knowledge with new ideas, considers alternative perspectives
    - Intellectual awareness and perseverance: cognizant of when s/he doesn't know something and takes steps to find out, asks questions in class for verification, asks for help from appropriate sources

Leadership

1. Has leadership skills
    - Can identify gaps and what needs to be done, knows when to step up and when not to, listens and communicates well, develops diplomatic and cooperative relationships with and between others, handles conflict appropriately, appreciates differences and can alter behavior according to what is needed, comfortable in

novel or unstructured social settings, a risk-taker who is willing to champion new ideas, balances personal goals with those of the group, thinks big and can negotiate the real world, can move or encourage others to accept a particular value or attitude, or to take action in a specific direction

2. Takes leadership role
    - Directly in a forefront leadership position and/or
    - Indirectly by influencing at a more molar level or providing a good example for others

Service

1. Commitment to public service: being a good scholar means making a positive difference in the lives of others.
    - At a minimum, scholars must volunteer each and every semester in service activities. The best scholars will conduct a needs assessment to identify gaps in a service organization, and then work with the organization to fill those gaps in innovative and creative ways that match his/her own interests/skills. Want scholars to begin to find service to be intrinsically rewarding, and to be capable of serving on a board for a nonprofit group upon graduation.

Character

1. Self-aware
    - Knows strengths and weaknesses, understands the ramifications of choices, manages self and resources well, finds an informed balance in activities, engages in activities "just for fun," engages in physical activity

2. Responsible
    - Knows the rules, is able to assess situations and to use good judgment

3. Integrity
    - Honest: doesn't cheat or steal or help others do so, true to his/her word
    - Conscientious: goes out of his/her way to deliver products as promised
    - Stands up for values and for what is right: would assist a fellow student who was being badly-treated

4. Adaptable/Resilient
    - Has the ability to profit from experience: recognizes the lessons in difficult situations and takes them forward by showing improvement over time in areas of weakness
    - Has good coping skills: won't fall apart when faced with failure, difficulty, or change, can regroup with minimal disruption and distress, has and uses a social network, handles separation from family appropriately
    - Tolerates ambiguity

5. Presence

1) Self–confident: adheres to views and values and can back them up, makes appropriate eye contact and shakes hands when meeting
2) Is sincere, well mannered, and possesses appropriate humility
    - Sociable: gets along well with others, looks out for classmates
    - Shows respect: for self and others

Appendix B
Situational Judgment Inventory Instructions

This inventory is designed to find out what kind of decisions you would likely make in complex situations. The inventory presents three (3) situations that may be encountered by scholars at XYZ University (XYZ). Each situation is comprised of 6-8 multiple-choice questions.

The questions within the three situations build upon each other. Therefore, unless instructed otherwise, you should assume that the details that are presented in each question are true for the remaining questions in that situation. For example, if one of the questions in Situation #2 indicates that the main character is majoring in engineering, then unless the following question(s) in that situation indicates otherwise, you should assume that the student is an engineering major for the remaining questions in Situation #2.

Please read the questions carefully. Choose the alternative that you would <u>most likely do</u> and enter the letter of that alternative on the answer sheet provided. Then choose the alternative you would <u>least likely do</u> and enter the letter of that alternative on the answer sheet provided. Please do not write on this inventory. Note that the situations you read about may not reflect a likely scenario for you as a student. For example, the student may have a different major or may have different interests than you. However, try to picture yourself in that situation and choose the answer that <u>you</u> would most likely do if you were in that situation.

One of the questions in each of the three situations is followed by an additional set of questions that asks about your answer to the previous question. Follow the instructions provided in answering those questions.

The first portion of your administration is timed. Once you begin, you will be given seven (7) minutes to work on the questions. Try to work quickly and accurately. At the end of 7 minutes, you will be asked to stop working and to circle the item on your answer sheet that you last completed. Then you will be instructed to resume taking the inventory, beginning where you stopped and working through the remainder of the inventory. After resuming work, please do not go back and change the answers on your answer sheet before the item you circled. The remainder of the administration is not timed. However, the entire administration should take less than one hour.

Appendix C
Sample SJT Item*

Kelly struggles with a difficult schedule. Kelly is a freshman scholar enrolled in the premed program at XYZ University. Kelly's long range plan is to become a pediatrician. Several courses that are required for the premed major are being offered in the fall of the upcoming school year, including Organic Chemistry and Biology. Although this makes for a difficult schedule, Kelly has registered for both classes in hopes of staying on track with finishing the premed major in four years. After a relaxing summer at home, Kelly returns to school in the fall ready to tackle the difficult course load. However, the workload is much more difficult than expected. Kelly worked very hard and felt well prepared for the first test in Organic Chemistry but got a C-. The next exam in is in two weeks. What would you do if you were in Kelly's situation?

a) Drop the class and not pursue it further.
b) Drop the class but plan to pursue your interest during the summer break.
c) Ask your friend Alex for advice.
d) Drop the class but plan to take it whenever it is offered next.
e) Keep the class.

*This item is a sample item only; it does not actually appear on the SJT used in this research. The SJT may be available for research purposes only.

Appendix D
Description and Descriptive Statistics for Cognitive Information Processing Model Variables

VAR	DESCRIPTION	MEAN	SD	N
Q11ds	SA - ID the Dilemma (Scholarship)	.98	.15	211
Q11es	SA - ID the Goal (Scholarship)	.60	.49	211
Q11fs	SA - ID the Future Projection (Scholarship)	.69	.46	211
SAS	SA - SA Summary Score (Scholarship)	75.51	27.52	211
Q11i	RETR - # of Similar Situations Recalled (Scholarship)	1.60	1.04	211
Q11j	RETR - Similarity of Most Similar Situation Recalled (Scholarship)	3.11	1.27	211
Q11g	GEN - #of Consequences Considered (Scholarship)	3.88	1.26	211
Q11h	GEN - # of Novel Alternatives Generated (Scholarship)	.99	.87	211
Q11ba	DM-Minimize Cost (Scholarship)	3.16	1.66	211
Q11bb	DM-Interpersonal Relationship Factor(s) (Scholarship)	2.73	1.53	211
Q11bc	DM-Probability of Outcomes (Scholarship)	4.70	1.53	211
Q11bd	DM-Guessing (Scholarship)	1.36	.93	211
Q11be	DM-Moral/Ethical Considerations (Scholarship)	2.42	1.76	211
Q11bf	DM-Maximize Returns (Scholarship)	4.95	1.54	211
Q11bg	DM-Self-Image Considerations (Scholarship)	2.66	1.71	211
Q11bh	DM-Personal Utility of Outcomes (Scholarship)	5.37	1.43	211
Q11bi	DM-Similarity of Alternative to Answer (Scholarship)	4.14	2.00	211
Q11c	DM - Expert vs. Novice Strategy (Scholarship)	1.85	.64	211
SJT1M	SJ - SJT M-correct Score (Scholarship)	48.72	21.64	211
SJT1L	SJ - SJT L-correct Score (Scholarship)	52.61	23.06	211
Q26ds	SA - ID the Dilemma (Leadership)	.59	.49	211
Q26es	SA - ID the Goal (Leadership)	.67	.47	211
Q26fs	SA - ID the Future Projection (Leadership)	.81	.39	211
SAL	SA - SA Summary Score (Leadership)	69.04	26.42	211
Q26i	RETR - # of Similar Situations Recalled (Leadership)	1.10	.97	211
Q26j	RETR - Similarity of Most Similar Situation Recalled (Leadership)	2.10	1.51	211
Q26g	GEN - #of Consequences Considered (Leadership)	4.22	1.24	211
Q26h	GEN - # of Novel Alternatives Generated (Leadership)	.75	.84	211
Q26ba	DM-Minimize Cost (Leadership)	2.35	1.66	211
Q26bb	DM-Interpersonal Relationship Factor(s) (Leadership)	5.88	1.11	211
Q26bc	DM-Probability of Outcomes (Leadership)	5.04	1.48	211
Q26bd	DM-Guessing (Leadership)	1.24	.63	211
Q26be	DM-Moral/Ethical Considerations (Leadership)	4.62	1.65	211
Q26bf	DM-Maximize Returns (Leadership)	4.44	1.66	211
Q26bg	DM-Self-Image Considerations (Leadership)	3.76	1.61	211
Q26bh	DM-Personal Utility of Outcomes (Leadership)	3.65	1.52	211
Q26bi	DM-Similarity of Alternative to Answer (Leadership)	3.21	1.87	211
Q26c	DM - Expert vs. Novice Strategy (Leadership)	1.76	.70	211
SJT2M	SJ - SJT M-correct Score (Leadership)	53.50	16.20	211
SJT2L	SJ - SJT L-correct Score (Leadership)	46.99	16.48	211

Appendix D, Cont'd
Descriptive Statistics for Cognitive Information Processing Model Variables

Q33ds	SA - ID the Dilemma (Service)	.96	.20	211
Q33es	SA - ID the Goal (Service)	.73	.45	211
Q33fs	SA - ID the Future Projection (Service)	.66	.47	211
SASV	SA - SA Summary Score (Service)	78.20	22.27	211
Q33i	RETR - # of Similar Situations Recalled (Service)	.82	.75	211
Q33j	RETR - Similarity of Most Similar Situation Recalled (Service)	1.89	1.61	211
Q33g	GEN - #of Consequences Considered (Service)	4.05	1.37	211
Q33h	GEN - # of Novel Alternatives Generated (Service)	.70	.84	211
Q33ba	DM-Minimize Cost (Service)	2.12	1.43	211
Q33bb	DM-Interpersonal Relationship Factor(s) (Service)	2.76	1.50	211
Q33bc	DM-Probability of Outcomes (Service)	4.90	1.54	211
Q33bd	DM-Guessing (Service)	1.43	1.04	211
Q33be	DM-Moral/Ethical Considerations (Service)	4.68	1.72	211
Q33bf	DM-Maximize Returns (Service)	4.65	1.69	211
Q33bg	DM-Self-Image Considerations (Service)	2.65	1.54	211
Q33bh	DM-Personal Utility of Outcomes (Service)	3.95	1.62	211
Q33bi	DM-Similarity of Alternative to Answer (Service)	3.47	2.07	211
Q33c	DM - Expert vs. Novice Strategy (Service)	1.71	.59	211
SJT3M	SJ - SJT M-correct Score (Service)	52.37	22.23	211
SJT3L	SJ - SJT L-correct Score (Service)	44.39	28.06	211

Appendix E
Correlation Matrix for Cognitive Processing Model – Scholarship Situation

	Q11ds	Q11es	Q11fs	SAS	Q11i	Q11j	Q11g	Q11h	Q11ba	Q11bb	Q11bc	Q11bd	Q11be	Q11bf	Q11bg	Q11bh	Q11bi	Q11c	SJT1M	SJT1L
MEAN	.98	.60	.69	75.51	1.60	3.11	3.88	.99	3.16	2.73	4.70	1.36	2.42	4.95	2.66	5.37	4.14	1.85	48.72	52.61
STD	.15	.49	.46	27.52	1.04	1.27	1.26	.87	1.66	1.53	1.53	.93	1.76	1.54	1.71	1.43	2.00	.64	21.64	23.06
N	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211
Q11ds	1.																			
Q11es	.06	1.																		
Q11fs	.10	***.39***	1.																	
SAS	***.28***	***.83***	***.81***	1.																
Q11i	.00	***.18***	.06	**.14**	1.															
Q11j	.11	.09	-.06	.04	***.58***	1.														
Q11g	-.04	.01	-.05	-.03	***.19***	***.25***	1.													
Q11h	.07	.01	.05	.05	**.17**	.12	**.16**	1.												
Q11ba	.01	**-.16**	-.08	-.13	-.11	-.11	.00	-.08	1.											
Q11bb	.03	.13	.07	.12	.05	.11	-.06	.04	.00	1.										
Q11bc	-.09	-.03	-.10	-.09	-.05	.01	.06	.08	.01	.05	1.									
Q11bd	.06	-.13	-.12	-.13	.13	***.20***	-.01	.08	-.01	***.23***	.00	1.								
Q11be	.02	.08	.10	.11	.01	-.02	.06	**.15**	.12	***.31***	**.16**	***.26***	1.							
Q11bf	-.01	-.02	.08	.04	-.02	-.04	.01	.12	.05	-.08	***.22***	-.12	.07	1.						
Q11bg	.02	.05	**.14**	.11	.01	**.14**	.06	.13	.06	***.19***	.12	***.23***	***.34***	.13	1.					
Q11bh	.06	.06	.05	.08	-.02	-.09	-.01	.06	-.02	-.08	.10	-.13	.01	***.42***	.12	1.				
Q11bi	.03	-.06	-.04	-.05	.02	.09	.01	.04	-.05	.07	-.03	.00	-.10	-.04	**.16**	-.02	1.			
Q11c	.06	-.08	.06	.00	-.05	.04	***-.28***	-.08	.06	.02	-.09	-.10	-.13	-.08	-.11	**-.17**	.10	1.		
SJT1M	.12	***.23***	***.18***	***.26***	.01	.01	.03	.10	-.12	.02	-.08	***-.20***	.02	.05	-.03	-.02	-.01	.00	1.	
SJT1L	.03	.02	.04	.04	.02	.06	**.15**	***.22***	**-.15**	.13	.03	.07	.09	**.17**	.09	.11	-.04	-.05	**.18**	1.

Bold = significant at $p < .05$

Bold and Italicized = significant at $p < .01$

Appendix F
Correlation Matrix for Cognitive Processing Model – Leadership Situation

	Q26ds	Q26es	Q26fs	SAL	Q26i	Q26j	Q26g	Q26h	Q26ba	Q26bb	Q26bc	Q26bd	Q26be	Q26bf	Q26bg	Q26bh	Q26bi	Q26c	SJT2M	SJT2L
MEAN	.59	.67	.81	69.04	1.10	2.10	4.22	.75	2.35	5.88	5.04	1.24	4.62	4.44	3.76	3.65	3.21	1.76	53.50	46.99
STD	.49	.47	.39	26.42	.97	1.51	1.24	.84	1.66	1.11	1.48	.63	1.65	1.66	1.61	1.52	1.87	.70	16.20	16.48
N	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211
Q26ds	1.																			
Q26es	-.09	1.																		
Q26fs	**-.16**	***.31***	1.																	
SAL	***.49***	***.69***	***.58***	1.																
Q26i	.02	-.03	.01	.00	1.															
Q26j	.03	-.05	.04	.01	***.79***	1.														
Q26g	.06	.04	-.02	.05	.11	.05	1.													
Q26h	.12	.04	.01	.11	***.29***	***.33***	**.17**	1.												
Q26ba	.07	.04	-.07	.03	.04	.01	.08	.05	1.											
Q26bb	.01	***.18***	.13	***.18***	.10	.06	.08	.05	-.12	1.										
Q26bc	.07	.08	-.05	.07	.03	.03	.06	-.06	.01	***.23***	1.									
Q26bd	-.01	-.06	-.03	-.05	**-.14**	-.13	.02	-.04	.07	.00	.02	1.								
Q26be	-.03	.11	**.14**	.12	.02	.05	.04	.10	.03	***.31***	***.21***	.04	1.							
Q26bf	.01	.05	.00	.03	**.18**	**.14**	.12	.02	***.37***	.04	***.20***	.07	.11	1.						
Q26bg	.02	-.02	-.07	-.04	-.01	.01	.00	-.06	.02	***.24***	**.14**	.06	***.18***	.09	1.					
Q26bh	-.01	-.05	-.08	-.08	.07	.08	.02	-.02	***.18***	.04	**.16**	.11	**.15**	***.27***	***.37***	1.				
Q26bi	.10	.12	.09	**.17**	***.19***	***.27***	.09	**.17**	.02	.12	.03	-.01	**.14**	.12	-.01	.12	1.			
Q26c	-.02	-.02	-.01	-.03	-.05	-.03	***-.28***	-.06	.01	-.07	.05	.02	-.05	-.08	.11	-.01	.03	1.		
SJT2M	.00	.10	.08	.10	.10	**.15**	.04	-.06	-.01	.07	.05	-.09	.08	.04	-.10	-.04	.01	.02	1.	
SJT2L	**.16**	-.10	.07	.08	.09	**.14**	.07	**.16**	-.04	-.06	.02	-.06	-.02	.01	**-.16**	-.09	.04	-.01	.10	1.

Bold = significant at $p < .05$

Bold and Italicized = significant at $p < .01$

Appendix G
Correlation Matrix for Cognitive Processing Model – Service Situation

	Q33ds	Q33es	Q33fs	SASV	Q33i	Q33j	Q33g	Q33h	Q33ba	Q33bb	Q33bc	Q33bd	Q33be	Q33bf	Q33bg	Q33bh	Q33bi	Q33c	SJT3M	SJT3L
MEAN	.96	.73	.66	78.20	.82	1.89	4.05	.70	2.12	2.76	4.90	1.43	4.68	4.65	2.65	3.95	3.47	1.71	52.37	44.39
STD	.20	.45	.47	22.27	.75	1.61	1.37	.84	1.43	1.50	1.54	1.04	1.72	1.69	1.54	1.62	2.07	.59	22.23	28.06
N	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211	211
Q33ds	1.																			
Q33es	.13	1.																		
Q33fs	-.10	-.06	1.																	
SASV	***.32***	***.67***	***.64***	1.																
Q33i	.07	.02	.02	.05	1.															
Q33j	.07	.02	.01	.04	***.81***	1.														
Q33g	.01	.01	-.01	.00	.13	.12	1.													
Q33h	.06	.09	-.06	.04	***.27***	***.25***	.08	1.												
Q33ba	-.03	-.04	.00	-.04	-.01	.07	**.14**	-.05	1.											
Q33bb	-.03	.02	-.14	-.09	.07	.12	.03	.05	***.23***	1.										
Q33bc	-.06	-.05	.10	.02	-.06	.03	**.15**	.00	.03	***.25***	1.									
Q33bd	***-.21***	**-.15**	.03	**-.14**	-.07	-.04	***-.28***	.10	-.05	**.16**	.09	1.								
Q33be	-.07	.06	.08	.08	.13	.13	.13	**.15**	.13	***.28***	***.18***	.08	1.							
Q33bf	.05	-.03	.13	.09	-.04	.08	.13	.06	***.24***	-.03	***.23***	-.08	.09	1.						
Q33bg	.06	-.12	**-.15**	**-.17**	-.01	.05	-.03	-.06	***.18***	***.26***	.13	.06	**.17**	**.17**	1.					
Q33bh	-.05	-.01	.02	-.01	-.05	-.01	-.02	-.04	***.22***	**.18**	.09	.08	**.14**	***.22***	***.39***	1.				
Q33bi	.07	.05	.04	.08	.07	.08	-.02	.02	**-.18**	.01	.05	.08	.06	-.01	.13	.01	1.			
Q33c	.03	.00	-.07	-.04	.03	.04	***-.30***	-.06	***-.22***	-.05	-.10	.10	.02	**-.16**	.03	.02	.11	1.		
SJT3M	.02	***.21***	.06	***.19***	.13	.08	.10	***.20***	-.06	-.02	.11	-.05	.08	.01	-.09	-.12	.12	.05	1.	
SJT3L	.00	***.23***	.10	***.23***	**.16**	.06	.06	.13	-.03	.06	.07	-.10	**.14**	**-.15**	-.07	.01	.08	.06	***.31***	1.

Bold = significant at $p < .05$

Bold and Italicized = significant at $p < .01$

Appendix H
Description and Descriptive Statistics for Integrated Model of Performance Variables

	VAR	EXP	MEAN	STD	N
1	HSGPA	g-High School Grade Point Average	4.55	.35	176
2	SATV	g-SAT-Verbal Score	693.18	64.92	176
3	SATM	g-SAT-Math Score	715.45	62.17	176
4	CLASS	EXP-Class	2003.82	1.09	176
5	Q11i	EXP-# of Similar Situations Recalled-Scholarship	1.56	1.05	176
6	Q11j	EXP-Rating of Similarity of Most Similar Situation Recalled-Scholarship	3.10	1.34	176
7	Q17	EXP-Rating of Amount of Experience with Dilemma-Scholarship	3.11	.98	176
8	Q26i	EXP-# of Similar Situations Recalled-Leadership	1.06	.91	176
9	Q26j	EXP-Rating of Similarity of Most Similar Situation Recalled-Leadership	2.09	1.50	176
10	Q29	EXP-Rating of Amount of Experience with Dilemma-Leadership	2.88	.84	176
11	Q33i	EXP-# of Similar Situations Recalled-Service	.82	.76	176
12	Q33j	EXP-Rating of Similarity of Most Similar Situation Recalled-Service	1.88	1.63	176
13	Q38	EXP-Rating of Amount of Experience with Dilemma-Service	2.49	.98	176
14	C1Comp	CON-NEO PI-R™ C1: Competence	22.88	3.68	176
15	C2Ord	CON-NEO PI-R™ C2: Order	17.54	5.72	176
16	C3Duti	CON-NEO PI-R™ C3: Dutifulness	23.14	4.69	176
17	C4Ach	CON-NEO PI-R™ C4: Achievement Striving	21.75	4.77	176
18	C5SelfD	CON-NEO PI-R™ C5: Self-Discipline	20.35	6.02	176
19	C6Del	CON-NEO PI-R™ C6: Deliberation	18.19	4.81	176
20	A1Tru	AGR-NEO PI-R™ A1: Trust	21.50	4.68	176
21	A2Stra	AGR-NEO PI-R™ A2: Straightforwardness	20.19	4.70	176
22	A3Altr	AGR-NEO PI-R™ A3: Altruism	24.97	3.40	176
23	A4Com	AGR-NEO PI-R™ A4: Compliance	18.52	4.28	176
24	A5Mod	AGR-NEO PI-R™ A5: Modesty	19.81	4.44	176
25	A6Ten	AGR-NEO PI-R™ A6: Tender-Mindedness	21.79	4.00	176
26	N1Anx	NforS-NEO PI-R™ N1: Anxiety	16.00	5.46	176
27	N2AngH	NforS-NEO PI-R™ N2: Angry Hostility	11.82	4.85	176
28	N3Depr	NforS-NEO PI-R™ N3: Depression	13.15	5.15	176
29	N4SelfC	NforS-NEO PI-R™ N4: Self-Consciousness	15.03	4.83	176
30	N5Imp	NforS-NEO PI-R™ N5: Impulsiveness	16.45	5.11	176
31	N6Vul	NforS-NEO PI-R™ N6: Vulnerability	10.22	3.89	176
32	SJT4	SJ-Overall SJT Score	47.95	9.68	176
33	SGPA	AP-Semester GPA from Spring 2002	3.76	.38	176
34	TGPA	AP-Total GPA for Time at University	3.79	.29	176
35	R1s1	PSP-Supervisor Rating 1 - Scholarship (Managing Academic Potential)	5.00	.97	176
36	R2s2	PSP-Supervisor Rating 2 - Scholarship (Critical Thinking Skills)	4.71	.77	176
37	R3l1	PSP-Supervisor Rating 3 - Leadership (Leadership Skills)	4.50	.97	176
38	R4l2	PSP-Supervisor Rating 4 - Leadership (Seeking/Accepting Leadership Roles)	4.50	1.00	176
39	R5sv1	PSP-Supervisor Rating 5 - Service (Service Behavior)	4.25	1.10	176
40	CSHr	PSP-Mean Hours of Community Service	29.99	21.31	176
41	R6c1	CP-Supervisor Rating 6 - Character (Self Awareness)	4.46	.90	176
42	R7c2	CP-Supervisor Rating 7 - Character (Integrity)	4.60	.79	176
43	R8c3	CP-Supervisor Rating 8 - Character (Adaptable/Resilient)	4.58	.95	176
44	R9c4	CP-Supervisor Rating 9 - Character (Presence)	4.54	.88	176

Appendix I
Correlation Matrix for Integrated Model of Performance Variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1.																
2	**.18**	1.															
3	**.19**	***.34***	1.														
4	-.01	-.04	-.06	1.													
5	-.01	-.02	**.18**	-.1	1.												
6	.04	.00	.11	-.05	***.64***	1.											
7	-.04	.08	.04	-.14	***.39***	***.41***	1.										
8	-.03	.10	.02	.06	**.18**	.13	**.16**	1.									
9	-.06	-.01	.00	.08	.09	.10	.13	***.79***	1.								
10	.04	.07	-.08	-.02	***.24***	***.24***	***.26***	***.40***	***.38***	1.							
11	.04	.13	-.02	.07	**.18**	**.18**	***.23***	.13	.07	**.16**	1.						
12	.11	.13	-.09	.11	**.16**	***.21***	***.26***	.08	.04	**.18**	***.82***	1.					
13	**.16**	.03	.04	-.03	.11	.10	***.21***	.08	.12	***.25***	**.18**	***.22***	1.				
14	**.16**	-.04	.03	**-.18**	-.02	-.07	-.05	.06	.02	-.01	-.08	-.11	.13	1.			
15	**.15**	-.14	.01	-.08	-.09	-.09	-.12	-.02	.06	-.07	**-.15**	-.11	.08	***.53***	1.		
16	***.24***	-.03	.09	-.08	-.09	-.07	-.10	-.01	-.04	-.11	-.08	-.07	.05	***.71***	***.54***	1.	
17	**.16**	-.03	-.08	.03	-.02	-.07	-.08	.01	.09	**.16**	-.02	.00	**.16**	***.55***	***.53***	***.57***	1.
18	**.19**	-.08	-.03	-.04	-.03	-.10	**-.15**	-.04	-.02	-.03	-.07	-.10	.13	***.71***	***.63***	***.76***	***.73***
19	**.17**	-.06	-.03	**-.16**	-.15	**-.17**	-.22	-.12	-.14	-.11	**-.16**	-.13	-.03	***.51***	***.43***	***.47***	***.36***
20	.02	-.13	-.02	-.02	.11	.13	**.17**	.04	.06	.01	.15	.04	.09	**.19**	.02	***.22***	.09
21	.12	-.04	.01	**-.15**	-.13	-.08	**-.15**	**-.19**	-.20	-.28	-.04	-.08	-.06	**.19**	.14	***.34***	.14
22	.00	**-.19**	-.22	-.01	.01	.07	.01	-.02	.02	-.07	**.18**	***.2***	-.01	.15	.04	***.20***	.07
23	.07	-.04	-.05	-.13	-.01	-.01	-.08	-.11	**-.17**	-.10	.05	.00	**-.19**	-.04	-.10	.09	-.09
24	-.05	-.11	-.12	.02	-.03	.01	-.11	**-.16**	-.12	-.25	.05	.09	-.07	**-.16**	.08	.15	.05
25	**-.16**	.00	-.11	.00	.03	.01	.11	.02	-.03	.00	***.23***	**.18**	.01	.03	-.14	.02	-.02
26	.00	-.09	**-.16**	.05	-.07	-.01	-.02	-.08	.01	**-.16**	-.09	.03	.00	-.29	.04	**-.15**	.03
27	-.08	.03	-.11	.04	-.05	.02	.05	-.03	.02	.00	-.12	-.04	.04	-.24	-.05	-.29	-.14
28	**-.16**	-.02	-.10	.08	.06	-.02	.02	.07	.14	-.07	.02	.09	**-.17**	-.48	-.10	-.32	**-.19**
29	-.05	-.04	-.08	.09	-.14	-.11	-.10	-.06	-.02	-.13	-.10	.00	-.06	-.44	-.11	-.22	**-.16**
30	-.08	-.03	-.06	.08	.04	.07	.11	.02	.12	.03	.03	.07	.09	-.45	-.21	-.51	-.31
31	.00	-.04	**-.15**	.06	-.09	-.06	-.09	-.13	-.06	**-.17**	-.12	-.02	-.10	-.57	**-.18**	-.41	-.33
32	-.09	-.06	-.10	.12	.04	.07	-.08	.12	.11	.01	.06	.03	.13	.12	.10	.08	.08
33	.12	.14	**.18**	-.12	-.02	.03	.04	-.09	-.07	-.03	-.02	-.06	.05	***.23***	***.26***	***.28***	***.28***
34	***.21***	***.20***	.14	-.09	-.06	-.05	.05	**-.19**	**-.16**	-.06	-.01	-.02	.03	***.20***	***.28***	***.23***	***.27***
35	.07	**.16**	-.02	-.27	.04	.03	.03	-.01	.01	**.17**	-.04	.01	**.16**	**.19**	***.24***	.14	***.35***
36	-.06	.15	-.03	-.29	.00	.00	.06	-.01	-.03	.12	-.03	-.02	.08	.08	.10	.06	***.24***
37	-.01	.00	-.09	-.33	.10	.02	.10	.07	.00	***.2***	-.09	-.07	.09	**.18**	**.17**	.08	.14
38	-.07	-.06	-.11	-.31	.06	.03	.04	.05	.08	**.19**	-.05	-.03	.07	**.17**	**.18**	.05	**.16**
39	.09	**.17**	.12	-.29	.13	.08	.13	**.18**	.15	**.17**	.03	.01	.15	***.23***	***.27***	**.19**	***.29***
40	-.03	-.04	.06	.10	.13	.10	.09	.04	.06	.02	.11	.05	.02	.06	**.17**	.07	***.23***
41	.00	.02	-.07	-.23	.01	-.06	-.01	.01	-.06	.07	-.10	-.08	.05	**.19**	***.20***	.14	**.16**
42	.03	.05	-.05	-.34	.07	.04	-.02	.06	.03	.12	-.09	-.07	.04	.15	**.18**	**.15**	***.22***
43	.00	.07	-.01	-.33	.00	.02	.02	.00	.01	**.15**	-.08	-.05	.08	**.18**	**.19**	.09	***.20***
44	-.07	.03	**-.16**	-.31	.03	.02	-.01	.07	.06	**.19**	-.12	-.11	.07	**.16**	***.22***	.06	***.25***

Appendix I, Cont'd
Correlation Matrix for Integrated Model of Performance Variables

	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1																	
2																	
3																	
4																	
5																	
6																	
7																	
8																	
9																	
10																	
11																	
12																	
13																	
14																	
15																	
16																	
17																	
18	1.																
19	***.47***	1.															
20	***.24***	-.01	1.														
21	***.28***	***.36***	***.25***	1.													
22	**.15**	.07	***.36***	***.27***	1.												
23	.08	.12	***.30***	***.44***	***.34***	1.											
24	.11	.00	-.01	***.39***	***.28***	***.26***	1.										
25	.02	-.08	***.38***	***.21***	***.45***	***.33***	**.17**	1.									
26	-.14	.00	-.1	.12	.08	-.07	**.16**	-.02	1.								
27	-.26	-.23	-.26	-.3	-.26	-.49	-.25	-.22	***.46***	1.							
28	-.31	**-.19**	-.26	-.07	-.10	-.11	***.25***	-.01	***.55***	***.35***	1.						
29	-.24	-.08	-.20	.12	-.13	-.04	***.25***	-.05	***.57***	***.33***	***.61***	1.					
30	-.44	-.56	.01	-.22	-.04	**-.18**	-.12	.02	***.36***	***.48***	***.27***	***.26***	1.				
31	-.41	**-.17**	-.11	.03	-.03	.05	.10	-.04	***.66***	***.48***	***.55***	***.51***	***.47***	1.			
32	.05	-.02	.14	.03	**.16**	.04	.10	.14	-.11	**-.15**	-.11	-.20	-.1	-.13	1.		
33	***.34***	.10	.12	**.17**	-.08	.13	.01	.00	**.19**	.03	-.05	.05	.01	-.03	-.1	1.	
34	***.30***	**.18**	.11	***.23***	-.15	.11	.01	-.03	**.16**	.02	-.09	.04	-.02	.00	**-.16**	***.82***	1.
35	***.27***	**.17**	.04	.15	-.09	.04	-.04	-.03	.01	-.02	-.12	-.12	-.09	-.07	-.02	***.30***	***.42***
36	.15	**.17**	.04	**.16**	-.09	.12	.01	.02	.01	-.08	**-.15**	-.02	-.05	-.11	-.03	***.27***	***.37***
37	.15	**.16**	.13	**.19**	.03	**.16**	.01	.11	.03	-.08	-.04	-.07	-.07	.00	.08	**.18**	***.22***
38	.15	**.16**	.10	.13	.00	.11	.00	.07	.03	.00	-.04	-.03	-.05	-.01	.05	**.16**	**.16**
39	***.22***	.09	**.18**	.13	.06	.10	-.02	.14	-.08	-.1	**-.17**	**-.19**	-.05	-.12	.05	***.23***	***.29***
40	**.17**	.06	***.26***	.06	.11	.07	.02	.15	-.04	-.11	-.12	**-.17**	.01	-.06	**.15**	.08	.07
41	**.16**	***.21***	.04	***.25***	-.04	***.21***	.06	.03	-.02	-.13	-.09	.02	-.14	-.03	.02	***.24***	***.36***
42	**.17**	.14	.09	***.20***	.01	***.20***	.09	.08	-.04	**-.15**	-.12	-.06	**-.16**	-.08	.01	**.17**	***.24***
43	**.15**	**.18**	.06	***.22***	-.04	.14	.00	.01	-.03	-.11	**-.17**	-.09	-.06	-.08	-.02	***.22***	***.28***
44	***.20***	.11	.13	.15	.02	.10	.00	.07	.08	.03	-.11	-.10	.03	.00	-.01	***.25***	***.29***

Appendix I, Cont'd
Correlation Matrix for Integrated Model of Performance Variables

	35	36	37	38	39	40	41	42	43	44
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										
26										
27										
28										
29										
30										
31										
32										
33										
34										
35	1.									
36	***.72***	1.								
37	***.49***	***.48***	1.							
38	***.46***	***.50***	***.81***	1.						
39	***.49***	***.45***	***.58***	***.54***	1.					
40	***.24***	***.20***	**.19**	**.19**	***.39***	1.				
41	***.61***	***.58***	***.67***	***.57***	***.52***	.13	1.			
42	***.56***	***.57***	***.62***	***.59***	***.55***	.14	***.77***	1.		
43	***.59***	***.63***	***.61***	***.61***	***.54***	**.17**	***.74***	***.72***	1.	
44	***.59***	***.61***	***.72***	***.70***	***.55***	**.17**	***.68***	***.71***	***.73***	1.

Bold = significant at $p < .05$
Bold and Italicized = significant at $p < .01$

Appendix J
Correlations between Experience Variables and Expert Processing Variables

	CLASS	Q17	Q29	Q38	Q11i	Q11j	Q26i	Q26j	Q33i	Q33j	Q11c	Q26c	Q33c
MEAN	2003.59	3.14	2.89	2.49	1.60	3.11	1.10	2.10	.82	1.89	1.85	1.76	1.71
STD	1.14	.99	.85	.99	1.04	1.27	.97	1.51	.75	1.61	.64	.70	.59
N	211	211	211	211	211	211	211	211	211	211	211	211	211
CLASS	1.												
Q17	**-.15**	1.											
Q29	-.03	***.26***	1.										
Q38	-.05	**.16**	***.23***	1.									
Q11i	-.12	***.43***	***.22***	.06	1.								
Q11j	-.03	***.38***	***.18***	.09	***.58***	1.							
Q26i	.02	***.22***	***.37***	.05	***.25***	.12	1.						
Q26j	.07	***.19***	***.38***	.11	**.15**	.11	***.79***	1.					
Q33i	.06	***.22***	***.21***	**.15**	**.17**	.13	***.2***0	.13	1.				
Q33j	.08	***.19***	***.22***	***.19***	.11	**.14**	.10	.06	***.81***	1.			
Q11c	.00	.03	.00	-.12	-.05	.04	-.05	-.06	.01	.00	1.		
Q26c	.03	-.10	.07	-.13	-.03	-.08	-.05	-.03	-.08	-.03	***.25***	1.	
Q33c	-.01	.06	.05	-.05	-.03	-.09	.00	-.05	.03	.04	***.33***	***.50***	1.

Bold = significant at $p < .05$
Bold and Italicized = significant at $p < .01$

Appendix K
Correlations Between Expert Processing Variables and SJT Summary Scores

	Q11c	Q26c	Q33c	MSCORES	MSCOREL	MSCORESV	LSCORES	LSCOREL	LSCORESV	SJTo3score
MEAN	1.85	1.76	1.71	48.72	53.50	52.37	52.61	46.99	44.39	50.19
STD	.64	.70	.59	21.64	16.20	22.23	23.06	16.48	28.06	10.94
N	211	211	211	211	211	211	211	211	211	211
Q11c	1.									
Q26c	***.25***	1.								
Q33c	***.33***	***.50***	1.							
MSCORES	.00	.10	.06	1.						
MSCOREL	.00	.02	-.01	.11	1.					
MSCORESV	.02	-.09	.05	**.16**	.04	1.				
LSCORES	-.05	-.08	.00	**.18**	**.15**	***.22***	1.			
LSCOREL	-.10	-.01	-.09	.12	.10	.09	***.28***	1.		
LSCORESV	.05	.06	.06	.13	.09	***.31***	***.20***	**.18**	1.	
SJTo3score	-.03	.00	.01	***.52***	***.52***	***.50***	***.64***	***.56***	***.52***	1.

Bold = significant at $p < .05$
Bold and Italicized = significant at $p < .01$

Appendix L
EFA Results for the Personality Inventory Items (n=18)

	C1	C2	C3	C4	C5	C6	A1	A2	A3	A4	A5	A6	N1	N2	N3	N4	N5	N6
MEAN	22.88	17.54	23.14	21.75	20.35	18.19	21.5	20.19	24.97	18.52	19.81	21.79	16.00	11.82	13.15	15.03	16.45	10.22
STD	3.68	5.72	4.69	4.77	6.02	4.81	4.68	4.70	3.4	4.28	4.44	4.00	5.46	4.85	5.15	4.83	5.11	3.89
N	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176
C1	1.																	
C2	.53	1.																
C3	.71	.54	1.															
C4	.55	.53	.57	1.														
C5	.71	.63	.76	.73	1.													
C6	.51	.43	.47	.36	.47	1.												
A1	.19	.02	.22	.09	.24	-.01	1.											
A2	.19	.14	.34	.14	.28	.36	.25	1.										
A3	.15	.04	.20	.07	.15	.07	.36	.27	1.									
A4	-.04	-.10	.09	-.09	.08	.12	.30	.44	.34	1.								
A5	-.16	.08	.15	.05	.11	.00	-.01	.39	.28	.26	1.							
A6	.03	-.14	.02	-.02	.02	-.08	.38	.21	.45	.33	.17	1.						
N1	-.29	.04	-.15	.03	-.14	.00	-.10	.12	.08	-.07	.16	-.02	1.					
N2	-.24	-.05	-.29	-.14	-.26	-.23	-.26	-.30	-.26	-.49	-.25	-.22	.46	1.				
N3	-.48	-.10	-.32	-.19	-.31	-.19	-.26	-.07	-.10	-.11	.25	-.01	.55	.35	1.			
N4	-.44	-.11	-.22	-.16	-.24	-.08	-.20	.12	-.13	-.04	.25	-.05	.57	.33	.61	1.		
N5	-.45	-.21	-.51	-.31	-.44	-.56	.01	-.22	-.04	-.18	-.12	.02	.36	.48	.27	.26	1.	
N6	-.57	-.18	-.41	-.33	-.41	-.17	-.11	.03	-.03	.05	.10	-.04	.66	.48	.55	.51	.47	1.
F1	.76	.72	.78	.75	.85	.57	.10	.26	.09	-.10	.04	-.09	.06	-.08	-.21	-.14	-.40	-.30
F2	-.43	.03	-.21	-.06	-.18	-.08	-.16	.13	.00	-.07	.28	-.03	.83	.50	.69	.70	.41	.76
F3	-.04	.00	.21	-.03	.11	.37	.00	.53	.19	.53	.52	.14	-.07	-.56	.06	.14	-.51	-.05
F4	.10	-.06	.13	.04	.15	-.14	.61	.33	.61	.45	.18	.59	.05	-.26	-.17	-.15	.17	.02

Appendix M
EFA Results for the Supervisor Ratings of Performance (n=9)

	R1s1	R2s2	R3l1	R4l2	R5sv1	R6c1	R7c2	R8c3	R9c4
MEAN	5.00	4.71	4.50	4.50	4.25	4.46	4.60	4.58	4.54
STD	.97	.77	.97	1.00	1.10	.90	.79	.95	.88
N	176	176	176	176	176	176	176	176	176
R1s1	1.								
R2s2	.72	1.							
R3l1	.49	.48	1.						
R4l2	.46	.50	.81	1.					
R5sv1	.49	.45	.58	.54	1.				
R6c1	.61	.58	.67	.57	.52	1.			
R7c2	.56	.57	.62	.59	.55	.77	1.		
R8c3	.59	.63	.61	.61	.54	.74	.72	1.	
R9c4	.59	.61	.72	.70	.55	.68	.71	.73	1.
Factor1	.73	.74	.33	.32	.42	.65	.62	.66	.56
Factor2	.28	.29	.82	.79	.52	.53	.54	.52	.64

Footnotes

[1]The cognitive processing model analyses did not include the four Character questions from the SJT. This is because the Character questions were embedded in the other three situations presented in the SJT and because the series of cognitive processing questions was not asked in conjunction with a Character question.

[2]The two SJTs are not identical. One of the items on the scholar SJT asks about a scholarship program specific issue, and therefore was inappropriate for the undergraduate sample. This item was removed from the SJT that was administered to the undergraduate sample, resulting in the undergraduate SJT having one fewer item than the scholar SJT. This difference is not likely to affect the results reported here for two reasons. First, the score reported here is a percent score correct, not a raw score. Second, overall summary scores are used in this comparison. This overall summary score includes 40 and 39 items for the scholar and undergraduate SJTs respectively. This is likely a large enough number of items that the difference is negligible.