HOW DO SITUATIONAL JUDGMENTS TESTS AND SITUATIONAL INTERVIEWS COMPARE? AN EXAMINATION OF CONSTRUCT AND CRITERION-RELATED VALIDITY

by

JAMES S. GUNTER
B.S. University of Central Oklahoma, 2002
M.S. University of Central Florida, 2006

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Psychology
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Fall Term 2010

Major Professor: Barbara A. Fritzsche

ABSTRACT

This study replicated and extended an earlier study by Banki and Latham (2010) and developed an equivalent SJT and SI in order to examine whether the two methods correlated differently with cognitive ability, personality, job experience, and job performance. The results of this study showed that the SJT and SI only correlated .20 and that the correlations for the SI with Extraversion, customer service experience, and overall work experience were significantly different from the correlations for the SJT. Participants felt that the SJT and SI provided the same opportunity to perform one's skills and level of scoring consistency. However, participants felt significantly more anxiety during the SI than the SJT. The practical and theoretical implications of these findings are discussed.

The completion of this dissertation is dedicated to my family who provided me the love and support required to complete this difficult endeavor. Most especially, this dissertation is dedicated to my father, James R Gunter, who passed away in 2007.

ACKNOWLEDGMENTS

I'd like to acknowledge the help, support, and feedback from my committee members Drs. Barbara Fritzsche, Gary Latham, Huy Le, and James Szalma. I'd also like to acknowledge the time, effort, and diligence of my research assistants who executed the study procedure and the support and help of my friends who showed me that hard work and persistence pays off.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

SJT – Situational Judgment Test

SI – Situational Interview

CHAPTER 1: INTRODUCTION

How Do Situational Judgment Tests and Situational Interviews Compare? An Examination of Construct and Criterion-related Validity

Two popular selection tools are the situational judgment test (SJT) and the situational interview (SI). The SJT and SI are each methods that present to applicants or employees situational dilemmas that are typically faced on the job and require the applicant or employee to respond with how they would handle the situation (Latham, Saari, Pursell, & Campion, 1980; Motowidlo, Dunnette, & Carter, 1990). Both are developed using the same job analysis method (i.e., critical incident technique; Flanagan, 1954) and both have been shown to predict performance across a wide range of jobs (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel, Whetzel, Schmidt, & Maurer, 1994). Because of these similarities the SJT and SI have been called, "close cousins" (Weekley & Ployhart, 2006).

Despite these similarities, SJTs and SIs differ in three important ways (Weekley & Ployhart, 2006). First, they differ in terms of how the items are administered. The SJT presents the items in a paper-pencil, video-based, or computer-based format whereas the SI presents the items orally, face-to-face. Second, they differ in the way in which the responses are constructed. In an SJT, between four and six possible response options to the situation are presented and test takers must select one or two of the response options or rate the effectiveness of each. In an SI, no responses options are given to the interviewees. Therefore, they must construct their own response.

Third, the SJT and SI differ in how the responses are scored. In the SJT, subject matter experts assign point values to each response option. The point values across all of the items are

turned into a scoring algorithm that is applied to all test takers in order to calculate an overall score. In an SI, one or more interviewers rate each interviewee's responses in terms of their overall quality using a behaviorally anchored scale.

Based on the similarities and differences, it is unclear whether the SJT and SI are interchangeable situation-based methods. The use of the same job analysis method and the implementation of job-related situational dilemmas would suggest that the SJT and SI are interchangeable methods for measuring job-related constructs and predicting performance. However, different cognitive functions may be required to construct your own response than to choose a response amongst a set of alternatives. Moreover, people may alter their behavior in a situation where someone who is rating their performance is sitting across from them.

As a result of measuring different constructs, SJTs and SIs may correlate with different aspects of performance such as task and contextual performance (Bergman, Donovan, Drasgow, Overton, & Henning, 2008; Borman, White, Pulakos, & Oppler, 1991; Motowidlo, Borman, & Schmit, 1997; Motowidlo & Van Scotter, 1994). This would allow for one or both of these methods to obtain incremental validity over the other, which would substantiate the use of both of these methods in the same selection system.

To my knowledge, only one study (Banki & Latham, 2010) has compared an SJT and SI. Banki and Latham's findings suggest that these two methods are not interchangeable. If SJTs and SIs are not interchangeable then, from a theoretical standpoint, this begs the question of why these two methods don't measure the same constructs and/or predict the same aspects of performance. Furthermore, if SJTs and SIs are not interchangeable then human resources personnel, I/O psychologists, and consultants would have to make important decisions on which

method to use or how they can be used in the same selection system. The current study sought to further the examination and comparison of SJTs and SIs by replicating and extending Banki and Latham's study by examining the correlations of equivalent versions of an SJT and SI with cognitive ability, personality, job experience, and performance.

Overview

SJTs and SIs have a large literature base behind them that supports their use. Both have been shown to correlate with performance (Latham & Sue-Chan, 1999; McDaniel et al., 2001) and have lower mean score differences between demographic subgroups than cognitive ability tests (Bobko, Roth, & Potosky, 1999; DeGroot & Kleumper, 2007; Hough, Oswald, & Ployhart, 2001). SJTs have been shown to provide incremental validity over other job-related constructs such as cognitive ability, personality, and social skills (Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; Morgeson, Reider, & Campion, 2005; O'Connell, Hartman, McDaniel, Grubb, & Lawrence, 2007), SIs have been shown to be more reliable than unstructured or psychological interviews (McDaniel et al., 1994) and to be viewed as more practical to use than past behavior or unstructured interviews (Latham & Finnegan, 1993).

Another major advantage of the SJT and SI is how closely the content of each resembles the job. SJTs and SIs share the same construction method, which is based on a job analysis method called the critical incident technique (Flanagan, 1954). Incidents typically encountered by employees on the job are collected from supervisors and incumbents. The incidents include a description of the incident, the behavior of the person in the incident, and the outcome of the behaviors. Critical incidents are then turned into SJT and SI items by writing a situational

dilemma that reflects the incident, requiring a response to be made in order to solve the dilemma or problem. For a full review of the construction methods for SJTs and SIs see Appendix B.

SJTs utilize two types of responses instructions, "should do" and "would do." A "would do" response instruction has been argued to be related to personality because test takers will respond with how they would tend to behave in job-related situations. Alternatively, a "should do" response instruction is argued to be related to one's knowledge of the right or most effective way to handle a situation. The type of response instruction used in an SJT is important because SIs use a "would do" response instruction. Thus, the findings for SJTs with a "would do" response instruction are most appropriate when comparing to SIs. In the pages that follow research that is specific to SJTs with "would do" response instructions will be highlighted.

Provided below is an example of an SJT item from Motowidlo et al. (1990) that uses a "would do" response instruction:

> You and someone from another department are responsible for coordinating a project involving both departments. This other person is not carrying out his share of the responsibilities. You would…
>
> _____Most likely _____Least likely
>
> 1. Discuss the situation with your manager and ask him to take it up with the other person's manager.
> 2. Remind him that you need his help and that the project won't be completed effectively without a full team effort from both of you.

3. Tell him that he is not doing his share of the work, that you will not do it all yourself, and that if he does not start doing more you'll be forced to take the matter to his manager.
4. Try to find out why he is not doing his share and explain to him that this creates more work for you and makes it harder to finish the project.
5. Get someone else from his department to finish the project.

An example SI item from Sue-Chan & Latham (2004) that uses a "would do" response instruction is provided below:

> You have been assigned to a group to complete an assignment. You feel that one of your group members is not doing any work at all, while others spend too much time gossiping. Overall, you feel that you are carrying all the weight for the group, and that no one else in the group cares very much about the project. Your professor has emphasized to you that the group must solve its own problems. What would you do?

As is evident from construction method and examples above SJTs and SIs are very similar. To compare and contrast these two methods one could consult the literature and draw conclusions from those findings. However, solid conclusions are difficult to make because the SJTs and SIs in the literature were constructed for different jobs, contained different content, and used different samples. In addition, many of the SJTs contained a different type of response instruction.

Arthur and Villado (2008) discussed the importance of comparing selection methods and made a distinction between what they called *predictor constructs* and *predictor methods*. A predictor construct is simply an individual difference variable such as cognitive ability or personality and a predictor method is the way in which predictor constructs are measured (e.g., SJT or SI). Predictor methods like SJTs and SIs are often thought of as flexible enough to be constructed with different content in order to measure a range of job-related constructs (Schmitt & Chan, 2006). Arthur and Villado stated that in order to compare predictor methods the content must be held constant. Once the content is held constant the only differences between the methods would be related to the characteristics inherent in each.

One study by Banki and Latham (2010) compared an SJT and SI by developing both with the same content, thereby, ruling out content as a confound. Their SJT and SI contained the same number of items and were written to measure several dimensions of sales performance including problem solving, team work capabilities, and work precision. They found that the SJT and SI correlated .31 and that both predicted performance similarly (r = .23 and .28 respectively). These findings suggest that the SJT and SI are not interchangeable because the correlation between the two methods is rather low. If the SJT and SI measured the same job-related constructs, then we would expect a much larger correlation between the two. This suggests that 1) they may correlate differently with the same job-related constructs and/or 2) they may correlate with different job-related constructs. Although their SJT and SI correlated similarly with overall performance, if the correlations with job-related constructs are different then the SJT and SI may correlate differently with specific types of performance.

The literature on the correlates of SJTs and SIs does allow for direct comparisons because the SJTs and SIs were built for different jobs and contained different content, but the literature can be useful for understanding whether or not SJTs and SIs correlates at all with certain job-related constructs. In the following sections, a review of the correlations with job-related constructs and performance for SJTs and SIs will be made by focusing on, where available, large-scale meta-analyses. Additional literature and theories will be relied upon in order to guide the comparison of SJTs and SIs and develop conceptual arguments and research questions regarding the correlations with cognitive ability, personality, and job experience and overall, task, and contextual performance.

Cognitive Ability

In an SJT, one's reading comprehension or verbal fluency and one's ability to analyze the situational dilemma and evaluate the effectiveness of each response option is key to responding effectively (Ployhart, 2006; McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006). The same can be said for SIs. Latham and colleagues (Latham et al., 1980; Latham & Skarlicki, 1995) stated that SIs measure one's behavioral intentions. Because SIs are typically conducted in a selection environment where applicants are motivated to score well, applicants will tend to respond to SI questions with what they believe to be effective ways of handling situational dilemmas. One's ability to interpret and understand the situational dilemma, mentally develop a set of possible responses, and evaluate those responses will likely relate to an interviewee's overall effectiveness in each question. As a result, individuals may rely on their cognitive ability to respond to the SJT and SI questions.

A second difference between SJTs and SIs (i.e., the way responses are constructed) is where a difference in the correlation with cognitive ability between the two methods may occur. For example, it may require more cognitive ability to "think on your feet" in an interview than it does to look at a set of response options and make a choice. However, it may require the same cognitive functions to answer SJT and SIs questions because individuals have to make decisions about how to solve the dilemma.

The literature is not clear on which cognitive functions are required for SJTs and SIs. Instead, the literature tends to focus on correlations with broader measurements of cognitive ability. Two meta-analyses have been conducted by McDaniel and colleagues (McDaniel et al., 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007) and both have found that there is a considerable cognitive component to SJTs. The mean observed correlation reported in both studies is about .30. McDaniel et al.'s (2007) follow-up meta-analysis is instructive because they divided the analysis by response instruction type and argued that SJTs with "would do" response instructions would correlate less with cognitive ability than SJTs with a "should do" response instruction. They found a mean observed correlation of .17 for SJTs with a "would do" response instruction and .32 for SJTs with a "should do" response instruction. Despite the smaller observed correlation, these more specific results suggest that there is a cognitive component to SJTs with a "would do" response instruction.

The literature for SIs is much sparser. Huffcutt, Roth, and McDaniel (1996) collected 10 correlation coefficients and reported a mean correlation of .21 (corrected .32) between SIs and cognitive ability. Similar findings were made by Salgado and Moscoso (2002). These authors didn't report the number of coefficients in the analysis or the mean observed correlation because

they collapsed their analyses across different types of interviews, but they do report a mean corrected correlation of .33 for SIs, which is similar to that found by Huffcutt et al. (1996) six years earlier.

Overall, the findings for SJTs and SIs suggest that the correlation with cognitive ability may be quite similar. However, given the relatively small number of coefficients used to estimate the correlation between SIs and cognitive ability and the fact the SJTs and SIs in the literature were not constructed to be parallel to each other renders any solid conclusions to remain unsettled. Research from other areas can be brought in to help inform the comparison of SJTs and SIs.

The educational psychology field has conducted research on the cognitive loading of multiple choice (MC) and constructed response (CR) tests (Martinez, 1999; Rodriguez, 2003; Ward, Fredericksen, & Carlson, 1980). Much of this research is on tests that contain declarative knowledge content such as knowledge of Calculus, Chemistry, Music, and Literature (Rodriguez, 2003). Even considering the explicit focus on declarative knowledge for MC and CR tests the literature can be helpful in understanding the cognitive functions that are likely to be activated in an SJT and SI.

MC tests are analogous to SJTs because a stem and several response options are provided to the test taker. CR tests (i.e, short answer, completion, and essay) are analogous to the SI because both present a stem and require the test taker to construct their own response. The amount and type of information in the stem of the SJT and SI used in the current study will be the same. Consequently, only the way in which the responses are made is where differences in cognitive ability may occur.

Several overarching conclusions have been made on the comparability of MC and CR tests (Martinez, 1999). First, MC items tend to activate more lower-level cognitive functions (i.e., memory and comprehension) in comparison to CR items which activate more higher-level functions (i.e., analysis and evaluation). Martinez (1999) also notes that CR items may tend activate a wider range of cognitive functions than MC items. CR items may be able to activate both lower-level and higher-level cognitive functions whereas MC items may tend to be limited to activating more lower-level functions. However, the extent to which higher-level cognitive functions are activated by MC items is based on complexity. The more complex MC items are the more likely they activate higher-level functions and, therefore, the more likely the overlap between MC and CR items in terms of the cognitive functions they activate will increase. Typical items in SJTs and SIs are of at least moderate complexity because of the detail provided and the complexity of the interactions that take place in the situations (McDaniel et al., 2006). Consequently, both methods may require a range of lower and higher-level functions such as recall, analysis, and evaluation.

An example study that used complex items to examine the cognitive loading of an MC and CR test is by Ward et al. (1980). The test they used is called the Formulating Hypotheses (FH) test, which presents situations that are typically encountered by behavioral scientists when conducting research and interpreting data. The FH items utilized a CR format and required test takers to read a brief description of a study, examine a graph or table of results, and write possible explanations (i.e., hypotheses) that may account for the study results. The items didn't have a single right or wrong answer. Rather, the hypotheses generated by test takers are

considered to differ in terms of overall quality. These characteristics make the FH test very similar to SJTs and SIs.

Ward et al. (1980) constructed equivalent MC and CR versions of the FH test by developing a set of items that were based on the same content domain, but not the exact same content. To ensure that both tests versions were as parallel as possible, they randomly assigned items to either the MC version or CR version. In the CR version test takers developed as many possible hypotheses that they could whereas in the MC version test takers were given a set of nine possible hypotheses and they were told to choose the ones that were likely to explain the results. Three types of quality scores were calculated for the hypotheses that were generated. Participants completed both the MC and CR versions of the FH test and their quality scores were correlated with measurements of specific cognitive functions such as quantitative reasoning, induction, logical reasoning, and cognitive flexibility.

The results of their study are very informative. They found that two of the quality scores for the MC version correlated about .30 with the same quality scores in the CR version. These findings are similar to those found by Bank and Latham (2010) who reported a correlation of .31 between an SJT and SI. Ward et al. (1980) also found that the correlation between these quality scores and the cognitive functions listed above were the same across the MC and CR versions (from .20 to .40). Similarly, they found that the correlation between the quality scores and scores on the quantitative section of the GRE were nearly identical (r = ~ .20) across the MC and CR version. Lastly, they found the same pattern with quality scores and content domain knowledge.

Overall, they found no large differences in the cognitive ability and knowledge saturation between the two versions. Ward et al. (1980) concluded that induction, logical reasoning, and

content domain knowledge are both used in the FH test regardless of the version the participants completed. In other words, the CR and MC quality scores are essentially the same in terms of providing information about cognitive ability.

Research examining correlation patterns is not the only research on MC and CR formats. Other research, summarized by Rodriguez (2003), has been conducted using factor analysis. Rodriguez cited four studies that used confirmatory factor analysis that found that MC and CR items loaded on two separate method factors, yet these method factors correlated, on average, .86. Of particular interest is that these studies did not use the same stem (content) for the MC and CR items. Thus, the correlations between the factors may be higher if the stems were the same. These findings support Wainer and Thissen's (1993) earlier discussion of MC versus CR formats which stated that although MC and CR items may activate somewhat different cognitive functions, the functions they do activate are highly correlated with each other and, therefore, make the two formats highly interchangeable.

Overall, the literature summarized above suggests that 1) SJTs and SIs do correlate moderately with measurements of cognitive ability, and 2) the educational literature, which has examined analogs of SJTs and SIs, suggests that MC and CR items tend to activate somewhat similar cognitive functions and that, overall, similar inferences can be made from both types of items about one's cognitive ability. However, this research doesn't tell us whether an SJT and SI with the same content correlate the same or differently with cognitive ability. The educational literature is helpful, but has largely focused on school-related declarative knowledge (see Ackerman, & Smith, 1988; Bennett, Rock & Wang, 1991; Hancock, 1994) rather than tests that include items with job-related situational dilemmas. Moreover, Ward et al.'s (1980) study, which

provides the closest evidence, did not use a test that utilized the same type of response instruction found in SJTs and SIs. Because there is no direct evidence on this issue, the following research question was addressed:

Research question #1: Do equivalent versions of an SJT and SI correlate significantly differently with cognitive ability?

Personality

A second individual difference variable of interest for comparing SJTs and SIs is personality. Personality is relevant to SJTs and SIs because they both try to understand and predict how someone would behave in various situations, and personality is core to one's behavior. The Big 5 is the most commonly-used personality taxonomy in research on SJTs. McDaniel and colleagues have completed two meta-analyses on the correlation between SJTs and the Big 5 personality traits. In their first meta-analysis of SJTs, McDaniel and Nguyen (2001) report a mean correlation of .26 for Conscientiousness, .25 for Agreeableness, .31 for Neuroticism, .06 for Extraversion, and .09 for Openness to Experience. In an updated meta-analysis that included almost three times as many correlation coefficients as their first study McDaniel et al. (2007) found a mean correlation of .23 for Conscientiousness, .22 for Agreeableness, .19 for Neuroticism, .13 for Extraversion, and .11 for Openness to Experience.

Similar to their examinations of cognitive ability, McDaniel et al. (2007) further divided their analysis by response instruction. They found that SJTs with a "would do" response instruction correlate with personality. Specifically, they found a mean correlation of .30 for Conscientiousness, .33 for Agreeableness, .31 for Neuroticism, .07 for Extraversion, and .09 for Openness to Experience. Overall, the traits Conscientiousness, Agreeableness, and Neuroticism

show the highest and most consistent correlations with SJTs while the traits Extraversion and Openness to Experience show lower correlations (for a full set of correlation coefficients see Table 4).

There aren't as many data for SIs as there are for SJTs, which has made it difficult for researchers to find many significant correlations between SIs and personality (see Table 4). Two meta-analyses (i.e., Sagado & Moscoso, 2000; Cortina, Goldstein, Payne, Davison, and Gilliland, 2000) have been conducted that have examined the mean correlation between interview scores and measurements of personality. Both of these studies, however, were performed using data that were gathered from both behavioral and situational interviews. Salgado and Moscoso's (2002) found no correlations with the Big 5 personality traits above .10. Cortina et al. (2000) accumulated only nine correlations from highly structured interviews and found a mean observed correlation of .21 with Conscientiousness.

These somewhat inconsistent meta-analytic results are also seen at the individual study level. Roth, Van Iddekinge, Huffcutt, Eidson, and Schmit (2005) developed an SI for sales associates in a retail organization and found that SI scores did not correlate significantly with any of the Big 5 traits. Most of the correlations they found were around .10 or less. In another study using retail sales associates, DeGroot and Kluemper (2007) correlated SI scores with two types of personality measurements. These authors measured the personality traits Extraversion, Agreeableness, and Conscientiousness with two types of frame-of-reference, how you are generally and how you are at work. This distinction is important for the current review because although recent research has emphasized the importance of frame-of-reference effects (Lievens, De Corte, & Schollaert, 2008; Schmit & Ryan, 1993; Schmit, Ryan, Stierwalt, & Powell, 1995;

Smith, Hanges, & Dickson, 2001), most, if not all, research on SJTs and SIs used the general frame of reference.

DeGroot and Kluemper's (2007) use of two frames of reference resulted in different findings for personality. They found that the SI correlated significantly with Extraversion in both the general (.22) and work (.28) frame of reference. However, the authors found that the SI did not correlate significantly with Conscientiousness in the general frame of reference (.13), but it did correlate significantly in the work frame of reference (.17). Lastly, the authors found that in neither the general nor work frame of reference (-04 and -.01, respectively) did the SI correlate with Agreeableness. Taken together, DeGroot and Kleumper's study did not provide strong support for a correlation between an SI and personality when using a general frame of reference.

The relatively sparse research, as well as the research that combined behavioral and situational interviews into the same analyses, makes it difficult to draw solid conclusions about the magnitude of the correlation between SIs and personality and how that compares to the findings for SJTs. The extant research, however, suggests that the correlations between SIs and personality are not as large as those for SJTs. The typical correlation between SJTs and many of the personality traits hover around .30 whereas the typical correlation between SIs and personality hovers around .10. For a full review and analysis see Appendix C.

The existing findings on the correlations for the SJT and SI with personality may seem somewhat counterintuitive considering the interpersonal nature of the SI. In addition, previous research has show that highly structured interviews like the SI are built with personality-related dimensions (Huffcutt et al., 2001). However, these results suggest that there may be something about the SI that is different from the SJT that prevents the SI from correlating with personality

as highly as does the SJT. Among the possible explanations for this are the nature of the environmental cues in an interview, the ability to tailor responses according to those environmental cues, and the ability to use impression management tactics.

Situational cues and situation strength. Mischel (1973) argued that across situations there are different cues that communicate which behaviors are likely to lead to higher performance. People perceive these cues and develop behavior-performance expectancies. He stated that individuals alter their behavior according to the expectancies they develop. In instances where the cues are clear the expectancies are likely to be very strong and, therefore, one's behavior will be altered a great deal. This is termed a strong situation (Mischel, 1973). In instances where the cues are rather ambiguous the expectancies are likely to be weak, and therefore, one's behavior will be altered much less or not at all. This is termed a weak situation (Mischel, 1973).

Trait Activation Theory (TAT; Tett & Burnett, 2003) makes similar assertions. TAT states that situational cues signal which traits are relevant to the situation. Lievens, Chasteen, Day, and Christiansen (2006) gave an example where aggression would not be a trait relevant to a religious service because the situation does not provide cues related to aggression. The two concepts of situation strength and trait relevance are distinct, but combine to provide an explanation for the restricted range of exhibited behavior in social situations. Tett and Burnett (2003) provided an analogy which states that while a radio frequency is distinct from volume both combine to affect the overall ability to listen to music. In essence, situations signal which traits to exhibit versus others (i.e., radio frequency) and the strength of the situation signals the intensity of the behaviors that should be exhibited (i.e., volume). For example, a situation may signal that Agreeableness is relevant, and that acting in a very courteous manner will lead to

positive outcomes. Thus, while trait-relevant cues will lead individuals to exhibit the same trait(s) the strength of the situation will lead to a reduction in the exhibited individual differences in that trait.

Two of the three differences between SJTs and SIs (i.e., method of administration and response construction) are likely to be points where trait-relevance and situation strength come into play. First, the social interaction that takes place in the oral administration of an SI may provide cues or information regarding which job-related traits are relevant because the interviewee is being judged on their employability (Caldwell & Burger, 1998). Courteousness, achievement orientation, sociability, emotional stability, and dependability are likely to be trait-relevant behaviors. Similar cues are likely to be present in the SJT because applicants would likely believe that the SJT is measuring something that is job-related (i.e., face validity; Chan & Schmitt, 1997). However, the social interaction in the SI should be able to convey more and richer information than the SJT because the only information conveyed in an SJT is that which is contained in the item whereas information is constantly being sent and received in the interview. As a result, the cues may be more plentiful and more unambiguous in the SI than in the SJT and, therefore, produce a much stronger situation.

Second, the way in which responses are constructed is the main point where trait activation and situation strength are most relevant. In the SI, interviewees have the opportunity to construct their response in any way they choose, and will likely alter their responses in order to perform well in the interview. In addition, interviewees can also exhibit other behaviors such as impression management (IM) tactics that may affect how someone perceives them, which, in turn, may affect their interview scores (Ellis, West, Ryan, & DeShon, 2002;Kacmar, Delery, &

Ferris, 1992; Kristof-Brown, Barrick, & Franke, 2002; Lievens & Peeters,2008; McFarland, Ryan, & Kriska, 2003; Peeters & Lievens, 2006; Stevens & Kristof, 1995; Van Iddekinge et al., 2007). By contrast, SJTs do not allow for the content of the response to be altered and do not allow IM tactics to affect their scores. Taken together, responses to SJT and SI items, as well as the behaviors that are expressed that are beyond the SJT and SI, may be a particularly relevant and strong mechanism that restricts the range of expressed personality.

Response distortion. A third difference between SJTs and SIs (i.e., method of response construction) is where the effects of response distortion may come into play. In contrast to SIs, test takers completing an SJT cannot alter the content (i.e., expressed behavior) of the responses. Although they can evaluate each response option according to its similarity to what they would do in that situation and alter their choice, the expressed behavior within the response option cannot be changed. This is not the case for the SI. Interviewees can alter their response in any way they choose to in order to score as highly as possible and convey a positive image. Conceptually, there is greater opportunity to distort one's responses in the SI than in the SJT.

Hooper, Cullen, and Sackett (2006) summarized the research on how susceptible SJTs are to response distortion. In addition, they cite two studies that show how response distortion affects correlations with personality variables. This research utilized "honest" vs. "fake" conditions as well as examined data for applicants and incumbents because applicants are more likely to distort their responses.

One study cited in Hooper et al.'s (2006) summary was by Nguyen, McDaniel, and Biderman (2002). Nguyen et al. had participants complete an SJT with a "would do" response instruction and a personality test. They instructed them to respond to both the SJT and

personality test honestly or as if they were a customer service representative applicant (i.e., fake condition). They found that when participants responded as if they were an applicant to both the SJT and personality test that that the correlations were not attenuated in comparison to when they completed the SJT and personality test honestly. In fact, for each trait the correlations in the fake condition were larger in magnitude. Ployhart, Weekley, Holtz, and Kemp (2003) administered a paper-pencil SJT to a sample of incumbents and applicants of a tele-service job. They only measured Agreeableness, Conscientiousness, and Emotional Stability (i.e., Neuroticism) and found that for two of these three traits the correlations were approximately the same across the two samples. The correlations for Conscientiousness dropped from .43 in the incumbent sample to .28 in the applicant sample, but the correlation in the applicant sample remained statistically significant. While two studies aren't enough to settle this issue, these findings suggest that when individuals fake on an SJT it may not significantly affect the correlations with personality.

As is typical in much research on interviews there is not much, if any, research on response distortion that is isolated for SIs. However, one study provides results that are somewhat informative. Van Iddekinge, McFarland, and Raymark (2007) conducted a study that showed how a strong, trait-relevant situation, such as an interview, can restrict the range of expressed personality-related behavior. They used a structured interview containing both behavioral and situational questions and told one group of participants to act like they were applicants applying for a customer service manager position. The other group was given no instructions on how to act, and served as a control group.

Their results are clear. The mean rating for those in the control group was 3.03 whereas the mean rating for those in the "applicant" group was 4.61. This is a standardized mean

difference of 2.0. Also, they found that measurements of personality correlated significantly with interview scores for those in the control group, but they did not correlate significantly with scores for those in the "job applicant" group. On average, the correlations in the "job applicant" group were .10 lower than those in the control group. These findings show that people can distort their responses and that the distorted responses may have lead to individuals' personality levels being obscured.

The little research there is for SJTs and SIs doesn't provide clear evidence of the extent to which response distortion affects correlations with personality. Two studies (Nguyen et al, 2002; Ployhart et al., 2003) on SJTs showed similar results. However, it is difficult to compare those findings with those of Van Iddekinge et al.'s (2007) because their interview was not strictly an SI. Even if it was, it would not have contained the same content as the SJTs described above.

Impression management. In addition to distorting one's responses to improve one's score, IM tactics are likely to be exhibited in an interview, leading to a restriction of individual differences, and affecting one's score (Ellis, West, Ryan, & DeShon, 2002;Kacmar, Delery, & Ferris, 1992; Kristof-Brown, Barrick, & Franke, 2002; Lievens & Peeters,2008; McFarland, Ryan, & Kriska, 2003; Peeters & Lievens, 2006; Stevens & Kristof, 1995; Van Iddekinge et al., 2007). IM tactics are most relevant to the way in which responses are generated and scored. For both factors, IM is only related to the SI because IM tactics cannot be demonstrated in the SJT; there is no social interaction. Also, IM tactics cannot affect an SJT scoring algorithm. However, it is possible for IM tactics to affect interviewer's perceptions of interviewees.

IM is defined as an applicant's attempt, consciously or unconsciously, to control one's image in social interactions (McFarland et al., 2005; Schlenker, 1980). There are two main

categories of IM tactics, assertive and defensive (McFarland et al., 2005). Assertive tactics include other-focused and self-focused tactics. Other-focused tactics are those where the interviewee is being complementary or conforms his or her attitudes and opinions to be similar to the interviewer's. Interviewees use self-focused tactics to promote their skills and abilities, take credit for positive outcomes, and to describe relationships with other important persons. Finally, defensive tactics are basically excuses designed to deny responsibility for negative outcomes. Other IM tactics include non-verbal tactics such as smiling, nodding, eye contact, and hand movements.

Peeters and Lievens (2006) argued that interview behavioral and situational interviews provide different opportunities for IM tactics. They found that other-focused tactics were more often used in an SI compared to self-focused and defensive tactics. Also, they argued that certain personality traits are linked to the different types of IM tactics. For example, they linked Extraversion with self-focused tactics because Extraversion is defined as being energetic, sociable, and assertive. They argued that Extraverts are more likely to promote themselves and their accomplishments. Agreeableness was linked with other-focused tactics because Agreeableness is about being courteous, likable, and good-natured. Therefore, highly agreeable people try to make themselves likable. Finally, they linked Neuroticism with defensive tactics because Neuroticism is defined as lacking self-esteem and feeling anxious. Therefore, those with high levels of Neuroticism attempt to repair any possible negative images that the interviewer may have of them.

Peters and Lievens, (2006) gave participants either a behavioral and situational interview and instructed participants to either respond honestly or to make the best impression. They

collapsed ratings for both interview types and, for those who were instructed to make the best impression, they found a significant negative correlation between Agreeableness and other-focused tactics and a significant positive correlation between Neuroticism and defensive tactics. Thus, those who have low levels of Agreeableness tend to express an image that they are likable and those who have high levels of Neuroticism tend to repair their image by using excuses and justifications. The result was that the range of these exhibited behaviors was restricted. In total, the trait activation and situation strength as well as the IM research suggest that the SI is likely to cause someone to alter their behavior and obscure one's personality traits.

Finally, the scoring of an SI may be affected by the use of IM tactics. Even though making interviews structured increases inter-rater reliability (Weisner & Cronshaw, 1988), systematic bias still affects the subjective ratings of interviewees (Imada & Hakel, 1977; Lievens & Peeters, 2008; Stewart, Dustin, Barrick, & Darnold, 2008). Recall that IM tactics are exhibited for the purpose of conveying a positive image. Therefore, IM tactics can be used to systematically bias interviewers beyond the content of the response itself. Peeters and Lievens (2006) found that other-focused IM tactics and non-verbal tactics correlated significantly and positively with performance in an SI. In another study, Lievens and Peeters (2008) found that overall SI scores made by student interviewers correlated positively with the use of defensive and non-verbal IM tactics, and overall SI scores made by professional interviewers correlated positively with the use of other-focused and non-verbal IM tactics.

Taken together, the studies by Peeters and Lievens (2006) and Lievens and Peeters (2008) show that SIs can be biased by IM tactics. By contrast, an SJT scoring algorithm always

scores a given response option in the same way. In other words, the SJT cannot be affected by behaviors unrelated to, or in addition to, the content of the responses.

The conceptual arguments made throughout this section suggest that one's personality traits are likely to be masked in an SI from distorted responses and IM tactics because of trait activation cues and situational strength. As noted previously, research needs to be conducted to fully estimate the effects of response distortion. In addition, the extant research has not compared SJTs and SIs that have the same content. Therefore, the following research question was addressed:

Research question #2: Do equivalent versions of an SJT and SI correlate significantly differently with each of the Big 5 personality traits (i.e., Agreeableness, Extraversion, Conscientiousness, Neuroticism, and Openness to Experience)?

Job Experience

SJTs and SIs are each conceptually linked to job experience because as employees work at a job they encounter a range of situations, and through dealing with those situations they likely gain knowledge about how to properly handle them. In other words, as employees make choices about how to solve job-related problems, they will figure out which ones tend to work and which ones don't. Therefore, to the extent to which the situational dilemmas are similar to applicant's experiences the more likely they will respond effectively.

The research for SJTs is mixed in terms of the correlation with experience (see Table 4). Clevenger et al. (2001) report non-significant correlations of .01, .03, and -.13 across three samples. However, Weekley and Ployhart (2005) found significant correlations between SJT scores and general work experience (.21) with job tenure (.13). Overall, McDaniel and Nguyen's

(2001) meta-analysis showed a non-significant mean observed correlation of .05 between SJT scores and measurements of job experience.

The research for SIs is similarly mixed. Three studies (Day & Carroll, 2003; Huffcutt, Weekley, Wiesner, DeGroot, & Jones, 2001; Pulakos & Schmitt, 1995) found near-zero correlations (.00, .08, and .04 respectively). However, Conway and Peneno (1999) obtained a significant correlation of .29 and Gibb and Taylor (2003) obtained a significant correlation of .49. Both of these correlations are much larger in magnitude than those found by Weekley and Ployhart (2005) who examined SJTs. Thus far, the research for both SJTs and SIs shows that they can correlate with measurements of experience (see Table 4 and Appendix C). However, the maximum correlation between scores and experience may be higher for the SI than for the SJT.

An explanation for why this might be the case may lie in the way in which someone constructs their response in an SJT and SI. For example, in both the SJT and SI a test taker may analyze the situational dilemma and evaluate the extent to which the situation is similar to experiences they recall from memory. Yet in the SJT, when generating a response, test takers cannot alter the content of the response options. In other words, specific and detailed information related to one's experiences cannot be communicated. The result is that the response options, and, therefore, the responses, in an SJT only represent an approximation of one's actual experiences.

The SI, however, presents a greater opportunity for more, detailed information related to one's experiences to be communicated because interviewees can construct their own responses. Pulakos and Schmitt (1995) conducted a study with employees from a large federal organization, some of which had up to six years of experience within their current position. They noted that

some interviewees were able to provide information about how they would handle possible contingencies that may have been present in the situation. This large experience base seems to have been used to create the longer, more complex, and detailed responses to the SI questions.

The data for job experience are among the fewest for SJTs and SIs across each of the constructs discussed so far. Moreover, the findings are more mixed than the findings noted above. Some studies show that SJTs and SIs can correlate with job experience, possibly .20 to .30. However, other studies show almost zero correlation. The research is not clear on what the magnitude of the correlation with job experience is for SJTs and SIs let alone what it might be if they had the same or similar content. The following research question was addressed:

Research question #3: Do equivalent versions of an SJT and SI correlate significantly differently with job experience?

Overall Job Performance

SJTs and SIs each have an ample amount of research that has been directed at estimating the correlation with performance (see Appendix D). This research shows that SJTs and SIs correlate very similarly with performance. McDaniel and colleagues' first meta-analysis (McDaniel et al., 2001) of the criterion-related validity of SJTs reported a significant mean observed correlation of .26, and their most recent meta-analysis (i.e., McDaniel et al., 2007) showed a significant mean observed correlation of .20. McDaniel et al.'s (2007) 80% credibility interval ranged from .13 to .39. Thus, criterion-related validity of SJTs generalizes across a range of different types of jobs.

SIs have also been shown to predict job performance. Latham and Sue-Chan (1999) collected 20 correlation coefficients and found a mean observed correlation of .29. Taylor and

Small's (2002) meta-analysis contained 30 coefficients and obtained a mean observed correlation of .25. They also reported a lower-bound credibility value of .09.

Despite the meta-analytic findings, some individual studies reported results that suggested that job complexity may moderate the correlation between SI scores and job performance (Huffcutt, Weekley, Wiesner, DeGroot, & Jones, 2001; Pulakos & Schmitt, 1995). Specifically, these two studies showed that an SI did not correlate with performance for a more complex job. Taylor and Small (2002) and Huffcutt, Conway, Roth, and Klehe (2004) followed up on these findings and examined whether job complexity moderated the correlation with job performance. Their studies showed conflicting results. Taylor and Small (2002) found no evidence of a moderating effect whereas Huffcutt et al. (2004) did. Although both studies utilized the same basic framework for making job complexity categorizations (Hunter, Schmidt, & Judiesch, 1990), the difference in the findings for these studies is likely due to the way in which the authors categorized the jobs.

This study collected data from employees from a range of customer service jobs. Based on a previously used job complexity framework (Hunter et al., 1990) it is likely that the customer service jobs within this study are either low or medium complexity. The range of the mean observed correlation reported for low and medium complexity jobs by Taylor and Small (2002) and Huffcutt et al. (2004) was .24 to .31. Taken together, the mean observed correlations reported in this section for SJTs and SIs as well as those for SIs in low to medium complexity jobs are very similar. As a result, the complexity of the jobs used in this study (i.e., customer service) is not likely to be an explanation for any differences in the correlations with performance for the SJT and SI.

There are no obvious reasons why the differences between the SJTs and SIs would cause a significant difference in the correlation with job performance. The extant literature shows that both SJTs and SIs predict job performance, and the magnitudes of the correlations are very similar. However, recall that the content of the SJTs and SIs in the literature is not the same. As a result, it is not clear if an SJT and SI with the same content predict performance similarly. Therefore, the following research question is addressed:

Research question #4: Do equivalent versions of an SJT and SI correlate significantly differently with overall job performance?

Task and Contextual Performance

Typically, research investigating the predictive ability of various employment tests uses overall job performance as the criteria. However, more recent examinations of the performance domain have highlighted that overall job performance can be divided into two specific types of performance. Motowidlo, Borman, and Schmit (1997) developed a theory of job performance wherein task and contextual performance are two important dimensions of job performance. They defined task performance as the activities that transform raw materials into goods and services. Such activities include selling merchandise and operating a piece of equipment. Contextual performance was defined as activities that promote the viability of the social and organizational network. Such activities include helping and cooperating with others, supporting the organization, and persisting with enthusiasm.

Their theory includes a set of linkages between individual difference variables and task and contextual performance. Specifically, cognitive ability and job experience are argued to be the antecedents of task performance because both are involved in the accumulation of job

knowledge and skill. Hunter (1983) related cognitive ability, job knowledge, and performance and found that cognitive ability predicted performance only through its effects on job knowledge. Borman et al. (1991) also found similar results. Schmidt, Hunter, and Outerbridge (1986) conducted a similar analysis for experience and found that experience is also influential in the development of knowledge and skill.

Personality is argued to be the antecedent of contextual performance (Motowidlo et al., 1997). For example, Conscientiousness contains the facets of achievement motivation and dependability, which is conceptually linked with the dimension persisting with enthusiasm. Similarly, the courteousness facet of Agreeableness and the sociability facet of Extraversion are linked with the dimension helping and cooperating with others.

In an examination of the theory, Motowidlo and Van Scotter (1994) found that task and contextual performance are distinct from each other and independently predict performance. In addition, they found that experience better predicted task performance whereas personality better predicted contextual performance. Their results did not show that cognitive ability better predicted task performance. However, a later study by Bergman et al. (2008) did provide support for the link between cognitive ability and task performance.

For SJTs and SIs, one would expect the same conceptual and empirical linkage between cognitive ability, personality, SJTs, and task and contextual performance. Recall that SJTs have been shown to correlate with significantly cognitive ability and personality (McDaniel et al., 2001; 2007). The research summarized in previous sections noted that SIs correlate significantly with cognitive ability, but there has not been much support for a correlation with personality. The cognitive ability and personality findings suggest that SJTs should correlate with both task

and contextual performance whereas SIs should correlate with task performance, but not contextual performance.

SJTs have been shown to predict task and contextual performance (Morgeson et al., 2005; O'Connell et al., 2007). However, the magnitudes of the correlations were quite different. O'Connell et al. (2007) found that their SJT correlated significantly with task (.14), and contextual performance (.10). Morgeson et al. (2005) found similar results and reported significant correlations with task (.36), and contextual performance (.32).

With regard to SIs, research has been conducted that showed a correlation with performance criteria that may be more related to task performance than contextual performance. For example, Latham et al.'s (1980) SI correlated significantly with performance (.50 and .46) of saw mill workers. Day and Carroll (2003) and Sue-Chan and Latham (2004) both found that their SIs correlated significantly with academic performance (.37 and .26, respectively). Some studies have looked at more contextual types of performance and found results that are not what would be expected given the research on personality. Sue-Chan and Latham (2004) built an SI to measure teamplaying behavior and found that it predicted performance operationalized as observed teamplaying behavior (.32). Similarly, Klehe and Latham (2005) found a significant correlation of .41 with teamplaying behavior. Finally, Latham and Skarlicki (1995) found that their SI significantly correlated (.30) with faculty member organizational citizenship behaviors.

The research above shows that SJTs and SIs can correlate with task and contextual performance. Many of the correlations are similar in magnitude except for the findings from the study by O'Connell et al., (2007). Although the magnitudes of many of the correlations are similar, the SJTs and SIs in this research had different content and were used for different jobs.

Thus, it is difficult to draw direct comparisons of the correlations with task and contextual performance. Therefore, the following research question was addressed:

Research question #5: Do equivalent versions of an SJT and SI correlate significantly differently with task and contextual performance?

Procedural Justice and Test Anxiety

The types of reactions that are most relevant to the selection context are those related procedural justice because it deals with the perceived fairness of the methods used to make selection decisions. Bauer and Truxillo (2006) indexed the aspects of procedural justice that are most relevant to SJTs. These include job-relatedness, opportunity to perform, consistency of administration, feedback, and two-way communication. This was considered to be a useful taxonomy because of the similarity between SJTs and SIs. However, in this study, consistency of administration was modified to refer to the consistency of scoring because an SJT is objectively scored through the use of a scoring algorithm that is applied in the same way to everyone whereas the SI is subjectively scored by raters.

Feedback is the timeliness with which test performance is communicated to the test takers. This aspect is more related to the larger selection system rather than the methods and, therefore, is not relevant to the scope of the study. Two-way communication is the extent to which the test taker and testing personnel communicate with each other and/or the extent to which the test taker has input into the selection process. This aspect is not relevant to the SJT because the SJT has very little, if any, two-way communication. Thus, the comparison isn't a legitimate comparison of the reactions to two-way communication. In addition, the current study does not allow for input into the methods used. Lastly, it is unlikely that the SJT and SI will be

perceived differently in terms of job-relatedness because they will be given the same situational dilemmas in both methods.

The way in which responses are made between the SJT and SI may cause differences in perceptions for opportunity to perform and consistency of scoring. For example, because the SJT has responses already generated, test takers may feel that they have less opportunity to show their skills and abilities than if they were allowed to freely generate their own responses. Similarly, because each interviewee generates their own response to the SI and the interviewer subjectively rates their responses they may perceive that the SI is not as consistently scored as the SJT.

Research question #6: Are equivalent versions of an SJT and SI perceived to provide significantly different opportunities to perform one's skills?

Research question #7: Do participants feel there is a significant difference scoring consistency across equivalent versions of an SJT and SI?

Banki and Latham (2010) measured the experienced test anxiety for those completing both an SJT and SI. They did not find a difference in the experienced anxiety in the SJT versus the SI. These findings are unexpected because of the interpersonal nature of the SI. Banki and Latham noted that the participants in their study had not completed any kind of selection test before. As a result, the participants may have felt similarly anxious when completing the SJT and SI because both methods were novel to them. Because of the unique nature of Banki and Latham's sample the following research was addressed:

Research question #8: Do individuals experience significantly different levels of test anxiety when completing equivalent versions of an SJT and SI?

CHAPTER 2: METHOD

Method

Participants

One hundred seventy-four students from a large Southeastern university participated in the study for course credit. The basic design was a concurrent validation study. Therefore, only students who had a customer service job were eligible to participate. A customer service job was defined as any job where an employee frequently interacts with customers in order to provide a particular service, handle complaints, solve problems, fulfill requests, and/or answer questions. Participants' customer service jobs included, but were not limited to, restaurant servers, cashiers, hosts and hostesses, customer service representatives, theme park attendants, retail salespersons, bank tellers, and front desk clerks. On average, participants had just under four years of experience in customer service positions (M = 46.49 months, SD = 40.80). Fifty-six percent of the participants were between 18 and 20 years of age, 39% between 21 and 25, and the remaining 5% were 26 or older. Seventy-three percent of the participants were female, 58% White/Caucasian, 18% Hispanic/Latino, 11% Black/African American, 5% Asian/Pacific Islander, and 8% Mixed/Other. Thirty-one percent of the participants were Freshman, 25% Juniors, 22% Seniors, 17% Sophomores, and 5% Fifth-year Seniors.

Procedure

All volunteer participants were treated in accordance with the Ethical Principles of Psychologists and Code of Conduct (American Psychological Association, 1992). Participants were instructed that the purpose of the study was to assess the relative effectiveness of various

tests used to predict performance in customer service jobs. A within-groups design was used to answer the research questions and was conducted across two sessions.

In session 1, participants first completed the measurement of cognitive ability followed by the measurement of personality and job experience. Participants then completed one of the situational tests and ended session 1 with the measurements of test anxiety and procedural justice. The measurement of text anxiety was always given immediately after the administration of the situational test. Session 2 consisted of participants completing the other situational test and the measurement of test anxiety and procedural justice.

The order of the SJT and SI was counterbalanced in order to account for testing and carryover effects by randomly assigning participants to one of two conditions. In condition 1, participants completed the SJT in session 1 and the SI in session 2. In condition 2, participants completed the SI in session 1 and the SJT in session 2. Because the content of the SJT and SI was exactly the same, session 1 and session 2 was separated by no less than one week and no longer than three weeks. To further safeguard against testing and carryover effects participants were not told about the nature of the test they were going to complete in session 2.

In order to simulate the conditions of a real selection environment all participants were instructed at the beginning of session 1 to complete the situational tests and measurements as if they were a real applicant applying for a customer service position and as if they wanted the job. To further represent a high-stakes environment, participants were given a monetary reward if they were the highest scorer on either situational test. Participants completing the SI were instructed that their responses should only be about one to two sentences in length. This served to make the SJT and SI comparable in terms of the content of the responses. If participants were

allowed to respond in the SI without any restrictions then the responses would likely contain more information, possibly alter the constructs that are measured, and make the SJT and SI less equivalent.

Job performance data were collected separately from the participant's participation in the study. I collected participants' supervisor email contact information in session 1. An email was sent to the supervisor with a link to a survey website that contained the job performance survey. Because job performance ratings were collected anonymously from participants' supervisors I was not able to train the supervisors to minimize rating errors. However, I instructed supervisors at the beginning of the survey to make their ratings honestly and accurately. Participants had a range of customer service jobs. Therefore, virtually none of the participants had the same supervisor. Only two participants had the same job and were rated by the same supervisor. As a result, interrater reliability estimates were not possible.

Several steps were taken in order to increase the survey response rate (Newman, 2010; Roth & BeVier, 1998). First, the emails were personalized for each supervisor and their subordinate (i.e., the participant). In the email, I used the supervisor's name, referred to the subordinate by name, and noted that the subordinate gave us his/her supervisor's name. Second, the email explained that the purpose of the study was to assess the relative effectiveness of various tests that are used to predict customer service performance. Not only did this provide information about the reason for the email, but also it acted as an incentive to complete the survey. Third, supervisors were told that the ratings they provided were completely confidential and that the subordinates did not have access to the ratings. Fourth, supervisors were given one week to complete the ratings. If the supervisor did not respond within the first week, reminder

emails were sent. Finally, supervisors were instructed that if they provided ratings then the subordinate would receive extra credit for the study. This provided an incentive for the participant to follow up with their supervisor to ensure that they provided ratings.

Measures

Cognitive ability. The Wonderlic Personnel Test-Revised (Wonderlic, 2002) was used to measure general cognitive ability. This 12-minute, 50-item commercially published measurement of cognitive ability has been shown to have high test-retest (.82 to .94) and internal consistency (.88) reliability (Wonderlic, 2002), positive validity results (Bell, Matthews, Lassister, & Leverrett, 2002; Dodrill & Warner, 1988), and is widely used in I/O research (Chan, 1997; Conway & Peneno, 1999; Sue-Chan & Latham, 2004; Wright, Kacmar, McMahan, & Deleeuw, 1995).

Personality. A free online instrument called The International Personality Item Pool (IPIP; Goldberg, 1999) was used to measure the Big Five personality traits. The IPIP not only has been used broadly in the I/O literature, but also specifically for SJTs (Colbert, Mount, Harter, Witt, & Barrick, 2004; Heggestad, Morrison, Reeve, & McCloy, 2006; LeBreton, Barskdale, Robin, & James; 2006; Lievens et al., 2008; Mount, Barrick, Scullen, & Rounds, 2005; Ng, Ang, & Chan, 2008; Oswald, Schmitt, Kim, Gillespie, & Ramsay, 2004). Ten items each were used to measure Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience. Coefficient alpha for the five traits were .90, .83, .81, .89, and .77 respectively.

Job experience. The measurement of job experience was divided into two types, general and specific (Weekley & Jones, 1997; 1999; Weekley & Ployhart, 2005), and was operationalized in months rather than years in order to create more variance in the

measurements. General work experience was measured with two items (e.g., "what is longest time you have worked for any one organization?" and "How many months have you worked in either full-time or part-time employment?"). Specific work experience was also measured with two items (e.g., "How many months have you worked in your current position?" and "How much experience do you have in customer positions?").

SJT. The SJT used in this study was a commercially available customer-service based SJT, and was developed using elements of both a content-oriented and construct-oriented approach. Two I/O psychologists with over 20 years of combined experience developed the SJT. They used the method outlined by Motowidlo et al. (1990) as a guideline in order to identify the tasks and duties related to customer service as well as the dimensions related to customer service performance. The SJT was originally video-based, but was modified to a paper-pencil format so that it could be delivered on a computer.

The purpose of the SJT was to measure an individual's overall ability to perform across different customer service situations. A job analysis was performed by the I/O psychologists across two companies in the automotive industry. The jobs analyzed ranged from cashiers to body shop advisors, rental car clerks, and receptionists. Based on the tasks and duties identified the dimensions positive customer relations, discovering customer needs, responding to customer needs, anticipating customer needs, working together to meet customer needs, and ensuring customer loyalty were developed. These dimensions are similar to previous research investigating the dimensions that comprise customer service (Frei & McDaniel, 1998; Hogan, Hogan, & Busch, 1984).

The I/O psychologists developed sixteen broad work scenarios were developed to measure the customer service dimensions. One work scenario dealt with a customer whose dry cleaning was lost. In another work scenario a customer comes into an automotive shop and needs his car repaired within a short period of time. The scenario scripts included narration, dialogue between the customer and customer service representative, and the items. The I/O psychologists wrote three to five items for each work scenario and four possible responses for each item, yielding 46 items. The I/O psychologists then had eight subject matter experts (SMEs), who were Human Resources managers and customer service managers from the organizations that participated in the job analysis, review the scenarios, narration, dialogue, and items for clarity, vocabulary level, and difficulty. In addition, the SMEs determined the effectiveness of each of the response options. SMEs rated each of the options independently. Any disagreements on the effectiveness of any option were settled by reaching consensus during a meeting with the I/O psychologists.

The original SJT presented rolling video of the customer service situations. To modify the format from video to a paper-pencil format that can be delivered on a computer the dialogue that was heard and the setting and context viewed in the video was provided in written form. At the beginning of each work scenario the context and dialogue for the customer service employee and customer was presented. This set up each work scenario and the items. The items in this SJT were linked sequentially within each work scenario. This was done by starting the work scenario with the written context and dialogue, and at a critical juncture presenting an item. Once a participant made a response the scenario moved forward from that point. This pattern continued until all of the items for each work scenario were presented. In essence, this particular SJT is a

middle-ground between a typical SJT that contains items that are independent of each other and a branching SJT that has items that are dependent upon each other (Olson-Buchanan, Drasgow, Moberg, Mead, Keenan, & Donovan, 1998). This SJT presented the same item to the participants regardless of how they responded in the previous item.

To ensure that the SJT had an opportunity to correlate with cognitive ability, personality, and job experience I modified the original items and response options. Each of the response options differed in terms of effectiveness, therefore, requiring the participants to evaluate each option. Sociable and gregarious response options were written to relate Extraversion. Courteous and helpful response options were written to relate to Agreeableness. Achievement-oriented response options such as tackling a problem and not shirking duties, and thoroughness response options such as dealing with a problem until it is solved were written to relate to Conscientiousness. Relaxed and stable response options were written to relate to Neuroticism. Finally, creative response options were written to relate to Openness to Experience. Finally, each of the response options was a possible way of handling the situation, thus, relating to different job-related experiences.

When no modifications or when small modifications were made to the response options the original effectiveness rating was used where appropriate. When large modifications were made to response options I re-rated the options. Participants received the maximum number of points for choosing the item that was rated as the most effective, fewer points for the next most effective, and so on. An overall score was calculated for each participant because a single factor often best explains the covariation patterns across SJT items (Chan & Schmitt, 2002; Oswald, Friede, Schmitt, Kim, & Ramsay, 2005).

The original SJT contained 16 work scenarios and 46 items, which is atypical length for an SJT, but much longer than typical for SIs. In order to manage the length of time participants would spend completing the SJT and SI only half of the SJT scenarios and items were used. Each of the scenarios and items were reviewed and a set of 6 work scenarios and 23 items were chosen based . The scenarios ranged from a fast food drive-thru to a cleaners, restaurant, and automotive repair shop. Appendix A provides an example SJT item.

Coefficient alpha was surprisingly low at .30. A follow up analysis was conducted by creating a total score for each of the six scenarios and estimating alpha on those scores. This did not significantly alter the reliability estimate, $\alpha = .31$. These estimates are lower than those found in meta-analytic reviews (McDaniel et al., 2001). However, similar low values have been found in other individual studies (Lievens & Sackett, 2006; Weekley & Jones, 1997). Because SJTs are considered multi-dimensional methods for measuring a range of job-related constructs low internal consistency estimates are not unexpected (Schmitt & Chan, 2006).

SI. To my knowledge, the only other study to make an SJT and an SI out of the same job analysis information was Banki and Latham (2010). Banki and Latham developed an SJT from an SI. However, in the current study, an SI was developed from an SJT. To translate the SJT into an SI the script for the SJT, which contains information on the setting, context, the situational dilemma, and the dialogue between the customer service representative and the customer, was narrated by the interviewer. This was followed by the interviewer asking the situational question (e.g., "what would you say to the customer?"). In this way, the type and amount of information given to the participants was the same across the SJT and SI.

The interviews were videotaped and scored after the interviews were completed. In total, there were 13 interviewers and 10 raters for this study. Each of the interviewers and raters were research assistants. Many of the interviewers were also raters. The raters rated the interview responses using the scale described above. Each interview was rated by two raters and the scores from both raters were averaged to derive the final score for each item. An overall score is most appropriate for SIs (Pulakos & Schmitt, 1995; Roth et al., 2005; Sue-Chan & Latham, 2004); therefore, the ratings for each of the items were summed to form an overall score. Interviewers attended a 1-hour training course on the use of the rating scale as well as to adopt a common frame-of-reference (FOR).

The SI items were rated on a 1 to 4 rating scale and all participants received one rating for each of the 23 questions. Recall that the SJT contained four response options for each item and each of the response options differed in terms of their overall effectiveness. These response options served as the anchors on the rating scale. In other words, the participants' responses were scored against a scale wherein the anchors at each interval were the SJT responses. The response options for the SJT served as examples of what the most effective response would be, the least effective response would be, and so forth. This allowed the SJT and SI to contain the same underlying interval scale and both to be scored using the same judgment standards.

Intraclass correlation is typically used to estimate interrater reliability. However, the structure of the measurement design in this study is one that has been termed an ill-structured measurement design because neither ratees nor raters were fully crossed or nested (Putka, Le, McCloy, & Diaz, 2008). Each SI was rated by two different raters, one in round 1 and one in round 2. SIs were randomly distributed to raters in round 1 then redistributed randomly to raters

in round 2. In order to properly account for the measurement design a variance components analysis (VCA) was conducted. The VCA was conducted by considering the ratees, raters, and SI questions as random effects, estimating the full factorial model with main and interaction effects, and using restricted maximum likelihood estimation to estimate the variance components. The ratee main effect was considered true score variance. Rater and SI question main effects were not considered error because they do not affect the rank ordering of ratees (DeShon, 2002). However, the interaction terms do affect the rank ordering of ratees, which in turn affects correlations. Thus, the interactions between ratees and items and ratees and raters were considered error along with random error. The VCA analysis resulted in a reliability estimate of .52 for 23 items and one rater, and an estimate of .62 for 23 items and two raters.

Test anxiety. The test anxiety scale used in this study was developed by McCarthy and Goffin (2004). These authors measured interview anxiety across five areas from communication anxiety to behavioral anxiety. Only two of the test anxiety areas were relevant to both SJTs and SIs (i.e., performance and behavioral anxiety). Performance anxiety deals with nervousness about one's performance as well as thoughts about failing. Behavioral anxiety deals with behavioral expression of anxiety. Each of McCarthy and Goffin's original items for these two dimensions were modified to refer generally to a test rather than a job interview. Seven items were used to measure performance anxiety (e.g., "I was worrying that my test performance will be lower than that of others"). Six items were used to measure behavioral anxiety (e.g., "My heartbeat was faster than usual during the test"). Ratings across the 13 items were summed to create an overall test anxiety score for the SJT and the SI. Coefficient alpha for the entire scale was .92.

Procedural justice. Bauer, Truxillo, Sanchez, Craig, Ferrara, and Campion's (2001) scale was used to measure procedural justice. As noted above, only two of their justice dimensions are relevant to the current study (i.e., opportunity to perform and consistency of scoring). Four items were used to measure opportunity to perform (e.g., "This test allowed me to show what my job skills are"). Ratings across the four items were summed to create an overall opportunity score. Coefficient alpha for this scale was .91. One item was used to measure consistency of scoring. This item was based on one of the items that Bauer et al. used to measure consistency of administration, and was modified to reflect whether participants felt that situational method allowed people to be scored consistently. (e.g., "The test allows people to be scored in a consistent manner").

Job performance. Job performance was measured by using Van Scotter and Motowidlo's (1996) operationalization of overall job performance as a general framework and adapting their scales to fit customer service jobs. Van Scotter and Motowidlo used the task and contextual performance delineation of overall job performance, but argued that contextual performance should be further divided into interpersonal facilitation and job dedication. They defined task performance as performance on job-specific tasks. Interpersonal facilitation was defined as interpersonally-oriented behaviors that contribute to achieving organizational goals such as helping and cooperating. Job dedication was defined as behaviors related to self-discipline, following rules, and taking initiative.

To derive the content of the task performance scale I used the work tasks that were identified from the job analysis. Examples of tasks on the scale include greeting customers, establishing rapport with customers, suggesting solutions to customers that solve their problem,

maintaining awareness of company policies, and identifying customer needs. A total of 13 items were included in this scale and ratings were made on a 5-point interval scale (1 = does not meet expectations, 3 = meets expectations, 5 = above and beyond expectations). An example item is, "Establishes rapport with customers." A participant's overall task performance score was the mean rating across the 13 items. Coefficient alpha for this scale was .95.

Content for the interpersonal facilitation and job dedication scales was also taken, in part, from the task information gathered from the job analysis. Van Scotter and Motowidlo's (1996) items were also used. Many of the items contained in these two scales were related to customer service and only minimally modified. Each of the modifications was done in order to reference customers or customer problems. Five items were included on the interpersonal facilitation scale (e.g., "Support or encourage a coworker with a work-related problem") and eight items on the job dedication scale (e.g., "Take initiative to solve a work problem or help customers"). Supervisors rated their employees on how likely they were to perform each behavior using a 5-point behavioral frequency scale was used (1 = not at all likely, 3 = somewhat likely, and 5 = very likely). Coefficient alpha for the interpersonal facilitation and job dedication scales was.89 and .91, respectively.

Interpersonal facilitation and job dedication ratings were combined into an overall contextual performance score for two reasons. First, these two aspects were argued to compose contextual performance (Van Scotter & Motowidlo, 1996). Second, only a broad-based perspective of contextual performance was of interest in this study. The mean rating across the 13 items for interpersonal facilitation and job dedication represented an employee's contextual performance score. The final item on the performance scale was an overall performance item that

was rated on a 5-point interval scale (1 = does not meet expectations, 3 = meets expectations, 5 = above and beyond expectations).

CHAPTER 3: RESULTS

Results

Basic Descriptives and Intercorrelations

Descriptives and intercorrelations of the situational tests and variables are presented in Table 1. The total sample size for the study was one hundred seventy-four. However, 33 participants did not complete all of the measures, most of whom did not return for session 2 data collection. An analysis of the differences in the measures variables and performance was conducted on those who completed session 1 and session 2 versus those who only completed session 1. There were no mean differences found between the groups except for Extraversion. Those who did not complete session 2 had a significantly higher a mean Extraversion (M = 38.50, SD = 5.26) score than those who completed both session 1 and session 2 (M = 35.72, SD = 6.74), $t(172)$, p = .03.

Supervisor performance ratings were gathered for 117 participants. This resulted in a response rate of 67%. On average, supervisors had observed their subordinate's performance for slightly less than 1.5 years (M = 16.41 months, SD = 16.69). Ninety of these 117 also had scores for the SJT, SI, cognitive ability, personality, and experience, and 85 had scores for the SJT, SI, cognitive ability, personality, experience, and anxiety and procedural justice. An examination of the performance data revealed range restriction. Overall performance had a high mean (4.39) and a small standard deviation (0.67). In addition, the most common overall performance rating was a five. Similarly, most participants' mean task and contextual performance ratings were between four and five. In order to create more range for the overall performance variable, a mean

performance variable was created by calculating a mean across all items on the performance scale.

As presented in Table 1, the SJT and SI correlated significantly with each other, $r(141) = .20$, $p < .05$. The SJT did not significantly correlate with any of the other measured variables. However, the SI correlated significantly with Extraversion [$r(141) = .21$, $p < .05$] and customer service experience [$r(141) = .18$, $p < .05$]. Neither of the situational tests correlated with overall, task, contextual, or mean performance.

Table 1: Descriptives and Intercorrelations

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
1. SJT	141	77.04	3.71	--											
2. SI	141	74.31	6.09	.20*	--										
3. Cognitive ability	141	24.59	5.20	.09	.04	--									
4. Extraversion	141	35.41	7.33	.03	.21*	.05	--								
5. Agreeableness	141	39.57	4.16	.15	.12	.10	.13	--							
6. Conscientiousness	141	37.88	5.01	.09	.09	.04	.18*	.33*	--						
7. Emotional Stability	141	33.38	6.09	.08	.13	.12	.31**	.17*	.15	--					
8. Openness to Experience	141	36.97	4.96	-.02	.10	.20	.37**	.44**	.22**	.46**	--				
9. Tenure in current position	141	19.92	19.31	-.04	-.05	-.15	-.13	.00	.01	-.04	-.12	--			
10. Customer service experience	141	46.49	40.80	-.13	.18*	-.27**	.10	-.06	.15	-.05	-.03	.38**	--		
11. Tenure in one organization	141	31.94	24.10	-.08	.04	-.13	-.05	-.10	-.01	-.07	-.15	.69**	.64**	--	
12. Total work experience	141	54.21	51.06	-.12	.16	-.17*	.06	-.05	.10	-.02	-.05	.38**	.82**	.65**	--

* $p < .05$, two-tailed. ** $p < .01$, two-tailed.

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
13. SJT Anxiety	132	19.08	6.73	-.04	-.02	-.20*	-.09	-.02	-.17*	-.18*	-.28*	.10	.11	.15	.14
14. SI Anxiety	132	25.50	9.44	-.04	-.07	-.15	-.33*	-.08	-.08	-.25**	-.36**	.11	-.10	-.01	-.80
15. SJT Opportunity to Perform	132	11.24	3.57	.07	.10	-.11	-.16	-.50	.07	.02	-.08	.15	.10	.00	.05
16. SI Opportunity to Perform	132	11.45	3.57	.10	.16	-.50	.01	.08	.04	.02	.07	.05	.09	-.02	.06
17. SJT Scoring Consistency	132	3.33	0.95	.05	.16	.04	.06	-.07	-.08	.03	.01	.08	.02	-.05	-.04
18. SI Scoring Consistency	132	3.18	0.92	.02	.19*	.05	-.07	-.04	-.07	-.02	.00	.00	-.08	-.08	-.15
19. Overall Performance	90	4.39	0.67	-.02	-.20	-.04	.01	.04	.14	-.04	-.04	-.02	.00	-.05	.02
20. Task Performance	90	4.27	0.61	.00	-.10	.05	.13	.15	.29**	.05	.03	.03	.03	.00	.12
21. Contextual Performance	90	4.39	0.56	.04	-.17	.05	-.02	.15	.17	.18	.01	-.01	-.03	.01	.02
22. Mean Performance	90	4.33	0.52	.03	-.15	.06	.06	.17	.26*	.12	.03	.01	.00	.00	.08

* $p < .05$, two-tailed. ** $p < .01$, two-tailed.

Variable	*n*	13	14	15	16	17	18	19	20	21	22
13. SJT Anxiety	85	--									
14. SI Anxiety	85	.58**	--								
15. SJT Opportunity to Perform	85	.05	.22*	--							
16. SI Opportunity to Perform	85	.10	-.04	.63**	--						
17. SJT Scoring Consistency	85	-.11	.05	.49**	.26*	--					
18. SI Scoring Consistency	85	-.05	.08	.47**	.41*	.55**	--				
19. Overall Performance	85	.07	.09	-.04	-.09	-.14	.15	--			
20. Task Performance	85	.03	-.02	.03	-.02	-.12	-.14	.70**	--		
21. Contextual Performance	85	-.02	-.05	-.20	-.17	-.14	-.22*	.65**	.60**	--	
22. Mean Performance	85	.01	-.04	-.09	-.10	-.14	-.20	.76**	.90**	.89**	--

* $p < .05$, two-tailed. ** $p < .01$, two-tailed.

Research Questions

Research questions one through five dealt with whether equivalent versions of an SJT and SI correlate differently with cognitive ability, personality, job experience, and performance. Steiger's Z (Meng, Rubin, & Rosenthal, 1982) was used to test for differences in the dependent correlation coefficients. Differences in the correlations for the SJT versus the SI were tested independent of whether individual correlations for either situational test were statistically significant because although an individual correlation may not be significantly different from zero two correlations may be significantly different from each other. Moreover, two correlations may be different from each other because the correlations are the same or a different sign. Thus, a one-tailed and two-tailed tests was used. Steiger's *Z* results are presented in Table 2. Figure 1displays the correlations for the SJT and SI in order to make eye-ball comparisons of the magnitudes. The results show that only the SI's correlations with Extraversion, customer service experience, and total work experience were significantly different from the correlation for the SJT.

Table 2: Within-groups Comparison of Correlations

	n	SJT	SI	*Z*	*p*
Cognitive ability	141	.09	.04	0.47	.319
Extraversion	141	.03	.21‡	-1.69*	.046
Agreeableness	141	.15	.12	0.28	.390
Conscientiousness	141	.09	.09	0.00	.500
Emotional Stability	141	.08	.13	-0.47	.319
Openness to Experience	141	-.02	.10	-1.12	.131
Tenure in current position	141	-.04	-.05	0.09	.464
Customer service experience	141	-.13	.18‡	-2.88††	.001
Tenure in one organization	141	-.08	.04	-1.11	.134
Total work experience	141	-.12	.16	-2.60††	.005
Overall performance	90	-.02	-.20	1.25	.106
Task performance	90	.00	-.10	0.69	.245
Contextual performance	90	.04	-.17	1.45	.073
Mean performance	90	.03	-.15	1.24	.107

‡ Significantly different from zero.
* $p < .05$, one-tailed. ** $p < .01$, one-tailed.
† $p < .05$, two-tailed. †† $p < .01$, two-tailed.

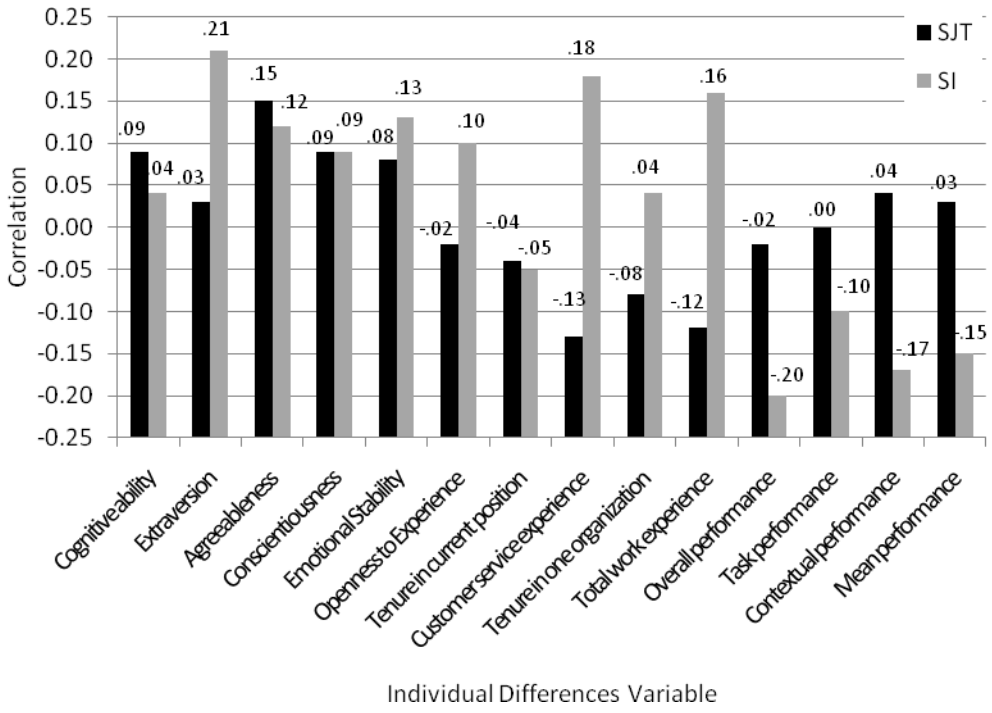

Figure 1: SJT and SI Correlations Across Study Variables

Research questions six through eight dealt with whether participants who complete equivalent versions of an SJT and SI perceive that both methods provide the same opportunities to perform one's skills, perceive that both methods provide the same level of scoring consistency, and whether they experience the same level of test anxiety. A repeated-measures *t*-test was used to test for differences in these means. Table 3 presents the descriptives and repeated-measures *t*-test results for the procedural justice and test anxiety ratings. The results show only the mean test anxiety ratings for the SJT and SI were significantly different from each other. Participants experienced significantly higher levels of test anxiety in the SI than the SJT.

Table 3: Mean Differences in Test Anxiety and Procedural Justice Ratings

	n	*M*	*SD*	*t*
Opportunity to perform (SJT)	133	11.23	3.56	-0.74
Opportunity to perform (SI)	133	11.43	3.57	
Consistency of scoring (SJT)	133	3.34	0.95	1.91
Consistency of scoring (SI)	133	3.19	0.92	
Test anxiety (SJT)	133	19.20	6.83	-8.99**
Test anxiety (SI)	133	25.65	9.57	

* $p < .05$, two-tailed. ** $p < .01$, two-tailed.

Exploratory Analyses

Suppose a selection system is designed with a cognitive ability test, personality test, measurement of job experience, SJT, and SI. Further suppose that the simple question, "Does the order in which the SJT and SI is administered affect the results of the selection decision?" is asked. Here, the issue is whether SJT and SI scores are affected by the order of administration. The experience of taking one situational test may affect how someone performs on the second. Participants may score higher on the second situational test because of the practice they receive from taking the first. Participants may also remember how they responded in the first situational test and repeat or provide very similar responses in the second situational test. This is likely to be most pronounced when participants are provided the SJT first, then the SI. In this sequence, participants are given a set of possible responses in the SJT that they could repeat in the SI. By contrast, participants who completed the SI first may not have been able to repeat their responses

on the SJT because the SJT responses may not have been similar to the responses they made in the SI.

To test for possible order effects on mean SJT and SI scores a 2 x 2 repeated measures analysis of variance (ANOVA) was conducted wherein the situational test was the within-subjects variable and order was the between-subjects variable. The results showed that the main effects for situational test was significant indicating that, on average, participants scored higher on the SJT than the SI, $F(1,140) = 28.10$, $p = .001$. The main effect for order was also significant indicating that participants, on average, scored higher on both the SJT and SI when the SJT was given first followed by the SI, $F(1,140) = 8.29$, $p = .005$. Finally, the interaction was significant indicating that participants' scores on the SJT and SI depended on the order in which they were administered, $F(1,140) = 6.13$, $p = .014$. Figure 1 displays these results. A follow-up analysis of simple effects was performed to examine the interaction.

Inspection of Figure 1 shows three mean differences that are of interest. First, the mean SJT score is higher than the SI mean score when the SJT is given first and when it is given second. However, the difference is larger when the SJT is given second. Second, the mean SI scores is larger when it is given second than when it is given first. To test for differences in these means a post-hoc Tukey HSD was conducted. The analysis revealed that when the SJT was given first the mean SJT score ($M = 77.16$, $SD = 3.63$) was not significantly higher [$q = 3.14$, $p > .05$] than the mean SI score ($M = 75.56$, $SD = 5.82$). However, when the SJT was given second the mean SJT score ($M = 76.77$, $SD = 3.87$) was significantly higher [$q = 6.95$, $p < .01$] than the mean SI score ($M = 72.35$, $SD = 6.02$). Lastly, the mean SI score when it was given second ($M = 75.56$, $SD = 5.82$) was significantly higher [$q = 5.64$, $p < .01$] than the mean SI score when it was

given first (*M* = 72.35, *SD* = 6.02). Taken together, these results show that the order in which the SI was administered significantly affected participants' performance. Specifically, participants performed significantly lower on the SI when it was given first and unaffected by previous exposure to the SJT.

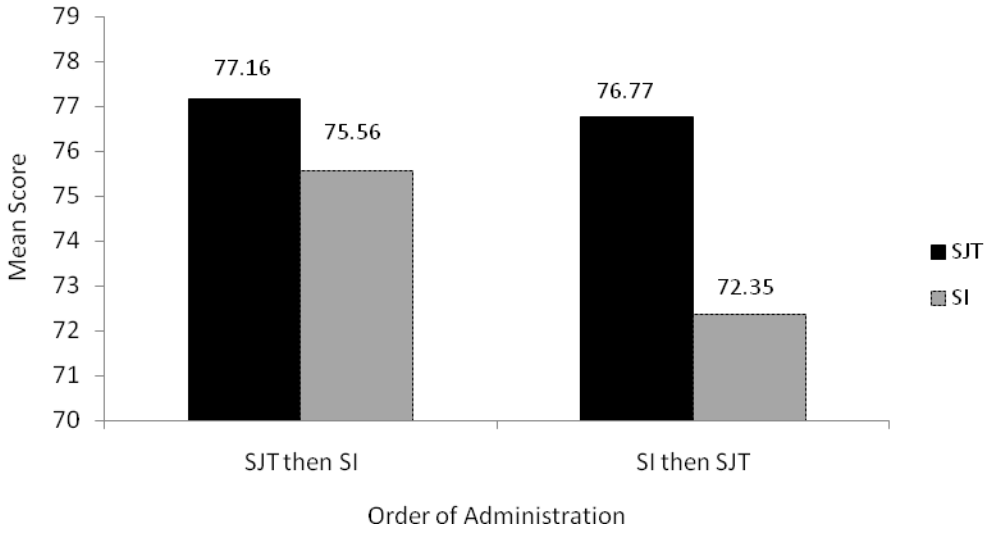

Figure 2: Mean Scores by Situational Method and Condition

CHAPTER 4: DISCUSSION

Discussion

The purpose of this study was to compare and contrast equivalent forms of an SJT and SI on construct and criterion validity, procedural justice reactions, and text anxiety perceptions. The results of this study showed that a equivalent SJT and SI correlated significantly with each other, but the practical magnitude of the correlation was small. This finding is similar to the finding from Banki and Latham (2010) who found a .31 correlation between the two. The relatively small correlation between the SJT and SI suggests that they are not redundant with each other because they do not share much common variance.

The correlations with the individual constructs support this finding. The SI correlated significantly higher with Extraversion, customer service experience, and overall work experience than the SJT. This suggests that there is something unique about the SI that allows it to correlate more highly with these constructs. Despite the similarities between the two methods, these findings suggest that the differences between the methods are likely the main reasons why the construct validity is different.

I speculate that a possible reason for the positive correlation for the SI is because the SI allowed participants to use their experiences and acquired knowledge to influence and construct their responses to the questions. In other words, the positive correlation suggests that participants were able to rely on their experiences in order to construct better responses. However, in the SJT, the response options were limited and the same for everyone. Participants were not able to modify the response options to match their experiences and knowledge. Even though the situational dilemmas may have been similar to their experiences the responses options may not.

As a result, participants may have been less likely to consistently choose a response that was similar to their experiences and knowledge, and less likely to consistently choose a better response.

Overall, there wasn't a significant difference between the SJT and SI with cognitive ability and the personality traits except for Extraversion. Thus, the SJT and SI may not be appreciably different in terms of the correlation with these traits. However, with regard to Extraversion, the results from this study (i.e., .22) are similar to the finding from DeGroot and Kluemper (2007) who found a .28 correlation between an SI and Extraversion. McDaniel et al.'s (2007) meta-analysis found that the mean observed correlation between SJTs with a "what would you do?" response instruction and Extraversion was .07, which is only slightly higher in magnitude than the correlation found in this study (i.e., .03). Thus, SIs may correlate more highly with Extraversion than the other personality traits and more highly than the SJT.

The significant difference that was found for Extraversion may have occurred from a similar process as described above. Participants responding to the SI questions may have been better able to construct their responses to include the facets of Extraversion than when choosing responses in the SJT. This may be the case because the correlation between the SJT and Extraversion is driven by the extent to which the response options include elements of Extraversion. The SJT in this study may not have had a particularly strong loading of Extraversion. Moreover, it may be difficult for an SJT to correlate with Extraversion because Extraversion is an outward expression of personality in social situations and this may be difficult to include in SJT response options.

There is another possible explanation for these findings. Extraversion is a personality construct that largely deals with social interaction (i.e., sociability, talkative, and assertiveness), and interviews are social interactions. Thus, more extraverted individuals may feel more comfortable in social situations like interviews. Findings from this study support this supposition. A significant negative correlation (-.33) was found for SI anxiety ratings and Extraversion. It seems possible that the social situation of the SI allowed more extraverted individuals to express themselves through their responses. Thus, it could be not only the ability to construct one's responses in the SI, but also the social context of the SI that promotes a higher correlation with Extraversion.

The criterion-related validity of the SJT and SI could not be evaluated in this study because neither correlated with any of the performance variables. The performance data were restricted in range wherein a significant majority of the participants received either a rating of four or five, thereby restricting the ability of the SJT and SI to correlate with these data. However, the pattern of correlations for the SJT and SI suggests that the SI may correlate more highly with performance than the SJT. In addition, the correlations patterns also suggest that the SI may correlate more highly with task and contextual performance than the SJT because the SI's correlations with Extraversion, customer service experience, and overall work experience were significantly different (i.e., more positive). Until better performance data are gathered, this leaves open the questions regarding whether a equivalent SJT and SI correlate similarly with overall, task, and contextual performance. However, Banki and Latham (2010) showed that an SI accounted for incremental variance in overall performance above and beyond the SJT, but that the reverse was not true.

The remaining research questions for test anxiety and procedural justice were answered. First, the results showed that participants felt they had the same opportunity in the SJT and SI to show their skills, and they felt that the SJT and SI had similar levels of scoring consistency. By contrast, participants felt significantly more anxiety in the SI than in SJT. A note must be mentioned about these findings. While participants did not provide significantly different ratings for opportunity and scoring consistency they also didn't provide very high ratings on these variables overall. On average, participants rated these variables about a 3 ("neither agree nor disagree"). Similarly, although participants felt significantly more anxiety in the SI they didn't report a very high level of anxiety overall. On average, participants rated each of the anxiety items as a 2 ("disagree"). In total, participants did not view the SJT and SI with high levels of procedural justice, were not anxious about their performance, and did not manifest behavioral expressions of anxiety even though they were instructed to act like a real applicant and were incentivized with a monetary reward for the highest scorer. Because participants did not feel much performance or behavioral anxiety it indicates that neither the SJT nor SI had particularly high levels of situational strength.

Practical Implications

Human Resources personnel, I/O psychologists, and consultants have to make decisions about which selection methods to include in a selection system based on how well they predict performance and how well they correlate with job-related constructs. These decisions are quite easy when the focus is on more uni-dimensional tests like cognitive ability and personality tests that measure distinct, non-overlapping constructs. However, the decisions are much more difficult for assessments like SJTs and SIs because they are multi-dimensional methods capable

of correlating with a range of job-related constructs (Conway & Peneno, 1999; DeGroot & Kluemper, 2007; Gibb & Taylor, 2003; Huffcutt et al., 1996; McDaniel et al., 2007; Weekley & Ployhart, 2005).

This study has several practical implications in this regard. First, the overall utility of a selection system is predicated upon a few things, one of them being the extent to which each instrument in the system measures relatively separate aspects of the job domain. To the extent that SJTs and SIs do not completely overlap in terms of the constructs they correlate with then Human Resources personnel and I/O psychologists may be able to use SJTs and SIs in the same selection system. This single study does not provide a definitive conclusion about the extent to which these two methods correlate with the same constructs, but the conceptual arguments and empirical evidence suggests that they are not completely redundant with each other.

Second, if SJTs and SIs are used in the same selection system, then this study sheds light on the order in which they should be administered. The exploratory analysis showed that the SI scores were significantly higher when it was administered after the SJT, than when it was administered before the SJT. This suggests that participants may have repeated their answers from the SJT in the SI. Thus, in order to avoid this type of contamination the SI should be given first.

Third, not only could the decision be to use both the SJT and SI, but also to use either the SJT or SI. If Human Resources personnel and I/O psychologists can only choose either the SJT or the SI then the findings from this study suggest that the SI is the best choice. However, this study did not adequately simulate a high-stakes testing environment. Situational strength and test anxiety that occur in high-stakes testing environments are important factors for interviews that

must be considered because they have been shown to lower correlations with personality (Van Iddekinge et al., 2007) and impair performance (Cook, Vance, & Spector, 2000; McCarthy & Goffin, 2004).

Theoretical Implications

Much of the support for the use of SJTs and SIs comes from empirical data showing a correlation between scores and measurements of job-related constructs and job performance. However, much less work has been done to develop theoretical explanations for why SJTs and SIs correlate with job-related constructs and predict job performance. Motowidlo, Hooper, and Jackson (2006) developed a theory to explain the correlation between SJT scores and personality traits. They argued that one's personality plays a role in the evaluation of the effectiveness of the response options. This occurs through an accentuation process whereby, for example, individuals with high levels of Agreeableness tend to view highly agreeable response options as more highly effective than individuals with lower levels of Agreeableness, and individuals with high levels of Agreeableness tend to view highly disagreeable response options as more highly ineffective than individuals with lower levels of Agreeableness. Motowidlo et al's (2006) theory also describes why experience plays a role. They argued that through the expression of one's traits individuals will begin to learn which trait-related behaviors are more effective than others. They also argued that by experiencing a range of situations individuals will learn which solutions are better than others. Thus, they argued that personality has an indirect relationship with the development of procedural knowledge and experience has a direct relationship with the development of procedural knowledge. The Motowidlo et al. (2006) study found mixed support for their theory.

However, the conceptual basis of the theory is promising and argues for more research to be conducted.

Theoretical explanations have also been made for SIs. However, the explanation is a bit broader than Motowidlo et al.'s. Latham and colleagues (Latham et al., 1980; Latham & Sue-Chan, 1999) argued that goal-setting theory (Locke & Latham, 1990) underlies the responses made in an SI. More specifically, they argued that the situational dilemma in an SI allows interviewees to express their behavioral intentions to act in certain ways. For example, when an SI questions asks, "what would you do?" it is argued that an interviewee responds with what they intend to do in order to solve the dilemma, and it is the intention to behave that relates to actual performance on the job (Latham & Saari, 1984). Latham and Sue-Chan (1999) stated that by finding a positive and meaningful correlation with job performance that the suppositions are supported.

The importance of these theoretical explanations to the current study is due to the fact that the equivalent SJT and SI shared only a small portion of variance and they did not correlate the same with each of the job-related variables. Perhaps because the SJT presents a set of responses from which an individual makes a choice, and the SI requires that an individual construct their own response means that there are different theoretical explanations for the processes that take place when completing an SJT and SI. On the one hand, the theory for SJTs is built on how individuals evaluate stimuli and on the other hand the theory for SIs describes the mechanism for goal-related behavior. Perhaps Motowidlo et al's (2006) theory and Latham et al's (1980) theory are actually compatible and describe the same phenomenon in different ways. In other words, perhaps there is only one theory that underlies both. Thus, one of the

implications of this research is that it brings to the forefront the need to explain why SJTs and SIs may correlate with different job-related constructs and, potentially, predict different aspects of performance.

In addition to the implications for the theoretical explanations there are implications for the response processes that take place when individuals complete an SJT and SI. To my knowledge there is no research that examines the actual responses that individuals engage in during the completion of an SJT and SI. However, there has been work that was done by Ployhart (2006) to describe and apply a response process model to SJTs. His model is called the Predictor Response Process Model (PRPR; "proper" model). It contains four processes or stages, namely, comprehension, retrieval, judgment, and response. This model may prove to be useful for examining the underlying processes of SJTs and SIs. The characteristics of SJTs and SIs affect one or more of these stages. For example, social desirability may affect the judgment stage SJTs whereas the strength of the situation may affect the response stage for SIs.

Related to understanding the response processes are the characteristics of the situational dilemmas and response options that activate certain construct-related response processes (Messick, 1995). There may be characteristics of situational dilemmas and response options that promote the activation of certain constructs versus others. Meyer (2010) has developed a taxonomy of situations containing two orthogonal dimensions (maintenance-development and formality-informality) that can be used to categorize situations into one of four types (i.e., bureaucratic, strategic, prosaic, and incubative). Motowidlo et al. (2006) wrote each of the response options in their SJT to reflect different levels of a single personality trait. The one thing the SJT and SI share is the situational dilemma. Thus, the implications of the findings from this

study are that it suggests that the response options in the SJT activate different constructs than the situational dilemmas alone in the SI.

Study Limitations and Areas for Future Research

The first limitation of this study was that it did not adequately simulate a high-stakes testing environment. Participants were instructed to act like a real applicant and they were incentivized with monetary rewards. In a real employment context, many applicants are likely to feel some amount of performance anxiety because many want to perform well so that they will get the job. However, participants in this study reported that they felt almost no performance anxiety. If participants did not feel much performance anxiety, then it's also likely they did not alter their behavior or responses in order to perform well and "get the job." Also, because this was a laboratory setting there may have been different situational cues in the environment than would occur in a real employment setting. Future research should replicate this study in a real employment context where true levels of performance anxiety may manifest and the testing environment is likely to be much stronger.

The second limitation is the reliability of the SJT. Coefficient alpha for the SJT in this study was .30. Although previous research (Lievens & Sackett, 2006) has found a similarly low alpha estimate, this is still is quite low, even for SJTs. The reason this is a limitation is because such low internal consistency prevents the ability of the SJT to correlate with other individual differences variables and performance. As dismal as this seems, it may point out another difference between SJTs. The low alpha estimate suggests that 1) participants were more inconsistent in their responses than were the raters in scoring their responses from the SI and 2)

the response options correlated less with each other than did the raters. Thus, this may point out that with well-trained raters who use a standard scoring guide SIs can be more reliable than SJTs.

A third limitation of this study is the poor criterion data that were collected. Neither the SJT nor the SI correlated with the performance criterion. What's more, cognitive ability also didn't correlate with performance, and previous research has consistently shown a robust correlation between the two (Schmidt &Hunter, 1998). There are two possible reasons for this. First, there may have been a selection threat to validity wherein supervisors only provided ratings for certain employees. An analysis of the variable means for those individuals who had performance ratings versus those who did not was performed to see if this was the case. The results showed that those individuals who had performance data had statistically significantly higher tenure in their current position, tenure in one organization, and overall work experience. Thus, there was a selection threat in the data and supervisors tended to rate their more experienced employees. In addition to this, a second possible reason is that supervisors may have also rated their employees leniently. In other words, many of the supervisors may have simply done their employees a favor by completing the performance rating scale, but did not make ratings in an accurate fashion. Both of these issues would serve to restrict the range of the performance data.

Future research should replicate this study in a real employment context where more accurate performance data can be collected. This would allow for research questions involving whether an equivalent SJT and SI predict task and contextual performance differently. If this study is replicated using the same method to collect performance data (i.e., anonymously from supervisors), then a different procedure could be used. Supervisors could be asked to rate their

employee's performance in relation to others instead of providing an absolute rating (Krajewski, Goffin, McCarthy, Rothstein, & Johnston, 2006). This would likely mitigate the possibility that supervisors provide similar and high ratings for their employees.

Finally, there are a couple of additional areas for future research. It was noted above that the findings from this study suggested that different response processes may be activated by SJTs and SIs. Future research should investigate the responses processes for these two methods. Ployhart (2006) and Messick (1995) suggested that a cognitive task analysis could accomplish this. For example, an analysis could be made on how individuals arrive at particular choice or response by asking them to verbally describe their thought processes as they read the situation and make their choice or response.

A second area for future research is on the characteristics of situational dilemmas and response options that activate certain response processes and constructs. Because SJTs and SIs are not redundant with each other there may be various characteristics of the situational dilemmas and response options that tend to activate different constructs. As noted above, Meyer (2010) developed a taxonomy of work-related situations that contained four orthogonal dimensions, and each work situation can be described in one of the four types. Moreover, Motowidlo et al. (2006) developed an SJT with response options that each differed in terms of the level of a particular personality trait that was exhibited. Future research could examine whether different types of situations tend to correlate with certain individual differences variables more than others. For example, a situation where a customer service provider must diffuse an irate customer may be more likely to activate Agreeableness and Emotional Stability than cognitive ability and Conscientiousness. In addition, future research could examine the

characteristics of behaviors exhibited in response options that tend to promote correlations with certain individual differences variables versus others. As an example, if a response option is written to reflect Agreeableness, then what are the best behaviors to include? Also, how agreeable or disagreeable should the response option be? To this point, little research has focused on why SJTs and SIs correlate with specific job-related constructs. Thus, future research could focus more on the underlying reason why one SJT or SI correlates more with, for example, personality than does another.

General Discussion

The importance of comparing predictors of job performance has been highlighted by Hunter and Hunter's (1984) meta-analysis on the criterion validity of various predictors and Schmidt and Hunter's (1998) extended meta-analysis that examined the criterion validity of various combinations of predictors of job performance. More recently, Arthur and Villado (2008) extended this discussion by noting that there is a distinction between the job-related constructs and the methods that are used to measure them. Banki and Latham (2010) further extended the discussion and research by comparing an SJT and SI with the same content in terms of their predictive ability. The current study built upon this foundation of research and discussion and provided evidence that showed that the SJT and SI may not be as similar as their surface similarities suggest. Weekley and Ployhart's (2006) description of SJTs and SIs as "close cousins" rings true here. However, Weekley and Ployhart made this description on the basis of the similarities between the two, not the important differences. Overall, the research and discussion on comparing predictors of job performance is still relatively new and there are still many questions left, especially for multi-dimensional methods like SJTs and SI. More specific

examinations should continue to help us refine our understanding of these commonly-used predictors of job performance.

APPENDIX A: EXAMPLE SITUATIONAL TEST ITEM

Context: *You are working as a drive-thru attendant at a local fast food restaurant. You are taking orders from customers today. You've just finished helping a customer and another customer drives up to make an order.*

You: Thank you, and have a nice day.

Your Mgr.: We just ran out of sausage. I'm going to call another store and send Jack over to get another case. It should take about 20 minutes.

You: OK, let me know when we can start serving sausage again.

Customer drives up to make an order.

You: Welcome to Tuckers. Can I take your order please?

Customer: Yes, I would like two sausage and egg sandwiches, and a small orange juice.

Question 1

If you were the drive-thru attendant, what would you say next?

A. I'm sorry, but we are out of sausage. Would you like something else?
B. Sir, unfortunately we are out of that breakfast item.
C. Ok, but it's going to be a few minutes on the sausage.
D. Thank you sir. Please pull around.

Context: *The customer has now driven up to the window to pick up his order. He's sitting in his car waiting and is somewhat impatient.*

(customer talking to himself)

Customer: Come on, come on. I haven't got all day.

You: Wow, what a morning I have had. Seems like nothing is going right. First, my alarm clock didn't go off, so I was late. I couldn't find my apron to bring, then, finally, I get here and it's been like a mad house around here. Everyone is rushing around . . .

(in an impatient tone)

Customer: I am in a hurry. Can I have my order now?

(taken back by the comment)

You: Sure, that will be $2.93.

Your coworker walks up to you as you're helping the customer.

Coworker: We don't have any clean trays. Can you go in the back and grab us some?

Question 2

If you were the drive-thru attendant, what would you do next?

A. Hand the customer his breakfast, apologizing for the time it took and then help your coworker.

B. Tell your coworker to wait until you have finished helping the customer.

C. Tell your coworker that you are too busy to help him right now.

D. Ask the customer to wait because you need to help your coworker.

Context: *You now have the customer's order and are handing it to him.*

You: Here is your order, sir. Is there anything else I can do for you?

(in a frustrated tone)

Customer: You know I've waited ten minutes just for two egg sandwiches and a small orange juice. I thought you guys were supposed to be fast food.

Question 3

If you were the drive-thru attendant, what would you say next?

A. Sir, we're busy today and going as fast as we can.
B. Sorry about that. We're trying our best.
C. There's no reason to yell sir.
D. You have your food now. You should be happy.

Context: *You have now given the customer their order and they are looking through the bag. The customer is still sitting in his car in front of you at the drive-thru window. You have just taken another customer's order through the intercom.*

(talking to the next customer)

You: That's a dollar eighty nine, please pull up to the window please.

(angry tone)

Customer: I thought I ordered a large orange juice not a small one; can't you guys get it right?

Question 4

If you were the drive-thru attendant, what would you do next?

1	Apologize, correct the order and do not charge Tim the difference in price for the large orange juice.
2	Apologize, correct the order and charge Tim the difference in price for the large orange juice.
4	Apologize, correct the food order, but explain to Tim that you are sure he ordered a small orange juice.
3	Apologize and call your manager in order to resolve Tim's complaint.

APPENDIX B: HISTORY AND DVELOPMENT

History of SJTs

Situational judgment tests (SJT) are measurement methods that present situations, problems, or dilemmas and require an individual to respond with what they should or would do. Therefore, we can trace the history and roots of SJTs back to early attempts by psychologists to measure intelligence. In the early 1900s, Binet and Simon conducted research on intelligence with the goal of identifying children who would not benefit from typical school instruction and created the Binet-Simon scale. E.L. Thorndike, another prominent psychologist in the early 20th century interested in the measurement of intelligence, developed his own intelligence test called *Intelligence Scale CAVD* that measured sentence completion (C), arithmetic (A), vocabulary (V), and the ability to follow directions (D) (Thorne & Henley, 2005). Around the same time, Yerkes, Terman, and Goddard worked to identify which individuals were fit to be trained to fight in World War I, which eventually resulted in the development of two intelligence tests called *Army Alpha and Army Beta*.

With regard to the roots of SJTs, the most influential person was E.L. Thorndike. He suggested that there is not one type of intelligence, but three types, namely abstract, mechanical, and social (Thorndike, 1920). Abstract intelligence was defined as the ability to understand and manage ideas and abstractions. Mechanical intelligence is the ability to understand and manage concrete objects of the physical environment, and social intelligence is the ability to understand and manage people (Thorndike & Stein, 1937). Seemingly, the division of intelligence into three more specifically defined types was based on anecdotal observations or stories of someone who excelled at, for example, activities that were more cognitive in nature, but struggled in activities that were more mechanical or social. R.L. Thorndike (1991) described his father, E. L.

Thorndike, as someone who would be rivaled by few in terms of abstract intelligence, but had almost no ability to fix or build machinery or apparatus. Hunt (1928) described examples of students who are exceptional in school, but failed in a work environment or people who are successful human engineers, but lack mental ability. Because of these examples, E.L. Thorndike and other psychologists realized that in order to explain the successes of individuals in real world endeavors they needed to develop methods that can measure these different aspects of intelligence, and it is in this vein that social intelligence tests and items depicting situations relating to understanding and managing people emerged. These social intelligence tests were the early forms of modern day SJTs.

1870s to 1930s. Prior to Thorndike's (1920) conception of social intelligence, there were a few examples of SJT-like tests. In 1873, for example, an exam used by the U.S. civil service contained items that presented work-related situation and one such item asked, "What action would you take?" (Weekley & Ployhart, 2006). About 30 years later in 1905, Binet's intelligence test contained items presenting real life situations or dilemmas asking, "What should you do?" (Weekley & Ployhart, 2006, p. 5). Because these items required open-ended responses from test takers, they were more like interviews than SJTs, in that respect (Weekley & Ployhart, 2006).

The situational tests that more closely resemble modern day SJTs arrived in the 1920s. The first of these tests was the George Washington University Social Intelligence Test. It claimed to measure six factors related to managing people and one of the scales measured an individual's judgment in social situations. As described by Hunt (1928), the items in the scale presented social relationship and business problems as well as problems relating to judging human nature. For each item, test takers were instructed to choose among four possible solutions.

Between the late 1920s and mid 1930s the George Washington test received a fair amount of attention by researchers (Thorndike & Stein, 1937). Hunt (1928) reported that it was given to high school and college students in addition to employees in positions ranging from administrative to executive, clerical to sales, lower grade industrial workers to engineers, and nurses. The George Washington test was given to, for example, sales employees and scores were correlated with supervisory ratings of their ability to manage people. Also, it has been given to college students and correlated with teacher and fraternity/sorority member ratings as well as the number of extracurricular activities engaged in by the student.

Because Binet's scale was used mainly for children, the George Washington test was really the first to systematically use situation-based items in order to measure a basic human attribute designed to predict adults' school and work success. In the criterion-related validity section of this paper, recent studies will be reviewed that use modern day SJTs in order to predict school and work performance. Current SJTs have much to owe to the George Washington test because of the way in which it was constructed and used.

1940s. The use of tests intending to measure either practical or moral judgment occurred throughout the 1940s. In 1941, Cuber and Pell sought to measure moral judgments by utilizing items that describe controversial moral situations. Some of the situations, cited by Jones (1943), resemble real life situations and one situation is rather long and complex describing a college student's troubles in intimate relationships. These items are different than those used in the George Washington test because of the moral dilemma that is placed within the situations as well as the fact that an individual is asked whether they agree with the behavior depicted in the scenario.

In 1942, Cardall developed and published his practical judgment test. This test was intended to measure judgment in business and social situations by using a set of 48 multiple choice items. The Cardall (1942) practical judgment test is often cited by other SJT researchers, but as of yet no example items have been published. Thus, it is difficult to understand the nature of this test and its attempt to measure judgment. However, what can be noted is that during the 1940s there was an alternative use of situation-based items in order to measure individual differences.

Late 1940s thru 1960s. Recall from the earlier review of situation-based items that they were used in and depicted school environments, business environments, and real life events. However, during the 1950s and 1960s, situation-based items and tests became more and more utilized, almost to the point of exclusivity, in business environments. The most notable of those tests was the *How Supervise?* test by File (1945). *How Supervise?* is a test that measures supervisor quality and is based on a set of assumptions that supervision is general in nature and is related to one's ability to understand human relations, and that to measure one's knowledge of supervision they could evaluate the responses they make to situations they will likely experience. Following these assumptions the test required individuals to respond to problems by indicating what they think should be done or whether they thought it was desirable to do one thing or another. The main applications of the *How Supervise?* test was to select supervisors from groups of employees with experience on the job, but who are not supervisors already and to evaluate supervisory training programs. To this day, *How Supervise?* continues to be sold by a test publishers and used by businesses.

The second major test used for selecting supervisors was called the *Supervisory Judgment Test* by Mandell in 1954. Instead of situation-based items, this test utilized a multiple choice format covering two main areas of supervisory responsibilities; interpersonal relations with superiors, subordinates, and peers and general duties relating to training and evaluating employees. Though the test used a different format from other situation-based tests it is still related to SJTs in concept because of its intention to measure judgment in work-related activities.

A similar type of test with the purpose of measuring "supervisor insight" was created by Mowry in 1957. His *Supervisor's Problems* test appears to be a blend of the two previously mentioned tests because it contains items representing problems faced by supervisors in their everyday experiences involving human relations and the format of the items was multiple choice. Mowry structured his test such that a set of multiple choice items were written for and nested underneath each supervisor problem. Presumably, though the response format is not identified in the article, if the test resembles the *Supervisory Judgment Test* then one may have been asked to respond to the multiple choice items by choosing the best option or the option they think they should do.

1970s to current day. Between the 1970s and mid 1980s research and development of situation or problem-based tests was relatively sparse. In reviewing the literature, it is difficult to find research that examined either ways of refining the existing tests, investigated alternative development methods, or collected additional data regarding correlations with performance for different types of jobs. In addition, it is difficult to find any new tests that were developed either by academicians or test publishing companies. It wasn't until the mid-to-late 1980s that the use of situation-based items reappeared in the literature. Wagner and Sternberg (1985) and Wagner

(1987) reintroduced the use of this type of item in their research on measuring practical intelligence. However, for industrial/organizational psychologists, Motowidlo, Dunnette, and Carter's paper in 1990 is often regarded as the article that fully reintroduced situation-based items and spurred much of the current interest in SJTs (Weekley & Ployhart, 2006).

Construction steps and methods for establishing the content validity of situation-based tests were difficult to find prior to the studies published by Motowidlo et al. (1990) and Wagner and Sternberg (1985); although, a study by Mowry (1957) provided a small amount of general information. Even in light Mowry's study, it was Motowidlo et al.'s article that provided the most specific information that enabled practitioners, test developers, and researchers to construct their own SJTs. In fact, before their paper, there were no established guidelines for constructing an SJT or what they should look like. Most current SJTs more or less follow Motowidlo et al.'s construction methods and some variant of their scoring procedure. The following is a detailed summary of their study.

In developing their "low fidelity simulation", Motowidlo et al. (1990) used, as a guideline, the steps used by Latham et al. (1980) in the development of their situational interview. The authors began the construction of the simulation with the goal of measuring general management performance within the telecommunications industry and had no explicit intention of actually measuring a set of constructs including social intelligence or supervisory quality. However, they organized the focus of their items around two aspects of management performance, namely, problem solving and interpersonal relations.

The first step utilized to derive the stem or situations for the simulation was to collect a set of critical incidents from a set of subject matter experts (SMEs; usually supervisors or

incumbents) that depicted effective and ineffective performance. Within these incidents is usually information regarding the incident itself, the behavior of the person, and the outcome of those behaviors (Flanagan, 1954). From the roughly 1,200 incidents collected, items were written that reflected the nature of the situation and presented a problem that must be dealt with or solved.

The second step included gathering another set of SMEs to generate possible responses to each of the situations written from the critical incidents. Often the responses that are generated and chosen aren't the same in terms of their effectiveness, and they shouldn't be. Thus, the perceived effectiveness of the responses needed to be determined. To do this, the authors had a group of SMEs rate each response on how effective they thought they were. Also, the SMEs were instructed to choose the best and worst options from the list of possible responses for each item. The items with a sufficient level of agreement amongst the raters regarding the effectiveness of the responses as well as which ones were the best and worst responses were included in the final version of the test. Through this process, Motowidlo et al. (1990) was able to establish the content validity and simulate in a paper-pencil test work-related situations and responses. Below is an example of one of their items:

> You and someone from another department are responsible for coordinating a project involving both departments. This other person is not carrying out his share of the responsibilities. You would…
>
> _____Most likely _____Least likely

1. Discuss the situation with your manager and ask him to take it up with the other person's manager.

2. Remind him that you need his help and that the project won't be completed effectively without a full team effort from both of you.
3. Tell him that he is not doing his share of the work, that you will not do it all yourself, and that if he does not start doing more you'll be forced to take the matter to his manager.
4. Try to find out why he is not doing his share and explain to him that this creates more work for you and makes it harder to finish the project.
5. Get someone else from his department to finish the project.

Many other response instructions are currently used besides the most/least likely format used by Motowidlo et al. (1990). Other response instructions include picking the one option that you would do or should do. Also, an individual could be asked to rate the effectiveness of each response. Lastly, some have asked individuals to pick the best and the worst response. The choice to use one response instruction versus another may not entirely be related to the nature of the instructions themselves, although some research shows that you can measure different things with different instructions (see the Construct Validity section for those studies). Some may use a particular response instruction because of a previous history of using one versus another, a perceived superiority of one versus another, or, in the case of the rate the effectiveness instruction, a desire to collect more data. Generally speaking, each response instruction works well within an SJT.

Most recently, a few researchers have begun to change the nature of the response options themselves. Instead of simply writing possible responses to a situation Motowidlo et al., (2006) wrote responses that not only were possible, but also differed in the level of the construct they depicted. For example, they set out to measure a single personality construct per situation and to

do this they wrote some responses that depicted a high level of the construct and some that depicted a low level of the construct. Though this response option format has not caught on in the literature or industry, it may prove to be a useful alternative. An example item measuring Agreeableness follows. The level of Agreeableness depicted in the response option is noted in parentheses.

> You are in charge of a meeting with six people from other departments. One of them has a very blunt way of announcing that something that was just said is stupid or that somebody's idea just won't work. By the time the meeting is half over, he has done this twice in connection with remarks made by two different participants. You should…
>
> a) During a break or after the meeting, explain to him that you appreciate his point of view, but that his comments are hurting the other coworkers (high).
> b) During the meeting, tell him to keep his rude comments to himself or he won't have a job anymore (low).
> c) During a break or after the meeting, tell him that his comments were hurting group participation, and ask him to phrase his criticisms differently (high).
> d) During the meeting, ask him to leave the meeting (low).
> e) During a break or after the meeting, tell him that you don't want to hear anymore comments from him unless they are positive (low).
> f) Address the group as a whole and state that it is important to keep comments constructive (high).

Other recent changes in the nature of SJTs have been made and include using a multi-media format instead of a paper-pencil format and administering an SJT over the Internet. In the

multimedia format the situation is depicted by actors in video and audio although the same written-type responses are used. The multimedia format adds an element of richness to the situations that cannot be depicted in a written form, which increases the perceptions of face validity (Chan & Schmitt, 1997; Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000). For SJTs administered over the Internet HR personnel see many perceived benefits such as the ability to administer a test to anyone anywhere and the reduced time-to-hire cycle for applicants, but there are also some important drawbacks such as test security and cheating (Tippins, Beaty, Drasgow, Gibson, Pearlman, Segall, et al., 2006). Research addressing these issues has already begun.

Since Motowidlo et al.'s (1990) study, the popularity of SJTs has grown, but it is not necessarily due to the multimedia technology or Internet capabilities. The main reason for the spike in popularity of SJTs is due to the research showing a high correlation with job performance. Research by Motowidlo et al. (1990), Phillips (1992; 1993), Motowidlo and Tippins (1993), Dalessio (1994), Weekley and Jones (1997; 1999), and Olson-Buchanan, Drasgow, Moberg, Mead, Keenan, and Donovan (1998) shows that an SJT correlates with performance across a wide range of jobs. Other reasons include face validity and smaller mean score differences between demographic subgroups (Hough, Oswald, & Ployhart, 2001). In the following sections, I will review the criterion-related and construct validity of SJTs.

History of Interviews

Few organizations or HR personnel would feel comfortable hiring applicants without ever having interacted with them, even if for only a short period of time. Not only does it make them feel more comfortable about the hire, but also people tend to be very confident in the correctness of their judgments (Fischhoff, Slovic, & Lichtenstein, 1977; Kahneman & Tversky, 1973). As a result, many organizations use some type of interview as part of the selection process (Ryan & Sackett, 1987; Spriegel & James, 1958). Various estimates have been made across the years regarding the prevalence of the interview. For example, Spriegel and James (1958) reported that over 93% of organizations surveyed utilized an interview in the selection process. Similarly, Ryan and Sackett (1987) found that about 94% of the respondents to their survey said they used an interview. Thus, the interview, second only to applications and resumes, has a very solid place in the beginning stages of organizations' employment systems.

Research has verified that people make reasonable judgments of character in brief periods of observation. Although Funder and Colvin (1988) found that the agreement between self perceptions and an acquaintance's perceptions of personality was larger than between the self and a stranger, Blackman and Funder (1998) showed that as the amount of time that a stranger spends with a person their perceptions of the person's personality becomes more and more accurate. Specifically, they show that the agreement between the self and the stranger rises as the time spent with the person increases from 5 minutes to 30 minutes. Interestingly, Colvin and Funder (1991) found that the ability of a stranger and an acquaintance to predict what a person will do in a situation that is similar to the one they viewed them is about the same. Considering that the typical interview is at least 30 minutes (Ryan & Sackett, 1987), the pervasive use of the

interview as a way to get an indication and predict how an applicant will perform on the job has merit. Several meta-analyses show that interviews can predict performance on the job (Huffcutt & Arthur, 1994; McDaniel et al., 1994; Wiesner & Cronshaw, 1988).

Certainly, throughout time an interaction between people for the sake of gathering information about the other is not new. However, within the fields of psychology or business the interview as used for therapy or to measure job-related attributes has only been around since the early 1900s. For Harry Stack Sullivan, the interview was a major component of his work in the 1920s and 1930s; Thorne and Henley (2005) state that his book, *The Psychiatric Interview*, is a standard text on the client-therapist relationship. Similarly, Bellows and Estep (1954), Spriegel and James (1958), Ulrich and Trumbo (1965), and Wagner (1949) reviewed the history and prevalence of the use of the interview in business and research and showed that, at least, back to the 1920s and 1930s the interview had become a major component of selection systems.

In early research many authors simply referred to the interview in a bland way without much detail given regarding its nature. However, more recently, the discussion of the interview has grown and become more specific in terms of what types of interviews are being used. Moreover, several researchers in the 1980s created two different types of interviews (Janz, 1982; Latham et al., 1980) As a result, current interviews take on several forms.

Types of Interviews. The first basic distinction in form is the level of structure. Huffcutt (1993) defined structure as the amount of procedural variation in the conduct of the interview across applicants. Thus, structure exists on a continuum ranging from completely unstructured where the interviewer is free to ask any types of questions on any topic and most likely does not have any type of formal procedure for scoring each of the answers, to moderate structure where

the interviewer must ask questions from pre-specified topic areas but not necessarily the same questions of all interviewees, is allowed to ask follow-up questions, and typically has a scoring procedure where they rate the interviewee on various job-related dimensions, to completely structured where the interviewer must ask the same questions of everyone in the same order with no follow up questions and rates each question.

Based on research by Wiesner and Cronshaw (1988) and McDaniel et al. (1994) the level of structure included in the interview affects its predictive ability. Both articles found that structured interviews to be both more reliable and predictive of job performance than unstructured. Using a more specific classification of structure, Huffcutt and Arthur (1994) found that as the level of structure increased the correlation with performance also increased, but to a point. These authors showed that interviews that had a specified set of questions to be asked in a certain order with no follow ups and a formal scoring procedure were not significantly more predictive of job performance than interviews that contained questions revolving around a set of pre-specified topic areas and scores for job dimensions rather than each question. Thus, they state that beyond a level of moderate structure there are diminishing returns. Yet, predictive validity is not the only benefit of structure. Williamson, Campion, Malos, Roehling, and Campion (1997) found that favorable legal outcomes were associated with the job-relatedness and structure of an interview. Thus, most I/O psychologists advocate the use of structured interviews for selection purposes.

The second basic form of the interview is whether it is completed by a single individual or by a panel of two or more. The reasons to conduct an interview with more than one person are many. As noted above, few organizations would hire an applicant without ever having met them.

Similarly, many managers may not want to have a new employee placed in their department or on their team without meeting them first. In addition, for many organizations HR personnel conduct the selection process. Also, not only do managers and HR personnel want to meet the applicant first, but also they would like to have input into the selection decision, especially if they will end up on their team or in their department. Consequently, at least two organizational members may typically be part of the interview.

Another reason for including several interviewers may be that some believe that it allows for the opportunity to evaluate everything said by the interviewee. In other words, if you have only one interviewer they may miss some information. Also, some interviewers may attend to various types of information that other interviewers do not attend to. However, with proper training these issues can be mitigated (Gatewood & Feild, 2001).

Thus, there may be some justification for using more than one person in an interview. However, these justifications must be weighed against the cost associated with utilizing several people. In their survey of I/O psychologists, consultants, and practitioners, Ryan and Sackett (1987) found that the average interview conducted by these groups was about one and one half hours. Considering the salaries of managers and HR personnel, the time spent in the interview can be quite costly. From an empirical standpoint, McDaniel et al. (1994) and Wiesner and Cronshaw's (1988) research shows that panel interviews are not more predictive of job performance than individual interviews. In fact, individual interviews may be slightly more predictive (McDaniel et al., 1994). In other words, in terms of predictive ability, the additional members don't add anything to the prediction over the first person.

The final major distinction in the form of the interview is based on content. Two major types of content are typically used in employment interviews; one form is behavioral (Janz, 1982) and the other is situational (Latham et al., 1980). There is a third form that is described by McDaniel et al. (1994) that is called a psychological interview. This type of interview will not be discussed in this section because McDaniel et al. (1994) categorize this type of interview as unstructured and they are typically conducted by psychologists not HR personnel or managers like the other two types.

The behavioral description interview (BDI) consists of questions based on job-related information and asks interviewees what they have done in various situations. In other words, BDI questions utilize the assumption that past behavior is the best predictor of future behavior and, therefore, the focus is on one's previous experiences and their choices of behavior in work-related situations (Janz, 1989). With this type of interview, probing or follow-up questions are permitted in order to gather more information about the person's behavior.

Both the BDI and the SI are very similar in nature. Both are based on job analysis information and both are designed to measure job-related knowledge, skills, and abilities. Research even shows that they have similar reliability estimates and correlations with performance (Taylor & Small, 2002). However, it is the underlying theory, the way in which the questions are worded, the way in which the interview is conducted, and the level of structure that creates the differences between the two. Some of these differences are not huge, but they do make for a different type of interview.

In contrast to BDIs, the SI, which is based on in goal-setting theory (Locke & Latham, 1990), proposes that one's responses to interview questions are actually intentions to behave in

certain ways. In other words, the SI presents a dilemma to interviewees and they are to respond with what they would do in that situation. Thus, they are very similar in form to SJTs. By asking what would you do, an interviewee must state what they intend to do, and because of this there is thought to be a link to what one would actually do on the job because intentions drive behavior (Fishbein & Ajzen, 1975; Klehe & Latham, 2006). The operational differences for the SI are that probing is typically not done and with BDIs if someone has never had a certain experience they won't be able to successfully answer the question, but this is not the case for an SI because all questions are hypothetical dilemmas.

Latham et al. (1980) described the construction of SIs and it starts with collecting job analysis information. Critical incidents that describe effective and ineffective behavior are sorted according to perceived similarity and then turned into dilemmas. In addition, the behaviors depicted in the incidents serve as behavioral anchors on rating scales used for scoring the interviewee's responses. Typically, although it is not always the case (Sue-Chan & Latham, 2004), ratings are made after each question. In comparison, for BDIs, ratings are made after all questions have been asked (Campion, Campion, & Hudson, 1994). An example item and rating scale from Sue-Chan and Latham (2004) who developed an SI for MBA students is presented below.

> You have been assigned to a group to complete an assignment. You feel that one of you group members is not doing any work at all, while others spend too much time gossiping. Overall, you feel that you are carrying all the weight for the group, and that no one else in the group cares very much about the project. Your professor has emphasized to you that the group must solve its own problems. What would you do?

(5) Discuss my concerns with the group. Work with the group to acknowledge and identify the problem. Devise a solution to the problem agreeable to the whole group.

(3) Get the group to acknowledge there is a problem and vote on the solution; OR, I would confront each member individually.

(1) Do nothing, or, go to the professor with my complaints.

History of SIs

In contrast to the history of SJTs which started in the early 1900s, the history of SIs is relatively short. Certainly, some kind of interview process has existed for many, many years, but an interview with the specific characteristics of the SI is recognized as starting in 1980 with the publication of the article by Latham and colleagues. The development of the SI was a response to the prevailing thoughts about the reliability and validity of the interview for decision making purposes. Specifically, drawing on the literature that showed that interviews were commonly conducted in an unstructured manner, the determination of which dimensions should be measured and the questions to measure them was relatively unstructured, and that the validity literature was mixed (Ulrich & Trumbo, 1965; Wagner, 1949), Latham et al. (1980) sought to remedy these problems by utilizing a structured, systematic job analysis to develop the interview questions and used a structured process for conducting and scoring the interview.

Several early reviews and commentary on this matter suggested that increasing the structure of the interview increasess predictive ability (Ulrich & Trumbo, 1965; Wagner, 1949). More recently, studies have confirmed the benefits of this suggestion (Huffcutt & Arthur, 1994;

McDaniel et al., 1994; Wiesner & Cronshaw, 1988). Thus, the SI has a gained a firm place as one of the major methods used in selection systems. The SI has been used for predicting customer service performance in retail positions (DeGroot & Kluemper, 2007), sales productivity (Weekley & Gier, 1987), performance for social workers and pulp mill workers (Campion, Campion, & Hudson, 1994) as well as to select individuals for teams in a manufacturing organization (Morgeson, Reider, & Campion, 2005).

APPENDIX C: CONSTRUCT VALIDITY

Construct Validity of SJTs

Construct validity, in part, can be evidenced by a set of empirical relationships that are specified within a nomological network (Borsboom, Mellenbergh, & Heerden, 2004; Cronbach & Meehl, 1955; Nunnally & Bernstein, 1994). A nomological network is a theory that describes a set of relationships that are expected between theoretical constructs, the theoretical constructs and the methods used to measure them, and the measurement methods. If, for example, the measurement method is measuring a set of job-related constructs then a set of correlations should be obtained between the measurement method and job performance. In other words, because cognitive ability (Schmidt & Hunter, 1998), personality (Barrick &Mount, 1991; Tett, Jackson, & Rothstein, 1991), and experience (McDaniel, Schmidt, & Hunter, 1998; Quinones, Ford, & Teachout, 1995) have been shown to correlate with job performance one could argue that SJTs measure these constructs; these are the constructs driving the correlation between SJTs and performance.

A nomological network for SJTs could include expected relationships between the SJT, cognitive ability, personality, experience, and job performance. Thus, the correlation between SJTs and job performance is only one of relationships could exist within the network. As a result, this is only a necessary, but insufficient condition for establishing the construct validity of SJTs

More, specific information in the form of correlations with the external constructs cognitive ability, personality, and experience as well as information about the psychometric characteristics of SJTs is required to make solid conclusions about construct validity. Cronbach and Meehl (1955) gave a number of suggestions regarding the types of psychometric information

that work towards establishing the construct validity of a measurement method. These analyses include inter-item correlations, internal consistency estimates, and factor analysis.

Cognitive ability. Of all the major individual difference variables relevant to I/O psychology it is most logical to turn to cognitive ability first. This trait has a rich history of investigation and has shown to be highly related to job performance (Schmidt & Hunter, 1998). Recall that when completing an SJT individuals must either indicate what they should do or would do, rate the effectiveness of each possible response, or pick the best and worst response. Thus, each response option must be examined in order to respond, and one's cognitive ability or reasoning could be used during this process.

In addition, SJT items are predominantly constructed in a written format and in order to effectively respond one must be able to comprehend the text presented. As a result, there is some amount of reading comprehension or verbal fluency that must be relied upon. McDaniel and colleagues (McDaniel et al., 2006; McDaniel et al., 2007) have suggested other factors that could relate to the cognitive loading of SJT items. First, they suggest that as the length of the text describing the situation increases the cognitive requirements may as well. Primarily, this would be because one would have to rely more heavily on their reading comprehension. Second, they note that some items may present more complex scenarios than others, and those items that are more complex may also require a larger amount of one's cognitive abilities. Lastly, they note that certain response instructions could be more related to cognitive ability than others (i.e., "should do" vs. "would do"). Thus, there are many reasons why SJTs should relate to measurements of cognitive ability.

In a meta-analysis by McDaniel et al. (2001) evidence regarding the correlation between SJTs and cognitive ability was gathered. They hypothesized that part of the predictor-criterion correlation was due, in part, to the cognitive loading of SJTs. In all, the authors collected 80 correlation coefficients and calculated the mean corrected population correlation between SJT scores and cognitive ability to be .39 with a 90 percent credibility interval ranging from .09 to .69. Thus, the correlation with cognitive ability is presumably due to the written nature of the items and because of the reliance upon one's reasoning ability when responding.

Subsequent studies echo these findings on the correlation with cognitive ability. For example, Clevenger et al. (2001) found a correlation between an SJT administered to entry-level employees in a government agency and cognitive ability of .11. In this study another sample of data was gathered from customer service personnel. For this sample the correlation between an SJT and quantitative reasoning (a facet of cognitive ability) was .17. Finally, the authors found that an SJT correlated .57 with cognitive ability for a sample of engineers. Although the findings vary from sample to sample each correlation is positive.

In another study, Weekley and Ployhart (2005) constructed an SJT and utilized incumbents from the retail industry as participants. The SJT items were constructed around two broad areas, general managerial situations, and loss prevention situations such theft or accidents. They found that their SJT correlation .36 with SJT scores. In addition to this correlation, they also provide evidence on the partial correlation between SJT scores and cognitive ability. The authors measured a set of individual difference variables in addition to cognitive ability including personality and job experience. Within their hypothesized model they showed that the partial regression coefficient (path coefficient) for cognitive ability was .37.

Although many studies will show a correlation between SJT scores and cognitive ability not all studies do. For example, Chan and Schmitt (2002) administered an SJT designed to measure one's overall adaptability to work-related situations to administrative employees in the Singapore civil service. They report a -.02 correlation between their SJT and measurements of cognitive ability. We see from the results of the meta-analysis by McDaniel et al. (2001) and the studies cited above that the cognitive saturation of SJTs differs. Some SJTs correlate moderately to highly with cognitive ability while other SJTs correlate lowly to not at all. Overall, however, the correlation between SJTs and cognitive ability is positive.

Personality. The second major set of correlates of SJTs is personality traits. Jobs that contain or require a large amount of interpersonal interaction are likely to provide opportunities or the context in which an individual may express their behavioral tendencies (i.e., personality), and it is these types of jobs where one's observed behavioral tendencies may affect job performance (Barrick, Stewart, Neubert, & Mount, 1998; Hough, 1992; Hurley, 1998; Mount, Barrick, & Stewart, 1998; Vinchur, Schippmann, Switzer, & Roth, 1998). To recite an assertion made earlier, if an SJT represents the job well in the scenarios depicted and the response options listed then it is also likely that one's behavioral tendencies will play a part in their responses. By extension, measurements of personality traits, as operationalized in terms of the Big Five, should tend to correlate with SJTs because many SJT items depict some sort of situation in which one works with or interacts with another coworker or customer. For example, an SJT constructed by Weekley and Jones (1997) was intended for a sample of employees in a retail organization. The items in the SJT were built around critical incidents involving friendliness, diplomacy, and teamwork. Thus, these types of items are likely to correlate with personality because of the

interpersonal nature of the items. Furthermore, one's personality may affect the way they perceive the situation and the effectiveness of various responses to the situation (Motowidlo et al., 2006).

Considering these arguments we should see a pattern of correlations across the extant research between measurements of personality and SJT scores. McDaniel and Nguyen (2001) conducted a meta-analysis on the correlations between SJTs and personality and found encouraging results. The authors found that, generally, three personality constructs correlated with SJTs. Conscientiousness, Agreeableness, and Neuroticism had the largest mean correlations (.26, .25, and .31 respectively) while Extraversion, and Openness to Experience failed to correlate highly (.06 and .09 respectively). The mean correlations reported in this study are both positive and rather high. However, the range of correlations was quite a wide. For example, the correlations for Conscientiousness ranged from -.1 to .43, but 5 of the 13 studies reported a correlation above .2. Similarly, the range of correlations for Agreeableness extends from -.03 to .49 with 4 of the 12 studies reporting a correlation above .2.

In addition to the range of correlations, a study included in the meta-analysis appears to be driving the large mean correlation with SJTs for Agreeableness and Conscientiousness. This study had a sample size over 4,000 and reported a correlation above .4 with these two personality traits. Thus, part of the high mean correlation is due to the findings of this one study. Consequently, the mean correlation for these studies may be better represented by excluding this study. McDaniel and Nguyen conducted analyses without this study included and found that the mean correlation for Agreeableness and Conscientiousness with SJTs was .13 and .17,

respectively. Although the exclusion of this study from the analyses lowered the mean correlation it is still positive, and many of the studies included have correlations above .15.

Many of the studies included in this meta-analysis are papers presented at conferences making it difficult to ascertain the nature of the SJTs or the samples used in order to understand why one study correlates more positively with personality than another study. Thus, a review of more recent studies will add to the findings reported by McDaniel and Nguyen (2001) and will provide a better gauge how well personality correlates with SJT scores.

Clevenger et al. (2001) conducted a study with three separate samples to investigate the incremental variance accounted for by SJTs above that for a set of other individual difference variables. Among that set of variables was measurements of Conscientiousness. The results of this study showed that for two of the samples Conscientiousness correlated .16 and .21 with SJT scores. In the third sample, Conscientiousness correlated zero.

Two additional studies, one by Chan and Schmitt (2002) and another by Oswald et al. (2004) show positive and consistent correlations between SJT scores and each of the Big Five personality constructs. Chan and Schmitt's study used an SJT designed to measure a test taker's adaptive ability and included content that was related to resolving interpersonal conflict. Oswald et al.'s study used an SJT designed to measure 12 dimensions of college student performance. Several of the dimensions of performance were interpersonal in nature with one dimension dealing interpersonal skills and another social responsibility and citizenship. Both studies show relatively similar correlations across each of the Big Five personality constructs. Conscientiousness correlated about .25, Agreeableness .30, Extraversion .20, Openness to Experience .20, Neuroticism -.20, and Neuroticism .17.

Another recent meta-analysis conducted by McDaniel et al. (2007) provides the most recent set of findings regarding correlations with personality. A large part of the data for this study comes from the McDaniel and Nguyen (2001) study cited above, but additional correlations were added to the original data set. Generally, they show that SJTs correlate positively with personality. Specifically, they found that Agreeableness has a mean corrected correlation of .25, Conscientiousness .27, Neuroticism .22, Extraversion .14, and Openness to Experience .13.

Overall, the two meta-analyses by McDaniel and colleagues (McDaniel & Nguyen, 2001; McDaniel et al., 2007) and the other studies cited above provide strong support for the assertion that one's personality affects their responses to SJT items. Although these correlations are not as large as those for cognitive ability they account for a meaningful amount of variance and they allow SJTs to measure a portion of the job domain that cognitive ability does not (i.e., contextual performance; Borman & Motowidlo, 1997; Motowidlo & Van Scotter, 1994; Van Scotter & Motowidlo, 1996).

Job experience. The next major correlate found in research of SJTs is job experience. Job experience is typically used a proxy measurement of job knowledge because of the ease with which experience can be measured versus job knowledge. In addition, the logic behind using experience as a proxy for job knowledge is as follows. As one gains experience on the job it is likely that one will also obtain some amount of knowledge about how to effectively perform the tasks, duties, and handle important situations.

As noted in the history section, early on it was hypothesized that in order to measure one's knowledge of how to handle common work-related situations items depicting these types

of situations could be administered and depending on one's performance in the items it would indicate how much knowledge one had (Decker, 1956; File, 1945). In essence, SJTs, in their initial form, were thought of as an indirect measurement of job knowledge. However, little research has been conducted to determine the extent to which SJT performance is based on one's knowledge of the job; only one study to my knowledge has specifically measured job knowledge and correlated it with SJT scores (e.g., Clevenger et al, 2001). Most studies instead of measuring knowledge directly have measured, in one form or another, one's job experience and related that to SJT scores.

Clevenger et al. (2001) cites two studies that measured job knowledge and correlated it with SJT scores. In one of those studies entry-level investigative agents in the government were given a job knowledge measurement consisting of 117 items covering knowledge of job procedures, agency structure, investigative techniques, and interagency relations. The second study utilized a 99-item measure of knowledge of business, customer service procedures, technical skills, and safety policies in a sample of incumbents in customer service positions. The correlation for the first study between the two was .13 and the correlation for the second study was .19. These findings lend credence to the assertion that SJTs measure job knowledge. This assertion may gather even more strength if the amount of job knowledge of these two incumbent samples was restricted in range thereby attenuating the correlations.

In reference to the data on the correlation between SJT scores and job experience, McDaniel and Nguyen (2001) collected studies published prior to 2001 that report correlations between the two. They show a rather small correlation between experience and SJT scores with the mean population correlation equaling .05. In addition, the Clevenger et al. (2001) article

shows correlations of .01, .03, and -.13 across the three studies they cite. These relatively unimpressive findings would lead many to conclude that SJTs don't meaningfully correlate with measurements of experience. However, there is considerable variability in the correlations reported by McDaniel and Nguyen (2001); the 95% credibility interval ranges from -.18 to .28. The credibility interval includes zero and thus some studies may have found a correlation between SJTs and experience by chance. Yet, 11 of the 18 samples reported by McDaniel and Nguyen have a correlation of .10 or higher.

In all, it simply may be that experience is not a good surrogate measure of job knowledge and this is why experience measurements don't correlate that well with SJTs. For example, referring back to the Clevenger et al. (2001) study, they cite correlations between job knowledge and job experience for the two studies noted in the paragraph above. In the first study with government investigative agents the correlation was .04 and in the second study with customer service employees the correlation was -.22. There is some evidence that a potential moderator of the relationship is the level of operationalization of job experience. The moderating effect of level of operationalization is beyond the scope of this review. However, future research would benefit if different levels of operationalization were utilized.

Psychometric data. Inter-item correlations and internal consistency estimates (i.e., coefficient alpha) can be used to support the construct validity of a measure. If one set of items intending to measure the same thing correlates more highly with each other than they correlate with another set of items intending to measure something different then it is possible that there is a common construct underlying the first set of items (Cortina, 1993). Similarly, because internal consistency estimates are affected by inter-item correlations, if internal consistency estimates are

reasonably high then it implies that there may be one construct underlying the items. However, it is also possible that more than one construct is being measured (Cortina, 1993).

Inter-item correlations and internal consistency estimates have both a conceptual as well as a mathematical relationship. Conceptually, if items are considered as observations then the more observations you make the more reliable your measurement. Mathematically, to the extent that items correlate more highly with each other then internal consistency tends to increase because it is an index of shared variance to total variance. However, internal consistency estimates also tend to increase with the number of items in the method as illustrated in the Spearman-Brown prophecy formula. Thus, we need examine each type of information (i.e., inter-item correlations, internal consistency estimates, and number of items) in order to get a picture of the underlying structure. These types of analyses have been performed on SJTs although there is not a lot of information. The following studies will review the data on inter-item correlations, internal consistency estimates, and the number of items typically found in SJTs. First, a study conducted by Oswald, Friede, Schmitt, Kim, and Ramsay (2005) will be reviewed which presents inter-item correlations for an SJT used within a college student environment. Disappointingly, this is the only study found that present such data.

Inter-item correlations. Oswald et al. (2005) constructed an SJT containing 57 items which were either written or selected to relate to 12 dimensions of college student performance. Items were rationally categorized into the dimensions and mean inter-item correlations were calculated for the items within each dimension. Their results show that the items that were rationally categorized within each dimension did not correlate highly with each other. The mean inter-item correlations ranged from .026 to .085.

In the same study, a second SJT containing 93 items was constructed to measure and predict the same 12 dimensions of college performance. For this SJT, the mean inter-item correlations ranged from .012 to .054. As can be seen, the two SJTs within this study did not yield strong evidence that a common construct underlies the items categorized within the same dimension. Although the items were perceived to be related to a given dimension the data did not bear the same result. Consequently, it is difficult to conclude that the dimensions that were set out to be measured by the SJTs actually did measure those dimensions.

It was stated above that another factor in addition to the inter-item correlations that may drive internal consistency estimates is the number of items within the measurement method. As will be shown, SJTs may yield high internal consistency estimates, but this is often for those that contain many items. However, this is not always the case. Some SJTs have quite low internal consistency estimates while containing many items. The next section reviews studies that present both internal consistency estimates as well as the number of items in the SJT in order to examine the typical reliability estimates for SJTs and how many items are required to achieve acceptable reliability. The review will start with studies published prior to 2001 and will then mover to more recent studies.

Internal consistency estimates. McDaniel et al. (2001) conducted a meta-analysis on the criterion-related validity of SJTs and as part of their analyses they reported the range of reliability estimates that were gathered from the studies they included in their meta-analysis. This distribution shows a wide range of estimates ranging from .43 to .94 with 50 percent of the studies (n = 16) having values of .79 or below. Considering that the rule of thumb for acceptable

reliability is .70 (Nunnally & Bernstein, 1994) many studies obtain less than acceptable reliability.

More recent studies show the same pattern of results. Chan and Schmitt (2002) constructed an SJT intending to measure one's overall ability to respond and adapt well to work-related situations. Their SJT contained 40 items and had an estimated alpha of .73. Similarly, Hunter (2003) constructed an SJT to measure pilot judgment. His test contained 51 items and yielded an alpha of .75. Lastly, Oswald et al.'s (2005) SJT predicting college performance contained 57 items and had an alpha of .85. Other studies show lower internal consistency estimates even though they contain many items. For example, Weekley and Ployhart's (2005) SJT designed to predict managerial performance contained 58 items and an alpha of .56. Lievens et al. (2005) constructed an SJT containing 30 items resulting in an alpha of .39.

Although several of these SJTs have what is deemed acceptable reliability these findings are rather disappointing in terms of suggesting a solid underlying structure because most SJTs contain 30 items or more. Cortina (1993) gives us a point of reference when examining reliability estimates and the number of items in SJTs. In this article, he shows that to obtain acceptable alpha levels a test does not need to have a large number of items, provided that each item in the test intercorrelates to only a moderate degree (i.e., .30 or greater). Recall from the Oswald et al. (2005) study that the mean intercorrelations for their two SJTs ranged from .012 to .085. He shows that alpha levels of .88 can be achieved with as few as 12 items and that when the items intercorrelate more highly (e.g., .50 or .70) alpha levels approach .95 and above. When looking back at the reviewed SJTs that contain 30 or more items in light of Cortina's (1993) findings we

can safely conclude that items within SJTs don't correlate even moderately with each other. In addition, these results suggest that there may be multiple constructs that underlie SJTs.

Cortina (1993) helps us out here again. He illustrated how multidimensionality affects coefficient alpha estimates. Within the two primary conditions of number of items and the intercorrelation of those items he also varied the number of independent dimensions underlying the items. He shows that when the number of dimensions increases alphas levels drop across each of the primary conditions. For example, under the condition of 18 items that intercorrelate .30 the alpha level decreased from .88 for one dimension to .75 for two dimensions to .64 for three dimensions. Thus, on the basis of these results by Cortina (1993) we can not only conclude that SJTs contain items that don't intercorrelate even moderately, but also that they are multidimensional.

With these findings in hand the next step to understanding what SJTs measure is to examine which items correlate most with each other and which items don't. In other words, factor analysis data should be investigated in order to get a glimpse of the underlying structure of SJTs. The next section reviews several studies that present factor analysis results on SJTs.

Factor analysis. Chan and Schmitt (2002) designed an SJT with the intent of measuring a single, general construct termed adaptability. In order to determine if they achieved this goal they conducted a principal axis factoring on the common variance, and found that a single, dominant factor emerged. However, the amount of variance accounted for by this single factor was only 16%. They noted that no other interpretable factors emerged from the data.

It seems that the extraction of a single interpretable factor from the SJT supports the case for construct validity. However, because such a small amount of systematic variance was

associated with the single factor it is difficult to see these findings as strong evidence for a single dimension underlying the SJT. If there was measuring a single, general dimension then we would expect this dimension to account for much more variance.

Similar results were found for an SJT built by Oswald et al. (2005). Their SJT was intended to measure 12 dimensions of college student performance. These authors conducted an exploratory factor analysis and found that a single factor could explain a larger amount of variance than the other, smaller factors that emerged from the data. They note that this large factor accounted for about three times the amount of variance as the smaller factors and that these smaller factors were not readily identifiable. In contrast to the study above, these authors intended to measure 12 dimensions, yet they failed to find evidence to support this a priori structure. In essence, both studies set out to measure a different number of dimensions, but found similar factor analysis results.

Although these two studies do not provide very strong construct validity evidence, not all SJTs have rendered the same results. Chan and Schmitt (1997) created a video-based SJT that was designed around the dimensions of work habits and interpersonal skills. They first conducted a content analysis of the items and found that four dimensions emerged from the rational categorization. These dimensions were work commitment, work quality, conflict management, and empathy. They conducted a multiple-groups confirmatory factor analysis in order to ensure that the same underlying structure existed across the two groups. The results of their analysis revealed that a four-factor model fit the data significantly better than a single-factor model.

Based on the review of the psychometric data underlying SJTs it suggest that SJT items correlate very lowly with each other and this is likely because multiple dimensions underlie

SJTs. Furthermore, these findings are suggestive that SJTs are not only multidimensional at the test level, but also at the item level because of the inability to extract a set of interpretable factors. This is likely because each item requires multiple constructs to be used when responding, but different sets of constructs are being utilized across each item. In other words, traits A, B, and C may be utilized for one item, but B, D, and E are used for another.

Construct Validity of SIs

Cognitive ability. One of the most studied and important individual difference variables in I/O psychology is cognitive ability. Similar to SJTs, cognitive ability has been proffered as one of the main reasons for the criterion related validity of SIs. Conceptually, this is because the interviewee must develop their own response to the hypothetical dilemma posed to them. During this process they must evaluate alternative courses of action in response to the information provided in the dilemma. As such, the interviewees are engaging in evaluative, problem solving, and decision making processes (Huffcutt et al., 2001; Huffcutt, Roth, & McDaniel, 1996). If this is the case that interviewees are engaging in processes related to cognitive ability then we would expect to see a relationship between SI ratings and cognitive ability test scores.

The correlation between general cognitive ability (GCA) and SI ratings is one of the most researched issues within the SI literature. However, not all research is specifically focused on SIs. Some of the research simply looks at structured interviews in general or examines a correlation between composite interview ratings that are composed of both SI and BDI questions (e.g., Bobko, Roth, & Potosky, 1999; Campion, Campion, & Hudson, 1994; Conway & Peneno, 1999; Konig, Melchers, Kleinmann, Richter, & Klehe, 2007). The nature of the GCA-SI correlation in the literature is such that while many individual studies have been conducted there have been several meta-analyses across the years that have in one way or another refined the previous analysis. As such, because of the number and recency of these studies it is more effective to review the meta-analyses one by one and describe the findings and how each is different from or builds on the previous.

The first major review of the correlation between GCA and interview ratings was by Huffcutt et al. (1996). Their meta-analysis included studies that contained correlations between interviews and measurements of GCA. They divided their analyses according to the level of structure implemented in the interview as well as the question content (SI or BDI). As was noted previously, SIs are the most structured type of interview followed by BDIs. The results they found for both the level of structure and the question content are virtually the same. Specifically for the SIs they calculate the observed correlation to be .21 and the fully corrected correlation to be .32.

In a follow up re-examination of Huffcutt et al.'s (1996) meta-analytic data Bobko et al. (1999) utilized only the data from the medium and high structure categories and weighted that data according to the sample sizes of the studies. In Huffcutt et al.'s study they found a mean observed correlation of .25 for medium structure interviews and .23 for high structure interviews. When re-analyzed by Bobko et al. the resulting correlation was .24. These results are also very similar to those found by Cortina et al. (2000). Considering that SIs contain at least a medium level of structure and that the results for medium and high levels of structure were essentially identical, Bobko et al.'s results appear to be a good initial estimate of the cognitive saturation of SIs.

Salgado and Moscoso (2002) were the next to examine the relationship between the SI and GCA. The purpose of their study wasn't to only examine SIs, but to examine several types of interviews and their correlation with other constructs. Among the types of interviews was the SI and among the constructs was GCA. It should be noted that the studies the authors included were not wholly redundant with those included in Huffcutt et al.'s (1996) meta-analysis. There are at

least a few studies that were included in this analysis that were not included in the earlier one. For this reason it is not surprising that this analysis show virtually identical results and Huffcutt et al. (1996). These authors found a .33 corrected correlation between SIs and GCA. Recall from earlier, Huffcutt et al. (1996) found a .32 corrected correlation. These specific findings don't add much to the literature because of the near identical results, but to this point they were the most up-to-date.

The latest meta-analysis was conducted by Berry, Sackett, and Landers (2007). This meta-analysis not only included an updated literature base, but also a more specific examination of the range restriction corrections that are typically used to estimate population parameters in meta-analyses. The updated literature base was an expansion of the studies included in Huffcutt et al. (1996), Cortina et al. (2000), and Salgado and Moscoso (2002). However, the authors took an extra step and excluded studies where it was evident or there was reason to believe that raters had access to or knowledge of interviewees' cognitive ability scores. The reason for the exclusion of these studies is because it would inflate the correlation artificially (Huffcutt et al., 1996).

The analyses they performed were different than those performed in previous studies because they chose to refine the correction of range restriction based on the type of restriction that occurred in the base studies. They provide one example where a study is conducted on data collected from a set of applicants who are chosen for a position on the basis of interview and cognitive ability test score, and while the applicants may have been screened with, for example, an application, these applicants represent an unrestricted applicant pool because they are considered the pool that is worthy for consideration of the target job. In this instance they did not

make a range restriction adjustment. In another example they note that a direct range restriction adjustment would be inappropriate in situations where a sample of job incumbents were used in a study when these incumbents were selected on the basis of, for example, an earlier interview. In this instance an adjustment for indirect range restriction was made. A number of other more refined range restriction adjustments were also made.

Through their more rigorous study selection process five studies where the authors were sure the interviewers did not have access to the interviewees' cognitive ability test scores and where it was possible to make known range restriction adjustments were retained. Their findings, however, given this relatively small study sample size was consistent with previous research. They calculated a .26 mean observed correlation and a .34 corrected correlation between SI ratings and GCA. Again, this is not surprising considering that much of their literature base was based on the previously reviewed meta-analyses.

Personality. The basic premise of SIs is that by presenting a dilemma and asking the interviewee to respond with what they would do allows for the interviewee to describe their intentions to behave in that particular situation (Latham et al., 1980). The idea is that there is some correspondence between what one says they would do and what one actually does. Some evidence supports this proposition. Latham and Saari (1984) found that one's descriptions of what they would do correlated with their observed behavior on the job. Although the main underlying theoretical mechanism is one's intentions, there may be another mechanism.

Personality has been defined as one's pattern of thought, emotion, and behavior (Funder, 2007). We can think of personality as a set of behavioral tendencies. Moreover, we can think of these behavioral tendencies as manifesting themselves across situations and as a function of the

situation (Funder & Ozer, 1983; Lievens et al., 2008; Mischel, 1973; Robie, Born, & Schmit, 2001; Schmit et al., 1995). In other words, an external observer will be able to see some consistencies in one's behavior while they are at home, at work, with friends, with family, and so on; people will tend to act similarly across these environments. However, the observer should also be able to spot some differences in their behavioral tendencies, for example, at work where the range of behaviors deemed appropriate is much more restrictive versus at home where the range of appropriate behaviors is much wider (Mischel, 1973).

With regard to SIs, the situations are contextualized. Therefore, one may respond to the situational dilemmas in a way that illustrates how they would typically behave in similar situations. A recent theoretical explanation of the personality saturation of SJTs by Motowidlo et al. (2006) will be useful for explaining why this may be the case for SIs. These authors state that one's personality trait levels determine how they perceive various responses to situations. For example, if a person has high levels of Agreeableness they will tend to think that agreeable responses are highly effective. Conversely, these individuals will tend to think that disagreeable responses are highly ineffective. However, individuals who have low levels of Agreeableness will not make such polarized distinctions. They will tend to perceive the effectiveness of agreeable and disagreeable responses more towards the middle of the scale.

SJTs have the responses already provided to the individual whereas in SIs the individual has to generate the response themselves. With regard to SIs, an individual with high levels of Agreeableness may describe behaviors that are highly agreeable in response to a situation where they are asked how they would handle a situation where a customer is, for example, upset and yelling at them. In contrast, an individual with low levels of Agreeableness may describe

behaviors that are not particularly agreeable or disagreeable. Other authors have made similar propositions that by asking someone how they would handle a given situation that their behavioral tendencies may be expressed (McDaniel et al., 2007; Roth et al., 2005).

For SIs it's not only that the interviewee expresses personality, but also whether the interviewer perceives personality from the interviewee during the interview. In order for there to be a correlation between SI ratings and measurements of personality there must be variance associated with the interviewee's personality in the interview rating, and this variance must come from relatively short interactions between an interviewer and interviewee. Funder and Colvin (1988) conducted research on the correspondence personality ratings across various sources (i.e., self, stranger, and acquaintance). They found that the largest correspondence (i.e., correlation) was between the self and the acquaintance rather than between the self and stranger. This makes sense because the acquaintance has a larger sample of the person's behavior than the stranger. The sampling of behavior (i.e., interview questions) is important for SIs. This will be discussed in more detail later.

Other research by Blackman and Funder (1998) shows the ability of people to get a glimpse of one's personality within a relatively short amount of time. These authors found that after five minutes of watching a videotape on an individual that judges tended to agree with each other about the person's personality. However, the ratings did not match the individual's self-reports of their own personality as much as if the judges viewed the person's behavior for 30 minutes. In other words, as the judges watched more and more of the person's behavior their perceptions of the person's personality began to match more and more the individual's own perceptions. Consequently, the length of interaction was a moderator.

A popular taxonomy of personality used within I/O psychology is the Big Five. These traits have been shown to predict performance for many different jobs and have also been shown to show up in interviews (Huffcutt et al., 2001; Van Dam, 2003). Huffcutt et al. (2001) conducted a meta-analysis on the dimensions typically found in interviews. They categorized the dimensions according to the Big Five model and found that of all the 5 traits Conscientiousness had the highest number of dimensions categorized at 55. Smaller numbers were obtained for Extraversion (21), Neuroticism (21), Agreeableness (10), and Openness to Experience (6). What this shows is that when interview dimensions are developed a large number of dimensions are related to Conscientiousness, Extraversion, and Neuroticism.

In a more recent study, Van Dam (2003) collected over 700 trait descriptions generated by interviewers who conducted relatively unstructured interviews. In reference to Huffcutt and Arthur's (1994) categorization, Van Dam states that the interviews were level 2 because only the topics to be discussed in the interviews were determined beforehand. The trait descriptions were notes taken during and after the interviews. These descriptions were then categorized according to the Big Five trait model. After the trait descriptions had been categorized the author calculated the number of descriptions that made it into each Big Five personality factor. She found that Extraversion and Agreeableness were the two most common noted trait descriptions, both about 22-25% of the time. Less frequently noted were Openness to Experience (18%), Neuroticism (18%), and Conscientiousness (17%). These two studies confirm that interviews do contain personality-related dimensions and interviewers do make personality-related notes on the interviewee.

The next issue to consider for personality perceptions is whether the correlations between the self and a stranger (i.e., interviewer) are the same across each of the Big Five personality traits. In other words, some traits may be more easily observed in interviews because of their inherent nature and, therefore, more likely to obtain higher inter-rater agreement (Funder & Dobroth, 1987; Russell & Zickar, 2005). For example, Extraversion is defined as sociability, gregariousness, and engaging the environment. These behaviors should be readily observable because they relate to someone interacting with others. Therefore, the perception is going to be heavily based on the extent to which a person is observed to be exhibiting these behaviors. On the other hand, a trait like Neuroticism may be less observable, especially within short interactions, because one may not have been put in a situation where they express or exhibit behaviors related to irritation, nervousness, or anger. In other words, the situation cues individual that these behaviors are not appropriate (Mischel, 1973). In addition, Neuroticism is more related to one's internal states or feelings (Russell & Zickar, 2005). Potentially, in an interview, there may be a correlation for Extraversion, but not for Neuroticism.

To address this issue, Connolly, Kavanagh, and Viswesvaran (2007) conducted a meta-analysis on a set of studies that contained correlations between various sources of personality ratings. The two sources of interest for this review are the self and a stranger. They accumulated data for these two sources across each of the Big Five personality traits. Their findings show that not all of the Big Five traits obtain significant correlations between the self and strangers. The mean observed raw correlations for Conscientiousness, Extraversion, and Openness to Experience were .23, .29, and .14 respectively while the correlations for Neuroticism and Agreeableness were .05 and -.01 respectively. These results show that while stranger's ratings of

an individual's personality do correlate with one's self ratings the predictive ability does not extend to all traits.

Coupling the findings from the studies reviewed so far leads to the conclusion that with a relatively short exposure to an individual one can get gain a perception of the individual that becomes closer to the individual's own perception, but this limited exposure tends to restrict the ability to gain an understanding of certain traits. In addition, personality trait descriptions and personality-related dimensions tend to show up in interviews. What these findings don't tell us is the ability to predict someone's behavior on the basis of limited exposure. Because SIs are used to predict behavior in certain contexts, an examination of one's ability to predict behavior well on the basis of relatively short observations is key.

Funder and Colvin (1988) conducted a study where they gathered a group of strangers and a group of acquaintances and had them judge a target person's personality. The acquaintance simply judged the person from their own memory whereas the stranger watched a five minute videotape of a person talking to another. These two groups then were asked to predict the target person's behavior in another situation similar to the one in which the strangers viewed earlier. Both groups were shown to be about as accurate in their predictions. We wouldn't necessarily expect such accuracy from the strangers because they only watched the person for five minutes, but because the context was similar to the one in which they had already viewed they were able to predict how they might act. Thus, even though the stranger had only viewed the person's behavior for a short period of time their understanding of how that person might act in another, similar situation was just as accurate as someone who knows the person much better. This has obvious implications for interviews and predicting behavior in a particular context (i.e., on-the-

job) because, in essence, an SI allows us to collect information across a range of contexts that are job-related. In other words, because an SI contains a range of situations, part of what may be driving the criterion-related validity is the predictions made on the basis of one's understanding of an individual's tendencies. The next set of reviewed studies deal specifically with interviews and personality.

A recent study by Barrick, Haugland, and Patton (2000) examining the ability to perceive personality by strangers was conducted in the context of an actual interview. In this instance, the strangers were either interviewers that conducted the interview or were individuals that watched a 15-second video of the interviewee. They also collected perceptions of the interviewee's personality from their friends. Seventy three students completed mock interviews performed by human resource practitioners using whatever interview type they were used to conducting. Thus, some interviews were behavioral and some interviews were situational, but there were no assignments made for which type the interviewers were to use. Both the interviewers and the interviewees completed a personality scale immediately after the interviews were completed, and the friends completed the personality scale later and mailed it back to the researchers.

The results of the study support the previous findings by showing that the interviewers were fairly accurate in their perceptions of the interviewee's personality in comparison to the interviewee's own perceptions. They found three significant correlations between the interviewer's ratings and the interviewee's (i.e., self) ratings; specifically, Agreeableness (.30), Extraversion (.42), and Openness to Experience (.34). What is interesting in these findings is that for two of the significant correlations they were higher for the interviewers than for the friends. The correlation for the interviewers for Agreeableness was .30 whereas the correlation for the

friends was .20. In a similar pattern, the correlation for Openness to Experience for the interviewers was .34, but for the friends it was .25. Not only is there a correspondence between the interviewee and the self ratings, but also the correspondence is as strong as or stronger than their friend's ratings.

To begin to address the correlation between SIs and personality we can look to an interesting set of findings within the Barrick et al. (2000) paper. The authors divided their analyses according to the level of structure and content (i.e., behavioral or situation) of the interviews used. In reference to situational interviews, looking across each of the sources and examining the correlation between interviewer-self and friend-self the results show that the interview source was as accurate as or more accurate than the friends. For example, for Extraversion the correlations for interview-self and friend-self were approximately the same, about .50. Ratings of Agreeableness showed stronger correlations with the interviewer than the friends (.27 vs. .16). The same thing resulted for Openness to Experience. The correlation between the interviewer and the self was .42 whereas the correlation friends and self was .37. The results were positive, but not the same pattern for Conscientiousness. In this case, the correlation between the interviewer and self was .28 and the correlation for friends and self was .36. While the correlation was higher for the friends, the .28 correlation stills suggest a fair amount of accuracy for this trait. The results for Neuroticism were not as encouraging because while the correlation for the interviewer was larger than for the friends it was much smaller than the others at .17.

These preliminary results suggest that interviewers (or strangers) can perceive an individual's personality from verbal responses, and that they can do so within a short period of

time. However, the results also suggest that not all traits are equally observable. To further establish if SI ratings correlate with self-report measurements of personality a set of studies that used SI questions will be reviewed next. Across the studies that have utilized SI questions in one way or another they have primarily done so in a way that prevents us from drawing definitive conclusions about the effectiveness of SIs in measuring personality. Many of the studies have not only included SI questions, but also BDI questions and calculated a composite interview score, which is then correlated with self-report personality ratings. Therefore, it is difficult to ascertain the personality saturation in SIs from these studies. Studies that have utilized only SIs will be reviewed later.

We know from the meta-analysis conducted by Huffcutt et al. (2001) that interviews often contain personality-related dimensions. In fact, the Big Five personality dimensions, as a group, are the most frequently included dimensions in interviews. Within the Big Five, Conscientiousness-related dimensions are the most frequent. To start the review, I'll begin by reviewing a meta-analysis by Cortina et al. (2000). This study investigated the incremental validity of interviews above cognitive ability and Conscientiousness. They gathered a large set of studies and used meta-analytical procedures to estimate the correlation between interview ratings and measurements of Conscientiousness. This study, however, did not divide the analyses according to interview type, only level of structure. The results can still be of value because they provide results for level 3-4 interviews, which include both situational and behavioral interviews. Thus, the results of this study can give us, at least, a starting point for the estimating the correlation between SIs and personality.

In all, the authors gathered nine correlation coefficients at the 3-4 level of structure. The mean raw correlation was .206, and .258 once corrected for range restriction. Of particular interest is that all of the variance across the nine coefficients was due to sampling error and range restriction. Thus, this estimate, although composed of only a few studies, does not show signs of any moderators and, therefore, the differences occur largely across studies. With more correlations, and studies that are specific to SIs, perhaps this estimate may differ.

Another meta-analysis examined the correlation between interviews and personality However, as noted above, this study does not contain results that are specific to SIs. Instead Salgado and Moscoso (2002) created a category of "behavioral" interviews that was almost exclusively composed of questions dealing with experience, past behavior, and future behavior. As such, the "behavioral" interview category is mixed with regard to content. Their results show relatively small correlations between interview ratings and measurements of personality. The largest raw observed correlation was for Extraversion at .10. Each of the other dimensions was below this value. The authors concluded that for practical purposes because the correlations were so small that interview ratings were unaffected by one's personality.

These two meta-analyses have somewhat conflicting results. Cortina et al. (2000) report estimated raw and corrected correlations with Conscientiousness of at least .20 whereas Salgado and Moscoso (2002) report estimated raw and corrected correlations of .08 and .17 respectively for Conscientiousness. Salgado and Moscoso included 13 correlations in their study whereas Cortina et al. included nine. Thus, much of the difference between the studies is likely due to the data sets. Salgado and Moscoso report that at least 98 percent of the variance in the correlations for each of the Big Five except Openness to Experience was due to sampling error. Similar to

Cortina et al.'s results, this suggests that correlations between personality traits and interview ratings will differ from study to study. Because these two studies can only be compared on finding for Conscientiousness a few more studies will be reviewed.

Huffcutt, Weekley, Weisner, DeGroot, and Jones (2001) conducted an interesting study with a sample of participants from higher-level positions. Pulakos and Schmitt (1995) argued that SIs may not be appropriate for higher-level positions because of the nature of the questions. They found that when participants responded to the situational questions they tended to answer with every possible scenario that could occur. However, some interviewees responded to the questions with a more straight-forward answer. Thus, there is some reason to believe that situational questions for higher-level positions present some complexities that make it difficult to rate, and this may also extend to ratings of personality.

Huffcutt et al. (2001) used a sample of district managers from a retail chain and asked of them both situational and behavioral interview questions. Both sets of questions were developed from critical incidents and were made to measure the same performance dimensions. They divided their analyses according to interview type and found that neither the situational nor the behavioral interview correlated highly with personality. For SIs, no personality trait measurements correlated with the interview ratings beyond -.13 or .14. For the BDIs only Conscientiousness correlated at .30. These results may give us a glimpse into the findings from the previous two meta-analyses. It may have been that BDIs were driving the correlation with Conscientiousness and not SIs.

A recent study by Morgeson et al. (2005) provides us with another example of a study where SI and BDI questions were developed, made similar to each other, and a composite score

was developed. These authors developed an interview containing both SI and BDI questions to measure social skills related to teamwork. Fourteen questions for each type were asked of incumbent steel mill workers and this interview score was correlated with measurements of Conscientiousness, Agreeableness, Neuroticism, and Extraversion. The authors found that the interview score correlated most highly but not statistically significantly with Agreeableness (.14) and Extraversion (.11) whereas Conscientiousness (.02) and Neuroticism (.02) correlated essentially zero.

In a later study, Van Iddekinge, Raymark, and Roth (2005) found discrepant results. They collected interview and personality ratings from students who participated in mock interviews for a customer service manager position. Through the use of a job analysis they identified Agreeableness, Neuroticism, and Conscientiousness as the most important personality traits. They constructed both behavioral and situational interview questions to measure each trait and calculated a composite overall interview score on the basis of ratings across these two types of interview questions. The results were positive. The authors found that interviewer ratings of Agreeableness correlated .33 with self-reports, and ratings of Conscientiousness correlated .29 with self-reports. However, the interview ratings for Neuroticism did not correlate significantly (.10).

So far, the results of the studies reviewed are mixed, but consistent with the sampling error found in the two meta-analyses. Across studies there are some positive findings and some negative findings. Two of the studies with higher level jobs (district managers and steel mill teams) found very small correlations whereas the just-mentioned study using students applying for a fictitious customer service position found larger correlations. The most important aspect of

these studies is that conclusions specific to SIs cannot be drawn because they included both situational and behavioral questions. However, as shown by Huffcutt et al. (2001) BDIs have not been found to correlate highly with personality (with the exception of Conscientiousness). Thus, the aggregated results reviewed thus far may not be a function of aggregation across interview types. Two later studies that specifically used SIs will give us a clearer indication of the personality saturation.

The first study is by Roth et al. (2005). They used a sample of sales associates within the retail clothing industry and built an SI containing six items to measure 10 dimensions of sales associate performance. Ratings for each question as well as an overall interview score were correlated with measurements of personality. None of the correlations with overall interview scores were statistically significant, and only two (i.e., Conscientiousness and Neuroticism) were larger than an absolute magnitude of .10 (or -.10 for Neuroticism). However, when the individual questions were examined two statistically significant correlations were found for Neuroticism (-.16 and -.18).

The latest study by DeGroot and Kluemper (2007) developed an SI to measure six aspects of customer service performance and gathered measurements of Agreeableness, Extraversion, and Conscientiousness from a sample of customer service employees. An interesting aspect of this study is that measurements of personality were made within two different frames of reference. Participants were asked to describe themselves as they are normally outside of work and to describe themselves as they are at work. They found that SI ratings correlated .22 with Extraversion in the normal frame-of-reference, .28 with Extraversion

in the at-work frame-of-reference, and .17 with Conscientiousness in the normal frame-of-reference. No significant correlations were found for Agreeableness in either frame-of-reference.

From these studies, it is reasonable to draw several conclusions. First, there is evidence that shows that individuals can perceive another individual's personality even within a relatively short amount of time. Second, the research is mixed regarding the personality saturation of interviews, generally, and SIs, specifically. While some research show correlations in the .20 range other studies show, essentially, a zero correlation. From the mixed results the most consistent findings are for Conscientiousness and Extraversion. Agreeableness showed positive results in some studies (i.e., Barrick et al., 2000; Van Iddekinge et al., 2005), but failed to be replicated in another study by DeGroot and Kluemper (2007). There isn't much information for Openness to Experience beyond Barrick et al.'s (2000) study, and the results for Neuroticism are not positive. These findings fit within the argument that Neuroticism may not be as easily observable as other personality traits.

Job experience. The last construct to review is job experience. A case can be made in either direction on whether SIs should correlate with measurements of experience. The first argument would be that SIs should not correlate with experience because the questions present hypothetical dilemmas that do not necessarily require someone to have prior experience in that situation (Pulakos & Schmitt, 1995). Because SIs are hypothetical one does not have to rely on their experiences to formulate a response. Instead, interviewees can analyze the dilemma presented, develop plausible response, and evaluate the effectiveness or appropriateness of those responses. The research presented above on the relationship with cognitive ability shows that interviewees are engaging in analytical and evaluative processes while developing their

responses. Thus, cognitive ability plays a major role in the development of their responses, not experience.

The counter argument is that SIs would correlate with experience because an individual draws on everything they have in order to develop their response, not just cognitive ability. In other words, while evaluating various responses to the dilemma they are also utilizing their experiences in various situations. These experiences may be similar to or identical to the dilemma presented and, therefore, one's experiences would play a role in the development of their response. To see which of these two arguments is correct a set of studies that present correlations between SI ratings and measurements of experience will be reviewed.

To begin the review, I will start with a later study, present the results, and then backtrack to an earlier study because the later study will help with the interpretation of the earlier study. Pulakos and Schmitt (1995) developed an SI to predict the performance of employees in high level positions with a federal organization. They operationalized experience in terms of the number of years on the job and found that most study participants had from one to six years of experience. They correlated SI ratings with both the measurements of performance and experience and presented two sets of results for these data. First, they showed that the SI correlated -.02 with performance and .04 with experience. Second, when they partialed out experience from the relationship between the SI and performance they found that the r-squared did not change significantly. This was expected because the SI did not correlate with either performance or experience measurements. While this information is not helpful beyond the bivariate correlations between the three variables, they do provide us with a framework to interpret the results of an earlier study.

Latham et al.'s (1980) seminal study on SIs provides similar data on the correlation between ratings and experience. Although they did not provide bivariate correlations between SI ratings and measurements of experience they did provide indirect evidence on these relationships. Latham and colleagues measured experience as the number of years on the job as a sawmill worker. They correlated their SI with an overall summed rating across nine performance dimensions as well as ratings on a behavioral observation scale (BOS) that included up to 13 items. They found that their SI correlated .50 with overall ratings and .46 with the BOS ratings. Next, they included the experience measurements in a model containing the SI and performance and examined the change in r-squared once experience was partialed out. They found that the r-squared changed from .50 to .46 for the overall performance measurements and from .46 to .41 for the BOS measurements.

These results can be interpreted by referring to the Pulakos and Schmitt (1995) article and examining them conceptually. Because the r-squared values did not change significantly for either the overall or BOS criterion it suggests that the SI is not correlated with performance, experience is not correlated with performance, experience is not correlated with the SI, or some combination of these. We know that the SI is correlated with performance so this explanation is not possible. Thus, we are left with measurements of experience being uncorrelated with either or both the SI and performance. Looking back to the small r-squared changes when experience is partialed out indicates that experience is neither correlated with the SI nor performance. If measurements of experience were significantly correlated with either the SI or performance then a portion of either of these variable's variance would be cut out and the r-squared value would change more than .04 or .05. In other words, experience does not overlap enough with either the

SI or performance to increase the prediction significantly. As a result, we can conclude that experience measurements in this study did not significantly correlate with SI ratings.

Pulakos and Schmitt (1995) argued that SIs may not be appropriate for high level jobs and showed that their SI did not correlate with performance. In addition, they also showed that their SI did not correlate with measurements of experience. One may interpret these results as an indication that SIs may have difficulty correlating with many individual difference variables for those individuals in high level jobs. One study is not conclusive evidence, but Huffcutt et al. (2001) report similar findings. Their SI was administered to a sample of district managers in the retail industry and was correlated with measurements of experience operationalized as years on the job. They also found a small correlation between the two (i.e., .08).

Thus, on the basis of these two studies, we may have evidence that suggests that there is something about high level jobs that prevents SIs from measuring experience. However, while these studies show null findings the evidence is not conclusive. Day and Carroll (2003) developed an SI for admissions to a fictional academic program. Their sample included currently enrolled students and experience was operationalized as the number of years in the university. This job, although fictional, would likely be considered a low level job. They report a correlation of .00 between SI ratings and measurements of experience. The evidence thus far is consistent in that is shows that SIs don't correlate with measurements of experience. However, the evidence is mixed regarding whether the nature of the job changes the correlation.

Thus far, the research appears to the support the proposition that SI responses are primarily a product of one's cognitive ability and not experience. However, later studies do not support this proposition. For example, Gibb and Taylor (2003) developed an SI to predict social

worker performance. They noted that Pulakos and Schmitt's (1995) findings may have suffered from range restriction because the experience was between one and six years. Thus, Gibb and Taylor sought to combat this and measured experience in terms of months as a social worker. Their sample ranged from one month to 15 years of experience. These authors found a .49 correlation between their SI and measurements of experience. In an earlier study, Conway and Peneno (1999) utilized a sample of students who applied for a resident assistant's position and found a correlation of .29. In Conway and Peneno's study they operationalized experience as the number of leadership functions they performed in extracurricular activities.

In these two studies the operationalization of experience was different; both were more specific than the previous studies. Thus, this implies that different operationalizations may yield different results. However, one point of interest is that for SJTs the general the operationalization of experience the more positive the results. Despite the different types of jobs or levels of specificity of the experience measurement the aggregated findings for the correlation between SIs and experience are mixed. At this point in the extant research it is safe to say that SIs can correlate with measurements of experience, but it is still too early to make a general statement that all SIs correlate with experience. In addition, it is difficult to estimate the magnitude of the correlation between the two.

Psychometric data. Following along the same path as for SJTs I will review the psychometric evidence regarding the construct validity of SIs. The intent is to make a review that is as comparable to SJTs as possible. In this way, more solid conclusions about the underlying structure of each method can be made. Moreover, this information, coupled with the research on the external correlates, which will be presented later, can give us some indication of what is

measured in each method and whether these constructs are the same. Evidence regarding both inter-item correlations and factor analysis will be reviewed. Inter-rater reliability estimates will also be reviewed even though this form of reliability does not exist for SJTs. The reason to make this brief review is because interviews often utilize two or more interviewers who do not always agree on the effectiveness of the responses. Theoretically, the combination interviewers used are from a universe of possible combinations of interviewers. Thus, the inter-rater reliability gives us an estimate of the correlation between a given set of interviewers and another set of interviewers drawn at random from the universe of interviewers. Following the review of the psychometric data for SIs will be a review of the correlations between SIs and various constructs. The constructs included in the review will be the same as those reviewed for SJTs, namely, cognitive ability, personality, and job experience.

Inter-rater reliability. Taylor and Small (2002) meta-analytically compared SIs and BDIs and as part of their analysis they reviewed the inter-rater reliability for both types of interviews. They collected a total of 15 studies published between 1980 and 2001 that supplied inter-rater reliability estimates for SIs. The estimates from these studies ranged from .55 to .90 with a mean of .79 and median of .81. This indicates SIs have acceptable and moderately high inter-rater reliability. Later studies support these findings on the stability of interviewers' ratings.

Sue-Chan and Latham (2004) conducted a study that used an SI to predict team playing behavior and academic grades. They report an inter-rater estimate of .83. Similarly, Morgeson et al., (2005) developed an interview that contained both situational and behavioral interview questions. They reported an inter-rater reliability for this interview of .89. Although the interview contained two types of interview questions, Taylor and Small (2002) have shown that SIs and

BDIs have similar inter-rater reliability. Two more studies, one by Peeters and Lievens (2006) who developed an SI to measure interpersonal skills, adaptability, and perseverance and Klehe and Latham (2006) who developed an SI to measure team playing behavior, report similar inter-rater reliability estimates, .78 and .90, respectively. Overall, these studies show that SIs have moderately high inter-rater reliability and that ratings are likely to be stable across multiple sets of interviewers.

Inter-item correlations, internal consistency, and factor analysis. In this section, inter-item correlations, internal consistency estimates, and factor analysis will be reviewed. In the same section for SJTs these three forms of psychometric evidence were split apart into separate sections. However, for SIs there is much less overall psychometric evidence and, therefore, it's much more efficient to summarize the research concurrently. Similar to SJTs, inter-item correlation evidence for SIs can be found in the form of inter-item correlations themselves, but also inferred from coefficient alpha estimates. Not all of the studies included in this review provide both inter-item correlations and internal consistency estimates. Therefore, if a study provides only internal consistency estimates then some amount of induction is needed to estimate the inter-item correlations.

It was mentioned that inter-rater reliability would not be reviewed and this is because, in contrast to inter-rater reliability estimates wherein the raters are of interest and form the columns of an intercorrelation matrix, with factor analysis the dimensions are of interest and form the columns in the matrix. In other words, because the focus is on the underlying structure of SIs the pattern of correlations among dimensional ratings must be examined. As noted earlier in the SJT section inter-item correlations, coefficient alpha, and factor analysis are related to each other.

Therefore, this section of the review will bring together each of these three aspects to show their relationship.

Because interviews contain fewer items than do SJTs if we find coefficient alpha estimates in the same range (.60 to .80) as was found for SJTs then this would suggest that the item ratings correlate fairly highly with each other. Several studies have examined the internal consistency of their SI by estimating coefficient alpha. Huffcutt et al. (2001) developed an SI containing six questions to measure important aspects of naval officer performance. The alpha estimate for this SI was rather low at .40. Recall that an alpha of this magnitude has been obtained for SJTs, but they often contain 30 or more items. As a result, this suggests that the items in this SI correlate more highly than do items in an SJT. Indeed, this is what the inter-item correlation matrix shows. While only a few correlations reach statistical significance many are above .10 and some reach magnitudes of .20 and .30.

In the second study within this article, Huffcutt and colleagues developed an SI to measure management skills for district managers of a retail organization. The SI contained 10 questions and the alpha was estimated to be .38. Within the inter-item correlation matrix 45 correlations are presented and most of these are .25 or lower. If we refer back to the inter-item correlations for SJTs the research shows that few, if any, items correlate higher than .10. In fact, many items were in the .05 range. Thus, these two studies by Huffcutt and colleagues shows that SI inter-item correlations can reach magnitudes much higher than SJTs.

Conway and Peneno's (1999) SI measuring five job performance dimensions for Resident Assistants also shows moderate to high inter-item correlations. The inter-item correlations for this SI ranged from .34 to .57 with an average inter-item correlation of .48. Klehe

and Latham's (2006) SI measuring teamplaying behavior obtained an alpha estimate of .50, which is similar in magnitude to Huffcutt et al.'s two studies. However, Klehe and Latham did not provide inter-item correlations. In a much earlier study by Latham and Saari (1984) an alpha estimate of .73 was found for their 20-question SI measuring clerical worker performance dimensions. And yet in an even earlier study by Latham et al. (1980) they estimated coefficient alpha of their two SIs to be .71 and .67.

In terms of how these inter-item correlations relate to factor analysis results a study by Roth et al. (2005) can provide us with information. They developed an SI to measure six dimensions of performance for workers in a retail organization. The SI contained six questions and the inter-item correlations ranged from .41 to 60 with an average inter-item correlation of .55. The authors also conducted an exploratory factor analysis are extracted a single factor that accounted for 63% of the total item rating variance. In an earlier study by Pulakos and Schmitt (1995), they also extracted a single factor from an SI measuring seven dimensions of performance for incumbents in a federal organization. Similarly, Sue-Chan and Latham (2004) was only able to obtain a single-factor solution for their 6-item SI measuring teamwork behaviors.

In summary, the psychometric data, some of it being relatively sparse, does show several things. First, SIs have acceptable and moderately high inter-rater reliability, which suggests that ratings across multiple sets of interviewers are relatively stable. Second, the SI items tend to correlate with each other much more highly than do SJT items. Third, although the item intercorrelations are higher for SIs than for SJTs it does not relate to higher internal consistency estimates because of the much smaller number of items. Fourth, because of the moderate to high

inter-item correlations only a single factor is able to be extracted from the intercorrelation matrix even though the SIs were intended to measure multiple dimensions of performance. Similar to SJTs, because of this it is difficult to draw conclusions about one's standing on a particular dimensions and also inappropriate to do so. Although the research shows that a single extracted factor explains a large part or majority of the ratings variance it does not tell us what SIs correlate with or measure.

APPENDIX D: CRITERION VALIDITY

According to Nunnally and Bernstein (1994), "No amount of apparently sound theory can substitute for a lack of a correlation between a predictor and criterion" (pp. 95). This statement couldn't be more important for SJTs and SIs because, as noted previously, the construction of SJTs is largely completed with a content oriented approach. Though content validity is important, especially for legal defensibility, it is a necessary, but insufficient condition justifying the use of a test to make selection decisions. In addition, until recently, there has been no guiding theory that underlies SJTs. As a result, without a correlation between SJT scores and a criterion of interest, there is little justification for them to be used. Not only does a correlation with a criterion satisfy legal requirements, but also as the correlation between a test and a criterion increases the utility of the decisions that HR personnel make increase (Hunter, Schmidt, & Judiesch, 1986; Cascio & Ramos, 1986).

Criterion-related Validity for SJTs

For many years, researchers have examined the relationship between SJTs with various job performance criteria across a variety of jobs. Much of this research has been conducted since the 1990 publication by Motowidlo, Dunnette, and Carter and summarized in a meta-analysis by McDaniel et al. (2001). Thus, this section will start with the paper by Motowidlo et al. and then move to the McDaniel et al. meta-analysis, and finish with the research published after 2001. In addition, the focus will be on I/O-related SJTs and not those that are tacit knowledge methods which look a lot like SJTs. Both the zero-order correlations with performance as well as incremental validity over other measures will be reviewed with the zero-order correlations section divided into service and non-service jobs.

Motowidlo et al's (1990) SJT was built around two aspects of manager performance in the telecommunications industry. They used three samples of managers, one sample included those who were promoted from outside the company, the second included those promoted from within, and the third included those who were applicants seeking employment into the company. The participants' jobs ranged from engineering and programming to administration and supervision to marketing/sales.

Performance ratings were collected for both incumbent samples, but not the applicant sample. Ten performance dimensions were included in the rating forms and were categorized into interpersonal, problem solving, and communication dimensions. In addition, overall performance ratings were collected. Although the authors created a short-form (30-item) SJT, the results from the initial 58-item SJT will be presented.

The results for the sample of managers hired from outside the company shows that the SJT correlated .35 with ratings of interpersonal skill, .28 with ratings of problem solving skill, .37 with ratings of communication skill, and .30 with ratings of overall performance. These findings provide support for the predictive validity of this low-fidelity simulation. When these correlations are compared with those found with assessment centers and work samples we do not see a tremendous difference in predictive ability (Schmidt & Hunter, 1998).

McDaniel et al.'s (2001) meta-analysis set out to determine the best estimate of the criterion-related validity of SJTs. They searched the literature to gather the relevant data on SJTs and job performance and found 39 SJTs with data that could be included in the analysis. The studies included were from publications in scientific journals, paper presentations at professional conferences, or unpublished doctoral dissertations. In addition, they also gathered data from

commercially published tests such as *How Supervise?* and the *Supervisory Judgment Test*. In all, their search and data gathering process yielded data from over 10,000 participants and 102 correlation coefficients for SJTs and job performance. The range of jobs studied included insurance agents (Dalessio, 1994), managers and marketing positions in the telecommunications industry (Motowidlo & Tippins, 1993; Olson-Buchanan et al., 1998), supervisors from across administration, academic, and business areas within an academic institution, (Olson-Buchanan et al., 1998), sales (Phillips, 1992), employees at a pulp mill and cardboard box plant (Stevens & Campion. 1999), and retail employees (Weekley & Jones, 1999).

Meta-analytical procedures were applied to estimate the population correlation coefficient between SJTs and job performance. The authors found that the mean, uncorrected correlation coefficient was .26 and the corrected population correlation coefficient was .34. An interesting finding was that the lower-bound 90^{th} percentile credibility value was greater than zero (i.e., .16) indicating that the correlation between SJT scores and ratings of job performance was generalizable across study characteristics, which includes different types of jobs and performance criteria. More specific findings from this study show that *How Supervise?* and the *Supervisory Judgment Test* also correlate meaningfully with performance (.21 and .41 respectively).

Overall, this study provides ample evidence regarding the criterion validity of SJTs across a wide range of jobs and types of performance criteria. So important are these results that they solidify the practicality of SJTs and also give some indication that job-related knowledge, skills, and abilities are being measured. Moreover, in combination with the content validity guidelines from Motowidlo et al. (1990), the legal defensibility of SJTs is bolstered.

Administrative and customer service jobs. One influential and often-cited study is by Clevenger et al. (2001). They collected data from samples across three types of jobs, one of which was customer service personnel from the transportation industry. The SJT, which used a rate the effectiveness response instruction, was built to measure dimensions related to customer service performance which included customer focus, organization, team leadership, decision making, communication, and business knowledge. On the criterion side, ten dimensions of job performance were rated by supervisors on a behaviorally anchored rating scale. The results show a positive correlation (.18) between SJT scores and job performance ratings. This correlation is consistent with earlier studies utilizing customer service samples (Weekley & Jones, 1997; 1999).

In another study, Chan and Schmitt (2002) administered an SJT, designed to measure the overall ability to adapt to work situations, to 160 civil service employees in administrative positions whose main duties were to provide staff support. Supervisors rated their performance on four dimensions including technical proficiency, job dedication, interpersonal facilitation, and overall performance. Chan and Schmitt (2002) showed that their SJT correlated .30 with core technical proficiency, .38 with job dedication, .27 with interpersonal facilitation, and .30 with overall performance.

Weekley and Ployhart (2005) conducted a study wherein they hypothesized several different ways that an SJT could relate to job performance (i.e., either no, partial, or full mediation) in a model with other measured individual difference variables. The most important finding in this study is the simple bivariate correlation between their SJT that contained an equal number of management-related situations and loss prevention situations, and a 47-item scale

measuring an individual's performance completing required tasks. The data were collected from 271 employees in a loss prevention management positions in a retail organization. To derive an employee's performance score the authors averaged the supervisor ratings across the 47 items to create one performance score. Their results are consistent with the studies reviewed above. They show a .22 correlation with the averaged performance ratings.

Finally, Chan (2006) set out to determine the conditions in which a proactive active personality may end in bad outcomes. To test his hypotheses he needed to develop and administer an SJT. His SJT was developed from job analysis information and there was no specific indication that a set of dimensions were being measured. However, he notes that his SJT was developed in a similar way to the SJT developed by Chan and Schmitt (2002). Thus, it may measure one's overall ability to adapt to work-related situations.

He administered this SJT to 139 employees within a rehabilitation agency. Performance ratings on three items were also gathered and made by the employees' supervisors. It is difficult to determine the nature of the performance ratings because ratings were not made on a set of dimensions and overall performance may or may not have been measured. Nevertheless, he obtained a .26 correlation between the SJT and performance ratings for rehabilitation officers.

In summary, recent studies provide evidence across a range of relatively different jobs that is consistent with McDaniel et al.'s (2001) findings regarding the predictive ability of SJTs. Although some of the SJT-performance correlations are higher than others, all are positive. These findings are very informative for issues such as validity transportability and selection system design. However, they do not represent the full range of jobs that SJTs have been applied to and criterion validity evidence has been gathered. To complete a review of the full range of

the jobs for SJTs an expanded review must be made beyond what could be considered a more typical set of jobs like those reviewed above. The next section completes the expanded review.

Non-administrative or customer service jobs. One study applied an SJT in an environment which is currently dominated by higher fidelity computer-based simulations and flight simulators (Fritzsche, Stagl, Salas, & Burke, 2006). Specifically, Hunter (2003) constructed an SJT to measure aviation judgment and predict flying safety. Items were constructed from critical incidents that were gathered from pilot and accident reports. Several themes emerged from these incidents and formed the content areas of the SJT. The areas include, but are not limited to, weather phenomena, mechanical malfunctions, and air traffic control requests.

The sample included pilots who volunteered to complete the SJT over the Internet while visiting a Federal Aviation Administration sponsored website. In addition, volunteer pilots also completed a scale measuring the number of hazardous events and accidents they have experienced. They hypothesized that good judgment, as measured by the SJT, should be correlated with fewer experienced hazardous events and accidents. In fact, SJT scores correlated -.215 with the number of hazardous events and accidents experienced by pilots.

These findings have implications for the use of SJTs. First, they provide evidence that SJTs can predict performance in non-managerial contexts. Second, they show that a low-fidelity SJT may be an alternative when the use of a flight simulator is infeasible or impractical. In essence, this study mirrors the purpose of the Motowidlo et al. (1990) study, which investigated whether a paper-pencil "simulation" could be used in place of a high-fidelity work sample or assessment center to predict performance.

The next three studies investigated SJTs in yet another, different environment than those that were previously reviewed. Studies by Lievens, Buyse, and Sackett (2005), Oswald et al. (2004) and Ployhart and Ehrhart (2003) examined the correlation between SJT scores and college performance. The study by Lievens et al. (2005) used a predictive rather than a concurrent design. The findings from the Lievens et al. (2005) study are particularly relevant to the practical side of using SJTs because college admission decisions were made on the basis of SJT scores as well as other measurements.

Lievens et al. (2005) collected data from several universities who used the same admissions system in which the SJT was one of the tests administered. The other tests administered were four tests covering physics, chemistry, biology, and mathematics, a cognitive ability test, and an exam requiring students to read and answer questions about a 10-page text covering medical topics. The SJT was intended to measure communication and interpersonal skills. Critical incidents were gathered in order to generate the content. The authors had access to students' grades across the five year span of the program and created two categories of GPAs. This was done because the authors noticed that across the universities some emphasized interpersonal skills within the course material more than others. Thus, a GPA was calculated for the curricula that did not emphasize interpersonal skills to a high degree and for the curricula that did emphasize interpersonal skills.

Their results show an interesting pattern. First, the authors noticed that for the curricula that did not emphasize interpersonal skills, the relative weight of the interpersonal courses in determining GPA was quite small. Conversely, the curricula that emphasized interpersonal skills, the relative weight of the interpersonal courses had higher weights in determining GPA. Second,

the authors noted that the weights for the interpersonal courses in the non-interpersonal curricula remained constant across the four years of the program. In contrast, the weights for the interpersonal courses in the interpersonal curricula increased across the four years.

These differences in the relative weights translated into a set of correlations between SJT scores and GPA showing that the SJT became more predictive across time for the curricula stressing interpersonal skills. Specifically, the authors found that the SJT predicted GPA .07 in year 1, .10 in year 2, .27 in year 3, and .38 in year 4. By contrast, the SJT did not predict GPA for any of the four years for the curricula that did not emphasize interpersonal skills.

Just one year earlier, Oswald et al. (2004) also set out to predict college student performance with an SJT. However, their study did not administer an SJT as part of an admissions process, but collected the data from current students. In their study they administered a personality scale, a biodata inventory, and collected ACT/SAT scores along with an SJT designed to measure 12 dimensions of college student performance. The major performance dimension categories were intellectual behaviors, interpersonal behaviors, and intrapersonal behaviors with more specific dimensions nested underneath each. For example, intellectual behaviors included mastery of knowledge, interpersonal behaviors included interacting with others and citizenship, and intrapersonal behaviors included life skills, perseverance, and integrity. Self-ratings and peer ratings were collected for the 12 dimensions. In addition to these two sets of ratings, GPA and absenteeism data were also gathered.

Oswald et al.'s (2004) findings parallel and add to those of Lievens et al. (2005). They report a significant correlations between SJT scores and GPA (.16), absenteeism (-.27), self-ratings (.53), and peer ratings (.16). Lievens et al. (2005) only collected GPA as the performance

criteria. However, in this study, we have evidence of the ability of an SJT to predict other aspects of college student performance beyond GPA.

The last of the college studies by Ployhart and Ehrhart (2003) used a different criterion for college students than was used by Oswald et al. (2004). They gathered self and peer ratings of study behaviors along with self-reported GPA and SAT scores. The authors built an SJT that contained situations that students were likely to face at college while studying. The authors also developed different forms of this SJT by using "would do" and "should do" response instructions. The results of this study showed that the different response instruction had an effect on the correlation between SJT scores and performance. However, an important finding is that the SJT with "would do" response instructions significantly correlated about .30 to .50 with GPA and SAT scores and self and peer ratings of study behaviors. Note that some of these correlations were corrected and some were not because of differences in samples sizes.

The last of the jobs, and the most different from those reviewed in this section, is within the manufacturing industry. McDaniel et al. (2001) included in their analysis a study by Stevens and Campion (1999) that used a sample of employees within a pulp mill and cardboard box plant. However, in contrast to more typical practice their study focused on situations relating to teamwork rather than individually-based situations.

O'Connell et al.'s (2007) paper gives us a more standard view of SJTs within this industry. O'Connell et al. sought out to measure the interpersonal aspects of jobs ranging from truck engine manufacturing to television manufacturing to glass manufacturing by writing 10 situations that dealt with giving advice, diffusing disputes, and demonstrating empathy. Performance ratings were made on both task and contextual performance dimensions.

The correlations they report between the SJT and performance are not as large as those reviewed above. The SJT only correlated with task performance ratings .14 and contextual performance ratings .10. The authors noted that a limitation of this study was that it used incumbents and exhibited range restriction in the sample. Thus could be a possible explanation for the small correlations. Another possible explanation for these findings is that these jobs are highly routine and don't require a large amount of interpersonal interaction. If this is the case then an interpersonally-based SJT will not tend to correlate highly with task performance because of the low variability of the behaviors exhibited and because task performance is not conceptually related to interpersonal dimensions. Moreover, if interpersonal behaviors are not frequently observed by supervisors then they may not be able to effectively differentiate the interpersonal behaviors exhibited by the employees. It is unknown from the descriptions of the sample and work in the study whether this possible explanation has merit, but these findings do call for more research to be conducted on SJTs in manufacturing jobs.

Criterion-related Validity of SIs

The interview, and its various forms, is so widely used in selection systems that we would hope that it has some ability to predict job performance. Otherwise, the time spent conducting the interview would be wasted (Reilly & Chao, 1982), and the interview itself would primarily serve to make the decision makers feel good because they had input into the decision making process (i.e., voice). For example, Phillips and Dipboye (1989) conducted a study to test several propositions for their dynamic model of the interview process. In this study, they found that when an interviewer's initial impression, which was based on an examination of an applicant's qualifications, was positive that they spent more time recruiting the applicant to the organization rather than collecting job-related information. In addition, they found that interviewers' initial and post-interview impressions did not add to the prediction of applicant's performance on a simulation of work tasks beyond that predicted by a test of numerical reasoning and knowledge of economics. In both cases, the interview didn't add much utility to the selection system.

As a result of this study and other early reviews of the interview there is the obvious question of whether the use of the interview is justified. A large body of research has tackled this question and will be reviewed here. Keeping in line with the previous section, a review of the criterion-related validity of interviews without regard to specific types of interviews (i.e., SIs) will be made first. Then a review of the criterion-related validity of SIs will be made followed by a review across a wide range of jobs in which an SI has been used to predict performance.

Early on the criterion-related validity of the interview was not well regarded (Hunter & Hunter, 1984; Reilly & Chao, 1982; Ulrich & Trumbo, 1965; Wagner, 1949; Wiesner &

Cronshaw, 1988). Generally speaking, the early reviews were mainly narrative, did not use statistical methods to aggregate findings, or did not have a large enough data set in which to draw firm conclusions (e.g., Hunter & Hunter, 1984; Ulrich & Trumbo, 1965). Thus, it was difficult to make a determination about the ability of the interview to predict performance across a wide range of jobs. In their narrative review, Ulrich and Trumbo (1965) stated that because the studies they reviewed were different in terms of purpose, sample, and setting that one should be cautious about drawing conclusions about the predictive ability of the interview. However, they did find that the most systematic interviews had the highest predictive validity. This offered some confidence toward the interview.

Despite this lukewarm review of the interview research continued. About seventeen years after Ulrich and Trumbo (1965) another review was made by Reilly and Chao (1982). They collected 12 correlation coefficients and found that the mean correlation with supervisor ratings was .19. A later study by Hunter and Hunter (1984) which included the correlations from Reilly and Chao (1982) found similar results. These authors meta-analytically aggregated interview data and examined the correlations across four types of criteria. In all, they collected 27 correlations and found that the interview correlated .14 with supervisor ratings of job performance, .08 with promotion, .10 with training success, and .03 with tenure.

Up to this point in the research one would be hard-pressed to say that the interview was one of the more predictive selection methods. Yet, some researchers were not convinced that the previous research was the definitive answer. In part, this seems to be because it was clear that not enough data had been collected and that data from interviews of differing levels of quality were combined. For example, in Reilly and Chao's (1982) review they cite many studies where the

interview was semi-structured, and therefore, probably suffered from a lack of reliability and possibly examination of consistent job-related content across interviewees.

In order to remedy the shortcomings of these reviews and build on the research already conducted, Wiesner and Cronshaw (1988) applied meta-analytic procedures to the data from a much larger set of interview studies to arrive at an overall conclusion of whether or not the interview had predictive ability and to examine the extent to which the structure of the interview and the number of individuals conducting the interview moderated the validity.

Wiesner and Cronshaw (1988) collected a large number of studies with correlations between interview scores and job performance ratings. Compared to Hunter and Hunter (1984) who only collected 30 correlation coefficients, these authors ended up with 150 correlations and aggregated them using meta-analytic procedures suggested by Hunter et al. (1982). In addition to estimating the mean criterion-related validity, they also conducted two more specific analyses to estimate the criterion validity of structured and unstructured interviews as well as individual and panel interviews.

In contrast to the earlier reviews, their results clearly support the use of the interview for predicting job performance. The mean uncorrected correlation was .26 and the corrected correlation was .47. However, the 95% confidence interval ranged from essentially zero to 1.0. With respect to the moderator analyses they found that structured interviews were much more predictive of performance than unstructured interviews (.17 vs. .34) and individual and panel interviews predicted performance to the same degree. Overall, this study showed that not only are interviews predictive of performance, albeit with substantial variability, but also that the

mean corrected validity is similar in magnitude with other tests like work sample and cognitive ability tests (Schmidt & Hunter, 1998).

Wiesner and Cronshaw's (1988) study is a valuable addition to the literature. However, more recent studies have made analyses at more specific levels of structure and across criterion types. Huffcutt and Arthur (1994) took the delineation of structure by Wiesner and Cronshaw (1988) a bit further and defined three levels of structure for both the interview question and scoring standardization. An additional benefit of the Huffcutt and Arthur (1994) study is that they examined the criterion validity of the interview for entry-level jobs. They reason that the interview is most likely conducted for entry-level jobs because this is where most selection occurs. Thus, they tackle the issues of how much structure do we need in the interview and if the interview is predictive of performance for jobs where it is most likely to be used.

The total data set they used for their analyses consisted of 114 correlation coefficients between interview scores and supervisor ratings of performance. They categorized the correlations according to their 3 x 3 structure matrix. The levels of structure are defined as follows. Level 1 was defined as a completely unstructured interview, Level 2 was defined as a rather small amount of structure in the form of a list of topic areas to cover in the interview, Level 3 had a pre-specified list of questions, but the interviewers could ask from this list whichever questions they wanted and ask follow-up probing questions, and Level 4 consisted of complete standardization wherein interviewers asked all applicants the same questions in the same order and could not ask follow-up questions. In reference to how different types of interviews fit within the categorization of structure, BDIs fit within Level 3 and SIs within Level 4.

Huffcutt and Arthur's (1994) results are in line with the findings of previous studies by showing that the mean uncorrected correlation with performance is .22. Recall from earlier that this correlation is slightly lower than Wiesner and Cronshaw's (1988), but slightly higher than Hunter and Hunter's (1984) and Reilly and Chao's (1982). The second set of their analyses showed that the level of structure did moderate the correlation with performance, but only to a certain level. Specifically, they found that as the level of structure increased from Level 1 to Level 3 the uncorrected correlation increased from .11 to .34. However, they found that from Level 3 to Level 4 the correlation did not increase. Thus, on the basis of these findings, there is an asymptote to the correlation as a function of structure.

McDaniel et al. (1994) further extended the examination of the criterion validity of the interview by categorizing and analyzing their data according interview and criterion type. The decision rules for classifying studies by interview type were the following. If the interview was conducted by a psychologist and measured traits like dependability then it was psychological, if the interview asked what interviewees have done in the past then it was job-related, and if the interview asked interviewees to describe what they have done in the past then it was situational. McDaniel et al. (1994) utilized the same criterion types as Hunter and Hunter (1984), namely, job performance, training success, and tenure.

These authors' findings further substantiate the predictive ability of the interview in general, but also, in reference to the purpose of this study, the situational interview. They found that the overall mean uncorrected correlation with job performance across all interview types was .20. When they specifically estimated the correlation for SIs they found that it was a bit higher at .27. In addition, once corrected for range restriction and criterion unreliability the

standard deviation of the population estimate for the SI was only .05 thereby indicating a fair amount of generalizability. While the correlation found for SIs is meaningful and justifies their use it was only based on 16 coefficients. Therefore, we cannot consider this to a robust finding.

In summary, there is ample evidence to suggest that structured interviews are predictive of performance. These studies built on each other by increasing the data set and examining more and more specific characteristics of the interview in order to find where boundary conditions for the criterion validity may lie. Because McDaniel et al.'s (1994) study included so few correlations for the SI, several more recent meta-analyses on the predictive ability of the SI have been conducted.

Two meta-analyses have been conducted on the criterion validity of the SI specifically. One was by Latham and Sue-Chan in 1999 and the second was by Taylor and Small in 2002. These two studies will be reviewed concurrently because both utilized a majority of the same studies to estimate mean and corrected population correlation coefficients. However, the Taylor and Small study went a step beyond Latham and Sue-Chan's by further estimating correlations across three levels of job complexity.

First, Latham and Sue-Chan (1999) examined previous attempts at summarizing the criterion validity of SIs and found several limitations in those reviews. They stated that previous reviews by Huffcutt and Arthur (1994) and McDaniel et al. (1994) were unclear in terms of the studies they categorized as SIs because there was no information regarding whether they were built from job analysis information, utilized formal scoring guidelines, and included a dilemma in the stem of the questions. Thus, they questioned whether the SIs included in their data set

were actually SIs according to the criteria listed by Latham et al. (1980) and this study. Furthermore, they stated the Huffcutt and Arthur's study only looked at entry-level jobs.

The data that Latham and Sue-Chan (1999) included were only those that met the full criteria of an SI, which are listed above. This created a relatively small data set. However, Taylor and Small's (2002) study did not make such restrictions. In other words, Latham and Sue-Chan's data was a subset of Taylor and Small's. Because of the different inclusion criteria Latham and Sue-Chan collected a total of 20 correlations whereas Taylor and Small collected 30.

Although the data sets differed the findings between the studies did not. Latham and Sue-Chan found a mean correlation with performance of .29 (corrected = .39) and Taylor and Small found a correlation of .25 (corrected = .45). The lower-bound 95% confidence intervals for both correlations were positive and did not include zero. However, in contrast to Latham and Sue-Chan who report a lower-bound estimate of .26, Taylor and Small report a lower-bound estimate of .09. As we can see that the more relaxed inclusion criteria used by Taylor and Small resulted in more variability in the population estimates. This is likely due to the fact that some studies utilized less structured or standardized SIs or the study itself was flawed.

Taylor and Small's additional analysis across three levels of job complexity also yielded positive results. They categorized jobs according to the type of work they do along two dimensions, namely, Data and Things. Examples of Data include actual data, numbers, or blueprints and examples of Things include tangibles objects like computers (Brannick & Levine, 2002). High complexity jobs require people to synthesize or analyze (i.e., Data) and set up, work with precision, or operate (i.e., Things). By contrast, low complexity jobs require people to copy or compare (i.e., Data) and tend to or feed (i.e., Things). A couple examples of jobs that are high

complexity would be management or professional jobs and low complexity jobs would unskilled jobs such as a filing clerk.

Considering the relatively large disparity in the types of work that high and low complexity jobs entail (with medium complexity residing somewhere in the middle) Taylor and Small did not find any significant differences in the mean correlations with performance between the three levels of complexity. Low complexity jobs showed a .26 correlation, medium complexity jobs .24, and high complexity jobs .23. Although some authors have posited that SIs may not correlate with performance for high complexity jobs (Pulakos & Schmitt, 1995) these findings did not support a moderating effect.

The results across the job complexity levels give us some information about the generalizability of the results. However, they do not address the extent to which the results may extend to specific jobs or industries, only jobs that may fall within several broad categories. For these findings to have the highest possible usefulness it would be beneficial if the studies conducted were across a diverse range of jobs. In Taylor and Small's (2002) analysis there was roughly the same number of studies across each job complexity level, and if you will recall all of the studies in Latham and Sue-Chan (1999) were included in Taylor and Small's analysis. Although Latham and Sue-Chan did not make such job complexity distinctions it is reasonable to assume that the studies they included were relatively evenly distributed across the three levels.

To address the more specific generalizability of these two studies we can refer back to Latham and Sue-Chan because they actually listed the range of jobs that data were collected for. In all, they had data for 16 jobs that included but were not limited to clerical employees, bank tellers, school custodians, pulp mill workers, sales representatives, first line supervisors, and

university faculty. With a brief read-through of this list of jobs the range of job complexity should be evident. Therefore, the studies by Latham and Sue-Chan (1999) and Taylor and Small (2002) suggest that SIs are predictive of job performance across a wide range of jobs.

Leadership, administrative, & customer service jobs. The findings in the criterion validity section are very useful for estimating the ability of the SI to predict performance across a range of jobs that differ in complexity, but what is more specific to generalizability is the ability of the SIs to predict performance for specific jobs. One limitation of the Taylor and Small's (2002) analysis of SIs across job complexity levels is that there were relatively few correlations found for each of the three levels. They found about 10 correlations for each level and while this may be enough to begin the discussion of generalizability we need more information in order to draw more solid conclusions. To build on their work a review will be made that not only accounts for different levels of job complexity, but also references specific job types. The jobs that will be reviewed include military officers (Huffcutt, Weekley, Wiesner, DeGroot, & Jones, 2001), police officers (Maurer, Solamon, and Lippstreu, 2008), managers (Huffcutt et al., 2001; Krajewski, Goffin, McCarthy, Rothstein, & Johnston, 2006), social workers (Gibb & Taylor, 2002), and customer service representatives (DeGroot & Kleumper, 2007).

Huffcutt et al. (2001) set out to investigate whether Pulakos and Schmitt's (1995) findings that SIs don't predict performance for complex jobs can be replicated with another sample. To accomplish this they conducted two studies in jobs that have at least moderate to high complexity. The first study was with candidates for Naval officer training that had completed their university education. A total of 59 candidates applied to the training program and 37 were

selected. Although it is not specifically stated it appears that SI scores were not used to make selection decisions.

The SI was built using traditional methods. The authors identified 10 dimensions of job performance and gathered ratings on a 5-point scale from staff members at the training academy for those selected. The general areas of the dimensions were leadership, administrative, and interpersonal. Once all of the data were gathered they found results similar to those by Pulakos and Schmitt (1995). They report a correlation of .20 between the SI and overall ratings of job performance. Although the magnitude of the correlation is consistent with previous research (Taylor & Small, 2002) it did not reach statistical significance. One reason for the lack of significance may be due to the power associated with the analysis. If more data points had been gathered then this correlation may have reached conventional significance levels. However, Huffcutt et al. (2001) conclude that their findings replicate those of Pulakos and Schmitt (1995) and suggest that SIs don't predict performance complex jobs well.

Their second study was conducted with district managers of a national chain of merchandise stores. Whereas it is somewhat difficult to determine the level of complexity for the first sample of military officers this study is clearer. The most appropriate level of complexity seems to be high because the manager's duties were reported to be oversight of several stores, and, in addition, Taylor and Small (2002) categorized the job as having a high level of complexity.

Just like the previous study the SI was build with traditional methods. On the basis of the job analysis results 10 dimensions of performance were identified and categorized into three main areas (i.e., Action, Leadership, and Interpersonal). This SI contained a question for each

dimension of performance and, in contrast to the previous study, was conducted over the phone. 93 district managers were interviewed and their performance was determined with a rank ordering by their regional managers. These ranking were then converted into quintiles in order to combine the ranking across regions.

In contrast to the previous study, however, which found a positive, but non-significant correlation, the results for this sample were rather disappointing. The authors found a .02 correlation between SI ratings and the performance rankings. This raises an obvious question. Is there something about the manager position that makes the SI inapplicable in terms of predicting performance? One explanation that has been put forth by Pulakos and Schmitt (1995) and Huffcutt et al. (2001) suggests that the situations for high complexity jobs can't adequately be described in SI questions. Considering the complexity of a district manager's position this suggestion may have some merit. Another study that used a sample of managers will shed some extra light on this issue.

Krajewski et al. (2006) noted that previous studies conducted with SIs for higher level jobs contained some methodological limitations. They note that the studies by Pulakos and Schmitt (1995) and Huffcutt et al. (2001) used a concurrent rather than a predictive design. Also, in contrast to the prescriptions made by Latham and Sue-Chan (1999), the study by Pulakos and Schmitt (1995) made an overall rating of the interviewee's performance at the end of the interview rather than after each question. In addition, the results from other studies (i.e., Taylor and Small, 2002) don't yield the same results. Thus, Krajewski et al.'s study was conducted in order to remedy some of the methodological flaws and to add more data to the literature base.

These authors conducted a predictive study involving 157 applicants for high-level management positions. As is typical with many management positions these manager's primary duties were to make sure that policies and procedures were followed and to direct and evaluate the performance of their subordinates. The SI built to measure job-related knowledge, skills, and abilities was done so according to typically prescribed procedures and the questions in the SI spanned six performance dimensions (i.e., planning and organizing, coaching, results orientation, willingness to learn, team orientation, and oral communication). Performance ratings across eight specific dimensions and one overall dimension were made by managers who did not have information about the individuals' test scores one year after the applicants completed the SIs and were selected. The performance ratings were derived in a different manner than is typical of most studies. Managers rated the applicants relative to each other rather than making an absolute rating of one's performance. Each of the relative performance ratings across the nine dimensions was aggregated to form a single performance value.

Krajewski et al. (2006) found a non-significant and fairly small correlation of .09 between overall SI ratings and job performance. The discussion by the authors of why the SI failed to predict performance can be summed up in terms of what the SI may have measured. The authors state that the responses to the SI trended toward work style and values rather than, for example, reasoning, deduction, and perhaps work experience. And it is because of this that a low correlation was likely to have resulted.

In another study utilizing a sample with higher level jobs, Maurer et al. (2008) findings with a sample of police officers whose main duties and performance dimensions were similar to the three samples described above are consistent with previous studies. They used a sample of

candidates for promotion to the rank of police and fire sergeants and lieutenants. These participants were asked 12 questions relating to duties relevant to each group's job. The performance dimensions were similar to the above studies (e.g., coordinating activities, evaluating employees, administrative duties, and supervising command activities). The SI was one method within a selection system that also included a knowledge test. In order to move on to the SI candidates had to pass the knowledge test. Thus, there was range restriction on the SI. They report a -.01 correlation between SI scores and ratings of performance.

What is interesting about this study and can possibly help answer the question of whether or not SIs can predict performance for higher level jobs is that about half of the police officer and fire fighter candidates attended a coaching session that included a description of the knowledge, skills, and abilities measured in the interview, participating in and observing interview role plays, and receiving sample questions and feedback from the role play. The authors noted that one of the possible outcomes of the coaching session would be to reduce error variance due to the interviewees not knowing what is being measured and, therefore, not exhibiting behavior related to the dimensions of the interview. As a result, if error variance is reduced then the correlation between SI scores and performance should increase. This is exactly what they found. For the group of candidates that were coached the correlation was .24.

Maurer et al. (2007) have brought forth an interesting issue that relates to the previous studies mentioned above. Pulakos and Schmitt (1995) noted in their study that many of the interviewees responded to the situations with every possible contingency that could have occurred. Thus, their responses may not have been specific to the dimensions of the interview. Recall from above in the Krajewski et al. (2006) study interviewees were reported to have

responded in a similar fashion. If coaching in the manner that was conducted by Maurer et al. (2007) can give the interviewees the proper perspective in which to respond and thereby reduces error variance then the reason why the previous two studies found null results may have been because of this issue. Future research could tackle this issue and determine whether or not SIs can predict performance for higher level jobs.

After an examination of four studies that focused on higher level jobs, an important next step would be to review studies using the SI to predict not only performance of employees that are one level below that of managers, but also performance of employees in different jobs. Given that the research is quite mixed on the ability of the SI to predict performance in higher levels jobs it is vital to know for which types of jobs the research on SIs is more consistent and supportive. The next two studies utilized an SI in order to predict performance for customer service employees and social workers. The purpose of the next set of reviews is to show that not all of the more current studies utilizing SIs to predict performance produce null or conflicting results.

DeGroot and Kluemper (2007) examined the SI in a concurrent design with a sample of customer service representatives from a retail organization. The purpose of their study was to examine why the SI correlated with performance. In other words, they measured other variables (i.e., vocal attractiveness and personality) that were hypothesized to be the mechanism by which SIs correlate with performance. Because this information is beyond the scope of this section, more specific information on the construct validity of the SI will be presented in a later section.

One hundred and fifty four current customer service representatives completed an SI that represented six important categories of customer service situations they typically face on the job.

A 7-point behaviorally anchored rating scale was used to rate each of the six SI questions. The performance rating scale was an adaptation of the SI questions wherein supervisors were asked to rate how well an employee has usually performed in that situation. Thus, the correlation was between interviewer ratings of how well a representative would handle a situation and supervisor ratings of how well representatives have handled the same situations on the job. Consequently, it is not surprising that the SI significantly correlated (.32) with job performance ratings.

With a different sample of employees (i.e., social workers) from New Zealand Gibb and Taylor (2002) correlated SI ratings on questions measuring four job-related dimensions with measurements of performance. The measurements of performance were similar to those made in the Krajewski et al. (2006) study. That is, employees were rated relative to each other rather than given an absolute rating of performance. However, in contrast to Krajewski et al., Gibb and Taylor (2002) found a significant correlation (.60) between SI ratings and performance. Recall that there are a couple main differences between these two studies beyond the differing results. First, is that the Krajewski et al. (2006) study utilized a predictive design whereas the Gibb and Taylor (2002) study used a concurrent design. The second difference is that the previous study used managers whereas Gibb and Taylor used an arguably lower complexity job in social workers. Thus, there are two possible reasons why these results differ; one is the design and the second is the job. Nevertheless, the studies reviewed thus far support the ability of the SI to predict performance across a wide range of jobs. However, some data suggest that the SI may not be applicable to all jobs, at least not in its currently administered form (i.e., coaching to reduce error variance). A further review of another, different set of jobs will help further clarify this issue.

Non-leadership, administrative or customer service jobs. SIs have not only been used to predict performance for rather typical jobs, but also to predict performance in new environments. Three recent studies have utilized a structured SI to predict various aspects of team behavior and performance. The first of these studies was by Sue-Chan and Latham (2004). In this study the authors developed an SI on the basis of information gathered from a critical incident technique in order to measure and predict teamplaying behavior. They focused on teamplaying behavior not only because teams are increasing in frequency within the workplace, but also because it was suggested as an important aspect of manager performance.

The SI, containing six questions and measuring teamplaying behavior, was administered to 75 professionals enrolled in an MBA program at a Canadian university. What is interesting with regard to this sample of participants is that many had job titles that were rather high level. For example, the sample included project engineers, managers of Human Resources, Vice Presidents of Marketing, and Pharmacists. Thus, the sample they used was experienced and educated. This is in contrast to some studies that utilized SIs with a sample of employees in lower level jobs. Also, considering the nature of the criterion (i.e., teamplaying behavior) and the level of the jobs that the participants had this study may have some implications for the job complexity issue reviewed above.

Two performance measurements were made in this study. The first was a course grade score that was a composite of both individual and team work and was derived at the end of one of the semesters of the program. The second was peer assessments of teamplaying behavior that were made on a behavioral observation scale. This was also completed at the end of the semester for the same course with the program. These two performance measurements yield slightly

different types of information because the course grade can be considered to be a bit like task performance because of the individual work, but also contain a little bit of contextual performance because of the team work. The teamplaying behavior, in this study, was considered to be more aligned with contextual performance. They present evidence for this by showing that cognitive ability (typically aligned with task performance) did not predict teamplaying behavior, but it did predict course grades.

The results of their study are positive. They found a .32 correlation between SI ratings and teamplaying behavior ratings. In addition, they found a .26 correlation between SI ratings and course grades. This study shows that SIs can indeed predict more complex behavior and performance. Additionally, it shows that SIs can predict different aspects of performance. Yet, this is not the only study that has examined this type of behavior. Klehe and Latham (2005) appear to have conducted another study examining the predictive ability of SIs for teamplaying behavior. This later study collected data from 79 MBA students, but it appears to be a different set of students than the study above. An SI with nine questions as opposed to six like the earlier study was correlated with peer assessments of teamplaying behavior. For this sample they found a correlation of .41 with the criterion. Thus, we have at least two studies that have found correlations between SI ratings and measurements of teamplaying behavior.

In the third and latest study Morgeson, Reider, and Campion (2005) examined the ability of a selection system composed of traditionally individually-based selection methods to select individuals for a team environment. One of the selection methods was an SI that contained 14 questions designed to measure social skills such as coordination, social perceptiveness, cooperation, and service orientation that are important for team-based activities. The study was

completed with current employees in a steel mill that worked in teams ranging from 5 to 10 members each. The work of the team was highly interdependent and required a significant amount of coordination because the team itself could not accomplish its tasks without the help of each of the team members.

Accordingly, Morgeson et al. (2005) gathered measurements of contextual performance (i.e., communicating, coordination, helping others, and team morale) as well as task performance (i.e., performing job duties, using tools/equipment, and performing routine maintenance). They made only a formal hypothesis for a correlation between SI ratings and contextual performance. However, they reported correlations for task performance as well. In all, 90 employees completed the SI and performance measurements. A .28 correlation between SI ratings and contextual performance and a .17 correlation between SI ratings and task performance was found for this sample. These are consistent with the two studies reviewed above and further show that an SI can predict the more complex, coordinated performance of teams.

In all, the studies reviewed support the view that SIs, in general, can predict performance for a wide range of jobs. Though some of the research is mixed on the ability of the SI to predict performance for higher level jobs this should not discourage us from using the SI in these contexts. Practitioners should ensure that, during an SI, interviewees' responses are consistent and that they relate to the dimensions that underlie the situations depicted in the interview. At minimum, this may help to guarantee that the SI is measuring the dimensions dictated by the job analysis information.

APPENDIX E: IRB LETTER

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901, 407-882-2012 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Notice of Expedited Initial Review and Approval

From **:** **UCF Institutional Review Board**
FWA00000351, Exp. 10/8/11, IRB00001138

To : **James S. Gunter** and Co-PI: **Barbara Fritzsche**

Date : **August 19, 2009**

IRB Number: **SBE-09-06368**

Study Title: **How do situational judgment tests and situational interviews compare? An examination of criterion and construct validity.**

Dear Researcher:

Your research protocol noted above was approved by **expedited** review by the UCF IRB Vice-chair on 08/19/2009. **The expiration date is 08/18/2010.** Your study was determined to be minimal risk for human subjects and expeditable per federal regulations, 45 CFR 46.110. The categories for which this study qualifies as expeditable research are as follows:

6. Collection of data from voice, video, digital, or image recordings made for research purposes.

7. Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

The IRB has approved a **consent procedure which requires participants to sign consent forms.** Use of the approved, stamped consent document(s) is required. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Subjects or their representatives must receive a copy of the consent form(s).

All data, which may include signed consent form documents, must be retained in a locked file cabinet for a minimum of three years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained on a password-protected computer if electronic information is used. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

To continue this research beyond the expiration date, a Continuing Review Form must be submitted 2 – 4 weeks prior to the expiration date. Advise the IRB if you receive a subpoena for the release of this information, or if a breach of confidentiality occurs. Also report any unanticipated problems or serious adverse events (within 5 working days). Do not make changes to the protocol methodology or consent form before obtaining IRB approval. Changes can be submitted for IRB review using the Addendum/Modification Request Form. An Addendum/Modification Request Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at http://iris.research.ucf.edu .

Failure to provide a continuing review report could lead to study suspension, a loss of funding and/or publication possibilities, or reporting of noncompliance to sponsors or funding agencies. The IRB maintains the authority under 45 CFR 46.110(e) to observe or have a third party observe the consent process and the research.

On behalf of Joseph Bielitzki, M.S., DVM., UCF IRB Chair, this letter is signed by:

Signature applied by Joanne Muratori on 08/19/2009 02:47:42 PM EDT

IRB Coordinator

APPENDIX F: TABLES

Table 4: Summary of Correlations for SJTs and SIs

	SJTs	SIs
Job Performance	.26[a], .18[b], .30[c], .22[d], .26[e], .10[f], .14[f], .20[k]	.34[aa], .27[bb], .29[cc], .25[dd]
Cognitive Ability	.31[a], .11[b], .17[b], .53[b], .36[g], -.02[h], .29[k]	.21[ee], .24[ff], .33[gg], .26[hh]
Personality		
Extraversion	.06[i], .20[j], .13 [k]	.10[gg] , -.01[jj], .22[kk]
Conscientiousness	.00[b], .16[b], .21[b], .25[c], .26[i], .25[j], .21[k], .30[k]	.08[gg], .21[ii], .10[jj], .13[kk]
Agreeableness	.30[c], .25[i], .30[j], .22[k]	.06[gg], .02[jj], -.04[kk]
Neuroticism	-.20[c], .17[c], .31[i], -.20[j], .17[j], .19[k]	.04[gg], -.12[jj]
Openness to Experience	.20[c], .09[i], .20[j], .11[k]	.04[gg], -.03[jj]
Experience	.01[b], .03[b], -.13[b], .21[g], .13[g], .05[i]	.04[ll], .08[mm], .00[nn], .49[oo], .29[pp]

Note. All values are uncorrected, raw correlations. [a] McDaniel et al. (2001), [b] Clevenger et al. (2001), [c] Chan & Schmitt (2002), [d] Weekley and Ployhart (2004), [e] Chan (2006), [f] O'Connell et al. (2007), [g] Weekley and Ployhart (2005), [h] Chan and Schmitt (2002), [i] McDaniel & Nguyen (2001), [j] Oswald et al. (2004), [k] McDaniel et al. (2007), [aa] Huffcutt and Arthur (1994), [bb] McDaniel et al. (1994), [cc] Latham and Sue-Chan (1999), [dd] Taylor and Small (2002), [ee] Huffcutt et al. (1996), [ff] Bobko et al. (1999), [gg] Salgado and Moscoso (2002), [hh] Berry et al. (2007), [ii] Cortina et al. (2000), [jj] Roth et al. (2005), [kk] DeGroot and Kluemper (2007), [ll] Pulakos and Schmitt (1995), [mm] Huffcutt et al. (2001), [nn] Day and Carroll (2003), [oo] Gibb and Taylor (2003), [pp] Conway and Peneno (1999).

REFERENCES

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*(2), 435-442.

Banki, S., & Latham, G. P. (2010). The criterion-related validities and perceived fairness of the situational interview and situational judgment test in an Iranian organization. *Applied Psychology: An International Review, 59*(1), 124-142.

Barrick, M. R., Patton, G. K., & Haugland, S. N. (2000). Accuracy of interviewer judgments of job applicant personality traits. *Personnel Psychology*, *53*(4), 925-951.

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

Barrick, M. R., Stewart, G. L., Neubert, M. J., & Mount, M. K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology, 83*(3), 377-391.

Bauer, T. N., & Truxillo, D. M. (2006). Applicant Reactions to Situational Judgment Tests: Research and Related Practical Issues. In J. Weekley & R. Ployhart (Eds.). *Situational judgment tests: Theory, measurement, and application.* (pp. 233-249). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology, 54*(2), 388-420.

Bell, N. L., Matthews, T. D., Lassister, K. S., & Leverett, J. P. (2002). Validity of the Wonderlic personnel test as a measure of fluid or crystallized intelligence: Implications for career assessment. *North American Journal of Psychology, 4*(1), 113-120.

Bellows, R. M., & Estep, M. F. (1954). *Employment psychology: The interview*. Oxford England: Rinehart & Co.

Bergman, M. E., Donovan, M. A., Drasgow, F., Henning, J. B., & Overton, R. C. (2008). Test of Motowidlo et al.'s (1997) theory of individual differences in task and contextual performance. *Human Performance, 21*(3), 227-253.

Berry, C. M., Sackett, P. R., & Landers, R. N. (2007). Revisiting interview-cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology, 60*(4), 837-874.

Blackman, M. C., & Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. *Journal of Experimental Social Psychology, 34*(2), 164-181.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52*(3), 561-589.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*(2), 99-109.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology, 76*(6), 863-872.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review, 111*(4), 1061-1071.

Brannick, M. T., & Levine, E. L. (2002). *Job analysis: Methods, research, and applications for human resource management in the new millennium*. Thousand Oaks, Calif: Sage Publications.

Campion, M. A., Campion, J. E., & Hudson, J. P. (1994). Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology, 79*(6), 998-1102.

Cardall, A. J. (1942). *Test of practical judgment; for 12th grade level and above*. Oxford England: Science Research Associates.

Cascio, W. F., & Ramos, R. A. (1986). Development and application of a new method for assessing job performance in behavioral/economic terms. *Journal of Applied Psychology, 71*(1), 20-28.

Chan, D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology, 82*(2), 311-320.

Chan, D. (2006). Interactive effects of situational judgment effectiveness and proactive personality on work perceptions and work outcomes. *Journal of Applied Psychology, 91*(2), 475-481.

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.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15*(3), 233-254.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates.

Colbert, A. E., Mount, M. K., Harter, J. K., Barrick, M. R., & Witt, L. A. (2004). Interactive Effects of Personality and Perceptions of the Work Situation on Workplace Deviance. *Journal of Applied Psychology, 89*(4), 599-609.

Colvin, C. R., & Funder, D. C. (1991). Predicting personality and behavior: A boundary on the acquaintanceship effect. *Journal of Personality and Social Psychology, 60*(6), 884-894.

Cook, K.W., Vance, C. A., Spector, P. E. (2000). The relation of candidate personality with selection interview outcomes. *Journal of Applied Social Psychology, 30*(4), 867-885.

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The Convergent Validity between Self and Observer Ratings of Personality: A meta-analytic review. *International Journal of Selection and Assessment, 15*(1), 110-117.

Conway, J. M., & Peneno, G. M. (1999). Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology, 13*(4), 485-506.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology, 53*(2), 325-351.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Cuber, J. F., & Pell, B. (1941). A method for studying moral judgments relating to the family. *American Journal of Sociology, 47*, 12-23.

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

Day, A. L., & Carroll, S. A. (2003). Situational and Patterned Behavior Description Interviews: A Comparison of Their Validity, Correlates, and Perceived Fairness. *Human Performance, 16*(1), 25-47.

Decker, R. L. (1956). An item analysis of How Supervise? using both internal and external criteria. *Journal of Applied Psychology, 40*(6), 406-411.

DeGroot, T., & Kluemper, D. (2007). Evidence of Predictive and Incremental Validity of Personality Factors, Vocal Attractiveness and the Situational Interview. *International Journal of Selection and Assessment, 15*(1), 30-39.

DeShon, R. (2002). Generalizability theory. In F. Drasgow & N. Schmitt (*Eds*), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 189-220). San Francisco, CA US: Jossey-Bass.

Dodrill, C. B., & Warner, M. H. (1988). Further studies of the Wonderlic Personnel Test as a brief measure of intelligence. *Journal of Consulting and Clinical Psychology, 56*(1), 145-147.

Ellis, A. P. J., West, B. J., Ryan, A. M., & DeShon, R. P. (2002). The use of impression management tactics in structured interviews: A function of question type? *Journal of Applied Psychology, 87*(6), 1200-1208.

Ethical Principles of Psychologists and Code of Conduct. (1992). *American Psychologist, 47*(12), 1597-1611.

File, Q. W. (1945). The measurement of supervisory quality in industry. *Journal of Applied Psychology, 29*(5), 323-337.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*(4), 552-564.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison-Wesley series in social psychology. Reading, Mass: Addison-Wesley Pub.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*(4), 327-358.

Frei, R. L., & McDaniel, M. A. (1998). Validity of customer service measures in personnel selection: A review of criterion and construct evidence. *Human Performance, 11*(1), 1-27.

Fritzsche, B. A., Stagl, K. C., Salas, E., & Burke, C. S. (2006). Enhancing the Design, Delivery, and Evaluation of Scenario-Based Training: Can Situational Judgment Tests Contribute? In J. A. Weekley & R. E. Ployhart (Eds.). *Situational judgment tests: Theory, measurement, and application.* (pp. 301-318). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 55*(1), 149-158.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52*(2), 409-418.

Funder, D. C., & Ozer, D. J. (1983). Behavior as a function of the situation. *Journal of Personality and Social Psychology, 44*(1), 107-112.

Gatewood, R. D., & Field, H. S. (2001). *Human resource selection*. Fort Worth: Harcourt College.

Gibb, J. L., & Taylor, P. J. (2003). Past experience versus situational employment: Interview questions in a New Zealand social service agency. *Asia Pacific Journal of Human Resources, 41*(3), 371-382.

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

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*(1), 9-24.

Hogan, J., Hogan, R., & Busch, C. M. (1984). How to measure service orientation. *Journal of Applied Psychology, 69*(1), 167-173.

Hooper, A. C., Cullen, M. J., & Sackett, P. R. (2006). Operational Threats to the Use of SJTs: Faking, Coaching, and Retesting Issues. In J. A. Weekley & R. E. Ployhart (Eds.). *Situational judgment tests: Theory, measurement, and application.* (pp. 205-232). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Hough, L. M. (1992). The 'Big Five' personality variables--construct confusion: Description versus prediction. *Human Performance, 5*(1), 139-155.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*(1), 152-194.

Huffcutt, A. I. (1993). *An empirical investigation of the relationship between multidimensional degree of structure and the validity of the employment interview.* ProQuest Information & Learning, US.

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*(2), 184-190.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*(5), 897-913.

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology, 81*(5), 459-473.

Huffcutt, A. I., Weekley, J. A., Wiesner, W. H., Jones, C., & Degroot, T. G. (2001). Comparison of situational and behavior description interview questions for higher-level positions. *Personnel Psychology, 54*(3), 619-644.

Hunt, T. (1928). The measurement of social intelligence. *Journal of Applied Psychology, 12*(3), 317-334.

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *International Journal of Aviation Psychology, 13*(4), 373-386.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 257-266). Hollsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*(1), 72-98.

Hurley, R. F. (1998). Customer service behavior in retail settings: A study of the effect of service provider personality. *Journal of the Academy of Marketing Science, 26*(2), 115-127.

Imada, A. S., & Hakel, M. D. (1977). Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *Journal of Applied Psychology, 62*(3), 295-300.

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67*(5), 577-580.

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.). *The employment interview: Theory, research, and practice.* (pp. 158-168). Thousand Oaks, CA US: Sage Publications, Inc.

Jones, A. H. (1943). A method for studying moral judgments--further considerations. *American Journal of Sociology, 48*, 492-497.

Kacmar, K. M., Delery, J. E., & Ferris, G. R. (1992). Differential effectiveness of applicant impression management tactics on employment interview decisions. *Journal of Applied Social Psychology, 22*(16), 1250-1272.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237-251.

Klehe, U.-C., & Latham, G. (2006). What Would You Do--Really or Ideally? Constructs Underlying the Behavior Description Interview and the Situational Interview in Predicting Typical Versus Maximum Performance. *Human Performance, 19*(4), 357-382.

Klehe, U.-C., & Latham, G. P. (2005). The Predictive and Incremental Validity of the Situational and Patterned Behavior Description Interviews for Teamplaying Behavior. *International Journal of Selection and Assessment, 13*(2), 108-115.

Konig, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U.-C. (2007). Candidates' ability to identify criteria in nontransparent selection procedures: Evidence from an assessment center and a structured interview. *International Journal of Selection and Assessment, 15*(3), 283-292.

Kristof-Brown, A., Barrick, M. R., & Franke, M. (2002). Applicant impression management: Dispositional influences and consequences for recruiter perceptions of fit and similarity. *Journal of Management, 28*(1), 27-46.

Krajewski, H. T., Goffin, R. D., McCarthy, J. M., Rothstein, M. G., & Johnston, N. (2006). Comparing the validity of structured interviews for managerial-level employees: Should we look to the past or focus on the future? *Journal of Occupational and Organizational Psychology, 79*(3), 411-432.

Latham, G. P., & Finnegan, B. J. (1993). Perceived practicality of unstructured, patterned, and situational interviews. In H. Schuler, J.L. Farr & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives.* (pp. 41-55). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Latham, G. P., & Saari, L. M. (1984). Do people do what they say? Further studies on the situational interview. *Journal of Applied Psychology, 69*(4), 569-573.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*(4), 422-427.

Latham, G. P., & Sue-Chan, C. (1999). A meta-analysis of the situational interview: An enumerative review of reasons for its validity. *Canadian Psychology/Psychologie canadienne, 40*(1), 56-67.

LeBreton, J. M., Barksdale, C. D., Robin, J., & James, L. R. (2007). Measurement issues associated with conditional reasoning tests: Indirect measurement and test faking. *Journal of Applied Psychology*, *92*(1), 1-16.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The Operational Validity of a Video-Based Situational Judgment Test for Medical College Admissions: Illustrating the Importance of Matching Predictor and Criterion Construct Domains. *Journal of Applied Psychology, 90*(3), 442-452.

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91*(2), 247-258.

Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology, 93*(2), 268-279.

Lievens, F., & Peeters, H. (2008). Interviewers' sensitivity to impression management tactics in structured interviews. *European Journal of Psychological Assessment, 24*(3), 174-180.

Lievens, F., & Sackett, P. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91*(5), 1181-1188.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance*. Englewood Cliffs, NJ US: Prentice-Hall, Inc.

Mandell, M. M. (1954). Ways to select supervisors. *Personnel Journal, 33*, 210-213.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

Maurer, T. J, Solamon, J.M., & Lippstreu, M. (2008) How does coaching interviewees affect the validity of a structured interview? *Journal of Organizational Behavior, 29*, 355-371.

McCarthy, J., & Goffin, R. (2004). Measuring job interview anxiety: Beyond weak knees and sweaty palms. *Personnel Psychology, 57*(3), 607-637.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63-91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*(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), 103-113.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). Job experience correlates of job performance. *Journal of Applied Psychology, 73*(2), 327-330.

McDaniel, M. A., Whetzel, D. L., Hartman, N. S., Nguyen, N. T., & Grubb, W. L., III. (2006). Situational Judgment Tests: Validity and an Integrative Model. In J. A. Weekley & R. E. Ployhart (Eds.). *Situational judgment tests: Theory, measurement, and application.* (pp. 183-203). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*(4), 599-616.

McFarland, L. A., Ryan, A. M., & Kriska, S. D. (2003). Impression Management Use and Effectiveness Across Assessment Methods. *Journal of Management, 29*(5), 641-661.

McFarland, L. A., Yun, G., Harold, C. M., Moore, L. G., & Viera, L., Jr. (2005). An examination of impression management use and effectiveness across assessment center exercises: The role of competency demands. *Personnel Psychology, 58*(4), 949-980.

McGraw, K., & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods,1*(1), 30-46.

Meng, X-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*, 172-175.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Meyer, R. D. (2010, April). *A taxonomy of work situations to help focus frame-of-reference personality tests*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80*(4), 252-283.

Morgeson, F. P., Reider, M. H., & Campion, M. A. (2005). Selecting Individuals In Team Settings: The Importance Of Social Skills, Personality Characteristics, And Teamwork Knowledge. *Personnel Psychology, 58*(3), 583-611.

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

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*(6), 640-647.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91*(4), 749-761.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). A Theoretical Basis for Situational Judgment Tests. In J. A. Weekley & R. E. Ployhart (Eds.). *Situational judgment tests: Theory, measurement, and application.* (pp. 57-81). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

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*(4), 337-344.

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.

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher-order dimensions of the big five personality traits and the big six vocational interest types. *Personnel Psychology, 58*(2), 447-478.

Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Five-Factor Model of personality and performance in jobs involving interpersonal interactions. *Human Performance, 11*(2), 145-165.

Mowry, H. W. (1957). A measure of supervisory quality. *Journal of Applied Psychology, 41*(6), 405-408.

Newman, D. (2009). Missing data techniques and low response rates: The role of systematic nonresponse parameters. In C.E. Lance & R.J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 7-36). New York, NY US: Routledge/Taylor & Francis Group.

Ng, K.-Y., Ang, S., & Chan, K.-Y. (2008). Personality and leader effectiveness: A moderated mediation model of leadership self-efficacy, job demands, and job autonomy. *Journal of Applied Psychology, 93*(4), 733-743.

Nguyen, N.T., McDaniel, M. A., & Biderman, M. D. (2002). *Response instructions in situational judgment tests: Effects of faking and construct validity.* Symposium presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Canada.

Nunnally, J.C., Bernstein I. H. (1994). *Psychometric Theory*. McGraw-Hill Series in psychology. New York: McGraw-Hill.

O'Connell, M. S., Hartman, N. S., McDaniel, M. A., Grubb, W. L., III, & Lawrence, A. (2007). Incremental Validity of Situational Judgment Tests for Task and Contextual Job Performance. *International Journal of Selection and Assessment, 15*(1), 19-29.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51*(1), 1-24.

Olson-Buchanan, J. B., & Drasgow, F. (2006). Multimedia Situational Judgment Tests: The Medium Creates the Message. In J. A. Weekley & R. E. Ployhart (Eds.). *Situational*

judgment tests: Theory, measurement, and application. (pp. 253-278). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Oswald, F. L., Friede, A. J., Schmitt, N., Kim, B. H., & Ramsay, L. J. (2005). Extending a Practical Method for Developing Alternate Test Forms Using Independent Sets of Items. *Organizational Research Methods, 8*(2), 149-164.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a Biodata Measure and Situational Judgment Inventory as Predictors of College Student Performance. *Journal of Applied Psychology, 89*(2), 187-207.

Parasuraman, A., Zeithaml, V., & Berry, L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing, 49*(4), 41-50.

Peeters, H., & Lievens, F. (2006). Verbal and Nonverbal Impression Management Tactics in Behavior Description and Situational Interviews. *International Journal of Selection and Assessment, 14*(3), 206-222.

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.

Phillips, A. P., & Dipboye, R. L. (1989). Correlational tests of predictions from a process model of the interview. *Journal of Applied Psychology, 74*(1), 41-52.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11*(1), 1-16.

Ployhart, R. E. (2006). The Predictor Response Process Model. In J. A. Weekley & R. E. Ployhart (Eds.). *Situational judgment tests: Theory, measurement, and application.* (pp. 83-105). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Pulakos, E. D., & Schmitt, N. (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology*, *48*(2), 289-308.

Putka, D., Le, H., McCloy, R., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*(5), 959-981.

Quiñones, M. A., Ford, J. K., & Teachout, M. S. (1995). The relationship between work experience and job performance: A conceptual and meta-analytic review. *Personnel Psychology*, *48*(4), 887-910.

Reilly, R. R., & Chao, G. R. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, *35*(1), 1-62.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology*, *85*(6), 880-887.

Robie, C., Born, M. P., & Schmit, M. J. (2001). Personal and situational determinants of personality responses: A partial reanalysis and reinterpretation of the Schmit et al. (1995) data. *Journal of Business and Psychology*, *16*(1), 101-117.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163-184.

Roth, P. L., & BeVier, C. A. (1998). Response rates in HRM/OB survey research: Norms and correlates, 1990-1994. *Journal of Management, 24*, 97-117.

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Schmit, M. J., & Eidson, C. E., Jr. (2005). Personality Saturation in Structured Interviews. *International Journal of Selection and Assessment, 13*(4), 261-273.

Russell, S. S., & Zickar, M. J. (2005). An examination of differential item and test functioning across personality judgments. *Journal of Research in Personality, 39*(3), 354-368.

Ryan, A. M., & Sackett, P. R. (1987). A survey of individual assessment practices by I/O psychologists. *Personnel Psychology, 40*(3), 455-488.

Salgado, J. s. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology, 11*(3), 299-324.

Schlenker, B. R. (1980). *Impression management: The self-concept, social identity, and interpersonal relations*. Monterey, Calif: Brooks/Cole Pub.

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.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71*(3), 432-439.

Schmit, M. J., Ryan, A. M. (1993). The big five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78(6)*, 966-974.

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology, 80*(5), 607-620.

Schmitt, N., & Chan, D. (2006). Situational Judgment Tests: Method or Construct? In J. A. Weekley & R. E. Ployhart (Eds.). *Situational judgment tests: Theory, measurement, and application.* (pp. 135-155). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Smith, D. B., Hanges, P. J., & Dickson, M. W. (2001). Personnel selection and the five-factor model: Reexamining the effects of applicant's frame of reference. *Journal of Applied Psychology, 86*(2), 304-315.

Spriegel, W. R., & James, V. A. (1958). Trends in recruitment and selection practices. *Personnel, 35*, 42-48.

Stevens, C. K., & Kristof, A. L. (1995). Making the right impression: A field study of applicant impression management during job interviews. *Journal of Applied Psychology, 80*(5), 587-606.

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.

Stewart, G. L., Dustin, S. L., Barrick, M. R., & Darnold, T. C. (2008). Exploring the handshake in employment interviews. *Journal of Applied Psychology, 93*(5), 1139-1146.

Sue-Chan, C., & Latham, G. P. (2004). The Situational Interview as a Predictor of Academic and Team Performance: A Study of the Mediating Effects of Cognitive Ability and Emotional Intelligence. *International Journal of Selection and Assessment, 12*(4), 312-320.

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology, 75*(3), 277-294.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*(3), 500-517.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*(4), 703-742.

Thorndike, E. L. (1920). Intelligence and its uses. *Harper's Magazine, 140*, 227-235.

Thorndike, R. L. (1991). Is there any future for intelligence? In *Improving inquiry in social science: A volume in honor of Lee J. Cronbach.* (pp. 285-303). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Thorndike, R. L., & Stein, S. (1937). An evaluation of the attempts to measure social intelligence. *Psychological Bulletin, 34*(5), 275-285.

Thorne, B. M., & Henley, T. B. (1997). *Connections in the history and systems of psychology*. Boston, MA US: Houghton, Mifflin and Company.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., et al. (2006). Unproctored internet testing in employment settings. *Personnel Psychology, 59*(1), 189-225.

Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin, 63*(2), 100-116.

Van Dam, K. (2003). Trait perception in the employment interview: A five-factor model perspective. *International Journal of Selection and Assessment, 11*(1), 43-55.

Van Iddekinge, C. H., McFarland, L. A., & Raymark, P. H. (2007). Antecedents of impression management use and effectiveness in a structured interview. *Journal of Management, 33*(5), 752-773.

Van Iddekinge, C. H., Raymark, P. H., & Roth, P. L. (2005). Assessing Personality With a Structured Employment Interview: Construct-Related Validity and Susceptibility to Response Inflation. *Journal of Applied Psychology, 90*(3), 536-552.

Van Scotter, J. R., & Motowidlo, S. J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology, 81*(5), 525-531.

Vinchur, A. J., Schippmann, J. S., Switzer, F. S., III, & Roth, P. L. (1998). A meta-analytic review of predictors of job performance for salespeople. *Journal of Applied Psychology, 83*(4), 586-597.

Wagner, R. (1949). The employment interview: a critical summary. *Personnel Psychology, 2*, 17-46.

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

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49*(2), 436-458.

Wagner, R. K., & Sternberg, R. J. (1987). Tacit knowledge in managerial success. *Journal of Business and Psychology, 1*(4), 301-312.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement, 17*(1), 11-29.

Weekley, J. A., & Gier, J. A. (1987). Reliability and validity of the situational interview for a sales position. *Journal of Applied Psychology, 72*(3), 484-487.

Weekley, J. A., & Jones, C. (1997).Video-based situational testing. *Personnel Psychology, 50*(1), 25-49.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*(3), 679-700.

Weekley, J. A., & Ployhart, R. E. (2005). Situational Judgment: Antecedents and Relationships with Performance. *Human Performance, 18*(1), 81-104.

Weekley, J. A., & Ployhart, R. E. (2006). *Situational judgment tests: Theory, measurement, and application*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*(4), 275-290.

Williamson, L. G., Campion, J. E., Malos, S. B., Roehling, M. V., & Campion, M. A. (1997). Employment interview on trial: Linking interview structure with litigation outcomes. *Journal of Applied Psychology, 82*(6), 900-912.

Wonderlic (2002). *Wonderlic Personnel Test & Scholastic Level Exam User's Manual.* Libertyville, IL: Author.

Wright, P. M., Kacmar, K. M., McMahan, G. C., & Deleeuw, K. (1995). P=f(M X A): Cognitive ability as a moderator of the relationship between personality and job performance. *Journal of Management, 21*(6), 1129-1139.