2011

The Effects Of Scoring Technique On Situational Judgment Test Validity

Daniel S. Miller
University of Central Florida

THE EFFECTS OF SCORING TECHNIQUE ON SITUATIONAL JUDGMENT TEST VALIDITY

by

DANIEL S. MILLER
B.S. University of Central Florida, 2005

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

Fall Term
2011

Major Professor: Kimberly Smith-Jentsch

ABSTRACT

Situational Judgment Tests (SJTs) are frequently used by organizations as a face-valid selection measure with low adverse impact and a relatively strong relationship with relevant criteria. Despite their common use, there remain several research questions regarding the theoretical foundations and characteristics of SJTs. Additionally, developments in SJT scoring provide fertile ground for research to validate new scoring techniques to better predict criteria of interest. Motowidlo and his colleagues (2006) recently developed a scoring technique for SJTs based on the principle of Implicit Trait Policies (ITPs) which are implicit beliefs concerning the effectiveness of different behavioral choices that demonstrate varying levels of targeted traits. Individuals high in these targeted traits will rate item responses that demonstrate high levels of that particular trait as more effective. Taking into consideration this new method, and also considering the multitude of scoring methods already available to test developers, it logically follows that these different scoring methods will have different correlations with constructs of interest, and that by using this new method it may be possible to achieve a much higher correlation with personality. The effects of scoring technique on relationships between SJT scores and constructs of interest such as personality will in turn have effects on the criterion validity of the SJT. This research explored how scoring methods affected the relationship SJT scores have with general mental ability, personality traits, typical performance, and maximum performance. Results indicated significant differential validity as a function of the respondents' race. For minority participants, SJT scores predicted "maximum performance ratings" in a simulation exercise but not "typical performance ratings" provided by familiar peers. However, the reverse was true for Caucasian participants. The two scoring methods demonstrated differential validity. However, the nature of these differences varied as a function of the

performance dimension in question (i.e., agreeableness, extraversion). Implications for future research will be discussed as well as the practical implications of these findings.

I would like to dedicate this project to my parents, Richard and Sheila Miller, and my sister, Andrea Miller, for their infinite support and encouragement.

ACKNOWLEDGEMENTS

I would like to thank all of my committee members for their advice, cooperation, and recommendations towards reaching this milestone. I would especially like to give thanks to my major advisor, Kimberly Smith-Jentsch for her guidance, encouragement, and mentoring during my time at the University of Central Florida. Without her tireless assistance, this project would not be completed. As my teacher and mentor, she has taught me more than I could ever giver her credit for here. I would like to thank the research assistants who aided me in the collection and coding of data for the extensive amount of time they invested in my study. Finally, I would like to thank my colleagues, who were an amazing support network. I feel honored to have had the opportunity to work with such incredible individuals.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Overview of Dissertation

As a low-fidelity form of simulation, situational judgment tests (SJTs) have been used by organizations for decades as an extremely face-valid method for the selection of employees. SJTs are defined as measurement techniques that share the following characteristics: they present the applicant with job-related situations, they present responses in a multiple choice format, and they have a scoring key which is developed a-priori (McDaniel & Nguyen, 2001; Weekley & Ployhart, 2006). Research has been broadening the knowledge base regarding the benefits of SJTs. These benefits include a relatively high predictive validity and having minimal adverse impact toward different races and genders while still costing a fraction relative to higher fidelity assessment methods (Motowidlo, Dunnette, & Carter, 1990). Despite these benefits, several questions have remained unanswered regarding the construct validity, theoretical underpinnings, and moderating variables of SJTs. Construct validity is an especially important issue. SJTs have been demonstrated to have inconsistent relationships with constructs of interest, suggesting there are several moderating variables to these relationships (McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006). Should researchers be able to understand which scoring technique can best capture particular constructs then tests could be developed to utilize these scoring techniques and in turn to better capture these constructs. Personality is an example of such a construct, as SJTs often have relatively low correlations with personality variables, even when they are designed to capture these variables (e.g. Motowidlo, Hooper & Jackson, 2006b; McDaniel et al., 2007). Personality is a particularly important variable to capture, as it is often a good predictor of typical performance (Marcus, Goffin, Johnston, & Rothstien, 2007) and personality traits tend to be relatively stable over time (Ferguson, 2010). However, there are also several problems

associated with using personality to predict work performance, such as self-report complications in which very transparent items will be easy to fake by participants (e.g. Viswesvaran, 2010). Thus, using SJT scoring techniques which may not directly rely on explicit self-report may afford researchers with a more valid method of personality measurement.

There are several variables that can be manipulated in order to explain variation in SJT scores and determine moderation between SJT scores and constructs of interest. Preceding work has already demonstrated that different SJT scores correlate differentially with constructs of interest (e.g. performance in different settings, teamwork, traits, experience, and declarative knowledge [McDaniel et al., 2007; Motowidlo & Beier, 2010; Smith & McDaniel, 1998]) depending on the SJT development process. There are several variables that can be manipulated that will affect these correlations, even after the development of a final SJT product. These variables include keying method, scoring method, and response instructions. Keying method refers to the method for assigning points (i.e. developing a scoring key) to the item responses of an SJT and is usually completed by SMEs familiar with the position for which the SJT is being developed. Scoring method refers to the questions asked of the participant (e.g. rate the effectiveness of each response or choose the best responses) and how the participant responses are assigned a score (e.g. correct/incorrect or distance from keyed level). Keying method has been determined to be a moderating variable when considering the correlates of SJT scores. Some research has demonstrated that different keying methods can moderate these correlations with constructs of interest such as personality (e.g. Motowidlo & Beier, 2010; Bergman, Drasgow, & Donovan, 2006). Specifically, keys developed using SMEs versus undergraduate students will cause differential relationships with declarative knowledge (Motowidlo & Beier, 2010). Other researchers have manipulated response instructions. There are two types or

response instructions. First, there are knowledge-based instructions, or instructions that ask participants, "What *should* you do in this particular situation?" Second, there are behavior-based instructions, or instructions that ask participants, "What *would* you do in this particular situation?" These response instructions have been demonstrated to affect SJT internal consistency, validity, and relationship with variables of interest (e.g. the Big Five and GMA) (McDaniel et al., 2007). However, to date few researchers have manipulated scoring method in order to determine the effects on the correlations SJTs have with constructs of interest. There are many potential findings of value to the field in the exploration of the moderating effects of scoring method, as there is a large potential for impact on construct validity. As scoring method is relatively easy to manipulate (even after an SJT has been developed) this is a fruitful ground for future research.

Increasingly, researchers are discussing the best methods for scoring SJTs in order to capture particular variables such as personality (e.g. Bergman, Drasgow & Donovan, 2006; McDaniel & Nguyen, 2005). SJT scoring method refers to the manner in which responses are elicited from participants (e.g. choose the best response; choose the best and worst response; rate the effectiveness of each) and the method through which overall scores are calculated from these responses (e.g. distance scores from key, dichotomously scored as correct/incorrect). Using different scoring methods, it may be possible to increase the construct validity of SJTs, especially with regard to particular constructs of interest such as personality.

In the past, SJTs have often employed a simple "best choice" scoring method, which is then dichotomously scored as correct or incorrect (e.g. Schubert, Ortwein, Dumitsch, Schwantes, Wilhelm, & Kiessling, 2008). Using this type of scoring technique, we can expect that SJT scores may not have a strong relationship to personality as the implicit mechanisms are not being

captured by the scoring technique. When employing this scoring technique, it may be expected that SJT scores will have a stronger relationship with general mental ability (GMA) and thus a stronger relationship to maximum performance. This is due to the fakability of a simple best choice SJT where the scoring method is very transparent. Transparency refers to the ability to accurately perceive rating dimensions (Kleinmann, 1993). In other words, it is evident to the individual that they will be rewarded for choosing the best answer, and they have the GMA to determine how the answers are varying and by extension determine the correct response. Previous research has demonstrated that personality measures can be faked and that the ability to fake is related to GMA. Pauls and Crost (2005) found GMA had a relationship to the ability to fake which was in line with the expectations of others (r = .31). Further, previous research has also demonstrated that SJTs can be faked. Peeters and Lievens (2005) found that in a 'fake good' condition in which participants were instructed to give the best answer possible as opposed to the honest answer, participants greatly increased their scores relative to an 'honest' condition (d = .89).

Research has also demonstrated that those who are higher in GMA will be more accurate at rating personality and performance. GMA is related to the accuracy with which an individual can rate levels of personality traits and performance effectiveness (e.g. Harris, Vernon, & Jang, 1999; Lippa & Dietz, 2000). For example, Lippa and Dietz (2000) found that individuals who are higher in GMA are more accurate when assessing Extraversion, Neuroticism, and Masculinity/Femininity. Harris, Vernon, and Jang (1999) found that intelligence was related to the accuracy with which one twin could answer a personality inventory rating their sibling twin. Research has demonstrated a strong link between intelligence and rating accuracy. Thus, one

might expect that GMA would be strongly associated with SJT scores when a best choice method is used, as is often the case in SJT research.

However, in a pivotal article, Motowidlo, Hooper and Jackson (2006a) proposed a new method of scoring SJTs that involves comparing a participants' evaluative judgment with a subject matter expert's (SME's) rating of the level of a predetermined trait represented by that particular item response. This method of scoring is based on an individual's implicit beliefs about the effectiveness of different levels of trait expression, which coined the term "Implicit Trait Policies" or ITPs. Implicit Trait Policies (ITPs) can be defined as "implicit beliefs about the effectiveness of different levels of trait expression" (Motowidlo, Hooper & Jackson, 2006a, p. 749). The theory behind ITP scoring is that the more effective an individual believes a trait is as expressed within the item response of an SJT, the likelihood increases that the individual is high on that particular trait (Motowidlo, Hooper & Jackson, 2006a). To illustrate this premise, individuals high on Conscientiousness would be more likely to rate an item response in an SJT as effective if that response demonstrates Conscientiousness. Individuals are disposed to believe that the actions they take are effective – thus, an individual's trait level will cause him or her to view the behavioral expression of that trait as effective. Researchers proposed that these Implicit Trait Policies mediate the relationship between personality traits and procedural knowledge (Motowidlo, Hooper & Jackson, 2006). Calculating ITPs involves determining the relationship between a participant's ratings of the effectiveness of response items with a subject matter expert's (SME's) ratings of the level of a trait that the items exhibit. This can be done by calculating the distance between the effectiveness ratings assigned to the highest and lowest response in terms of trait expression, or by computing correlations between effectiveness and SME ratings of trait expression.

Several theoretical underpinnings help explain the relationship between effectiveness ratings and individual trait levels. First, Contrast Effects (Hovland & Sherif, 1952) cause changes in ITP scores whereby individuals high in a particular trait will rate all item responses that display a low amount of that particular trait as extremely ineffective. Specifically, Van der Pligt and Eiser (1984) found that negative information, or information judged to be opposite of the individuals own viewpoint or attitude, was judged even more negatively than warranted or even further from the participant's own attitude in participant's evaluation of target individuals.

Second, Assimilation Effects (Van der Pligt & Eiser, 1984) cause changes in ITP scores whereby individuals high in a particular trait will rate all item responses that display a high amount of that particular trait as extremely effective. Hovland and Sherif (1952) found in their research that in an SJT task, responses which reflect viewpoints or attitudes somewhat similar to the participant's actual viewpoint will be rated even more favorably than is warranted, and will be judged to be more similar to the participant's own perspective.

Finally, Accentuation Effects (Tajfel, 1957) affect scores whereby value judgments will cause an amplification of the distance between ratings of responses with high and low displays of a trait. For example, in a study were participants are rating the differences in weights of different objects, the rating of valuable objects (i.e., golden coins) will cause individuals to overestimate the differences in weight between the items relative to non-valuable objects, such as lead weights (Tajfel, 1957). This is relevant to the ITP scoring method in that individuals who value the particular trait being measured or being portrayed in the item responses of an SJT may be more likely to exaggerate the differences between an item response high in the particular dimension versus an item response low in that particular dimension (Motowidlo, Hooper, & Jackson, 2006).

Theoretically, through these mechanisms, it should be possible to determine an individual's level of a particular implicit trait without the need for more transparent self-report by measuring the influences of these theories within the effectiveness ratings of different response options that vary as the inherent levels of particular traits that are expressed. Specifically, using a distance scoring technique that compares the effectiveness rating of the highest scored item to the effectiveness rating of the lowest scored item, we can capture an ITP score that will be more highly reflective of personality. Specifically, comparing the highest and lowest rated scores will measure the accentuation or exaggeration of scores that should occur when an individual is high in that particular trait. For example, someone who is high in Agreeableness will exaggerate his or her ratings of item responses low or high in Agreeableness. However, while it is expected that there would be a high distance between these scores, Assimilation Effects will cause individuals high in a particular trait to have difficulty distinguishing between item responses high in that particular trait, and thus may not rate the correct response as the best response. Best response scoring is the scoring technique that is most often used in current SJT research. This difficulty associated with distinguishing items that are both high in a particular trait is theoretically based on Hovland and Sherif's (1961) "latitude of acceptance," or the concept that an individual who values a particular trait will assimilate responses similar to the response that he or she would give such that they are less able to distinguish between those responses high on the trait of interest. Supplemental research building on this theory has found some support for the concept that individuals who have an extreme position would also have wider latitudes of acceptance (e.g. Mascaro, 1969) and may thus be more likely to cluster their effectiveness ratings item responses that represent traits that they possess.

Due to the theoretical mechanisms of contrast, assimilation, and accentuation, it can be expected that the different the ITP scoring method will better capture personality relative to typically used best choice methods. The differences in complexity between these two scoring techniques and the increased accuracy of individuals high in GMA will also cause ITP scores to less reflect GMA, which will be better captured by classically used best choice methods. It can also be expected that the different scoring techniques will have different relationships with typical and maximum performance based on the degree to which they are related to GMA or personality.

Typical performance refers to day-to-day performance over an extended period, while maximum performance refers to the optimal level of performance in a short period when the performer is doing their best (Sackett, Zedeck, & Fogli, 1988). Research has demonstrated that personality is more strongly related to typical performance and GMA is more strongly related to maximum performance (Marcus, Goffin, Johnston, & Rothstien, 2007). Thus, it logically follows that the scoring method which most reflects either personality or GMA will also more strongly predict either maximum or typical performance.

The present research demonstrates the potential for scoring method to moderate the relationship between SJT scores and constructs of interest. Specifically, these constructs of interest are: GMA, personality, maximum performance, and typical performance. In other words, different scoring methods will affect the strength of correlations with specific antecedent variables of interest and also have an effect on the criterion validity through the prediction of maximum or typical performance. We would normally expect SJT scores to be capturing GMA and work experience, but by using the ITP scoring method it becomes possible to measure personality. Because of these differential relationships, the distinct scoring methods are expected

to provide differential incremental validity beyond the constructs of interest. Ultimately, evidence will be provided to support the hypotheses that different scoring techniques will be more strongly related to typical or maximum performance, as personality will be better captured by ITP scores, where GMA will be better captured by best choice scores. Evidence will be provided to demonstrate the differential validity resulting from distinct SJT scoring methods.

The second chapter of this paper includes a literature review of the typical validity of SJTs, in addition to discussions of variables that have been shown to affect this validity. In chapter three, this paper discusses the proposed relationship between SJT scores and correlates, as moderated by scoring method. In chapter four, the development of this particular customer service SJT is discussed as well as results from pilot administrations. Chapters five will focus on the methodology of this study and chapter six will focus on the results demonstrating differential effects for SJT scoring method. Finally, chapter seven will discuss the theoretical and practical implications of these results.

CHAPTER TWO: LITERATURE REVIEW

Statement of the Problem

Selecting the most qualified individual to fill a position has been a constant challenge for organizations. However, in the current economic climate, the task of making the correct selection decisions from a large labor pool becomes even more daunting. The current economic climate is a buyer's labor market, meaning that there is a large ratio of applicants to open positions. Research has demonstrated that when there is a buyer's labor market, scores on applicant's personality tests can be inflated on average by up to .52 standard deviations relative to scores in times of normal labor market (Ones & Viswesvaran, 2007). Fortunately, a well-developed Situational Judgment Test (SJT) can aid organizations by reducing faking and serve as a complement to other selection devices. New scoring methods can also help SJTs serve as implicit measures of useful constructs such as personality. SJTs are becoming increasingly popular due in part to high perceived face validity, incremental validity over other selection devices, positive responses by job applicants, and strong correlations with criteria.

Operational Definition of SJTs

The history of SJTs is relatively long compared to some other methods of selection that are currently used (e.g. simulation, conditional reasoning). The origin of SJTs can be traced back to a scale in the George Washington University Social Intelligence Test published in 1926. Widespread use began during World War II, where the SJT measurement technique was used for civil service and military examinations (McDaniel & Whetzel, 2005b). In the following decades, SJTs were used in a variety of fields. Examples include: Practical Judgment Tests (Cardall, 1942), draft tests from Richardson Bellows and Henry in 1948, and in the 1960s SJTs were used

at the Civil Service Commission (McDaniel & Whetzel, 2005b). More recently, Motowidlo (1990) reinvigorated interest in the SJT, and described them as "low-fidelity simulations." In the last 20 years, research on SJTs has grown dramatically; however, several areas of research still require further exploration.

The literature has demonstrated some disagreement over the breadth of the definition of an SJT. For example, McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001) define an SJT as any paper and pencil test that measures judgment in work settings. Others provide more detailed definitions. Muros (2008) defined SJTs as "a simulation based method of assessment which presents domain-specific situations that require a response" (p. 9). Some authors have defined SJTs in a manner that does not explicitly state that SJT item responses must be presented in a multiple choice format, and thus blend the line between SJTs and Situational Interviews (e.g. Labrador, 2007). However, for the purposes of this paper, SJTs are defined as measurement techniques which share the following characteristics: they present the applicant with job-related situations; they present responses in a multiple choice format; and they have a scoring key that is developed a-priori (McDaniel & Nguyen, 2001; Weekley & Ployhart, 2006).

Steps of SJT development

Although there are variations in the development of an SJT, the basic process involves three steps. First, a group of subject matter experts (SMEs) or job incumbents are consulted to collect critical incidents of the criteria of interest. These critical incidents are used to write brief descriptions of task situations relevant to the specific criteria of interest. Second, a separate group of SMEs or job incumbents are asked to write a few sentences describing how they would handle each task situation presented. These incumbents are instructed to only write the situation

responses believed to be the best or most effective response. These responses are compiled into between five and seven different response strategies for every task situation. Third, a group of job experts or senior incumbents in the job area are asked to evaluate the effectiveness of the alternate strategies for each task situation developed by the previous group of incumbents. Specifically, depending on the scoring key, these managers are asked to select the best response, select the best and worst response, or rate/rank the effectiveness of each response on a scale from one to five. From this information the SJT is developed, although further validation is necessary before the SJT can be used in the field (Motowidlo, Dunnette, & Carter, 1990). The scoring method which will be used to measure applicant success is important to consider during this process. Should a scoring method capture personality, then this should be considered when eliciting critical responses from SMEs. Questions should then be framed in order to ask SMEs about times when personality traits became a factor in the workplace. Additionally, the development of the keying method should be based on the scoring method that will be employed. SMEs should be asked to rank, rate, or 'best choice/worst choice' item response with consideration of the scoring method that will be used in the final product. Understanding the effect of scoring method on SJT correlates has a direct impact on the development process.

Keying Methods

As previously stated, SJTs have scoring keys that are developed a-priori. When considering keying SJTs, many researchers seek to find a formulaic definition of situational judgment with the exact correct mix of variables involved, from GMA to practical intelligence to personality traits, and believe that these traits may be further correlated with SJT scores depending on the keying method used. Other researchers disagree; McDaniel and Nguyen (2001) believe that SJTs are more of a measurement method that can be built to measure a variety of

constructs depending on the type of questions used and the domain sampled. Despite the controversy, it seems that keying method would logically affect the inter-correlations for SJTs, as was recently supported by Motowidlo and Beier (2010).

When considering keying methods for SJTs in general, there are several different approaches that can be taken. One of the most popular manner of doing so is by having subject matter experts (SMEs) evaluate each of the responses for their effectiveness (Motowidlo et al., 1990). This might be considered a *rational approach* to keying. Another approach consists of using a given theory to identify the correct answers (Bergman et al., 2006; Weekley, Ployhart, & Holtz, 2006). Bergman and colleagues (2006) described this *theoretical approach* as follows: "Items and options can be constructed to reflect theory, or theory can be used to identify the best and worst options in a completed test" (p. 225). Alternatively, some authors have employed an *empirical approach* to identify correct responses. Dalessio (1994) employed the "horizontal percent method," where he calculated the percentage of insurance agents surviving and terminating after their first year for each response. The response with the highest number of agents surviving was considered the best, while the response with the lowest number of surviving agents was considered the worst. Four other studies have determined the best and worst responses using the mean criterion performance for each response (Ployhart & Ehrhart, 2003; Weekley & Jones, 1997, 1999; Weekley, Ployhart, & Harold, 2004). A fourth approach for identifying the correct responses has been suggested in the literature as well, although this approach is used less frequently. Legree and colleagues (2005) suggest the use of a *consensus-based approach,* which determines the best and worst answer choices based on means from the sample of interest rather than a smaller subset of SMEs, although this approach is not ideal as it requires participation from the entire population of interest.

The keying method can have a drastic effect on the relationship between SJT scores and relevant correlates. For example Weekley and Jones (1997) compared an empirically derived answer key with a rational key developed using customers as SMEs. They found that SJT scores resulting from the two keys were correlated, but that this correlation was only moderate (r = .48). Additionally Legree and colleagues (2005) reviewed a multitude of studies comparing rational keys and consensus-based keys. Their review demonstrated some extent of convergence of these keys across studies. They reported a correlation between the rational and consensus-based keys in the .70s to .90s as well as correlations between the SJT scores based on those keys in the .80s and .90s. Given the strength of this convergence, these results support Legree and colleagues' assertion that consensus-based keys may be a reasonable substitute for rational keys when the costs of employing the latter are too great. Bergman (2006) compared different scoring and keying methods. This research demonstrated that SJT scoring keys provided a wide range of validity coefficients (-.03 to .32). Interestingly, in this particular study, the results generally had stronger relationships with specific criteria (in this case leadership) than to the measures of overall job performance.

Of particular importance for this research, Motwidlo and Beier (2010) compared scoring keys where the SMEs used to develop the keys were drawn from different populations. In one condition, the key was developed using undergraduate students with little job knowledge. In the other condition, the key was developed by graduate students with more extensive work experience. The scores obtained through these different keys had different scores with the participants work performance (r = .37 for the expert key, r = .29 for the novice key). Additionally, the authors found that after partialling out the variance from these keys with regard to Conscientiousness and Agreeableness, the keys had drastically different correlations with

supervisor-rated work performance (r = .25 for the expert key, r = .05 for the novice key). As illustrated here, the keying method can have a direct effect on the criterion validity of an SJT. Keying method is directly related to scoring method, and the current research. An SJT key should be developed with consideration of the scoring method. This research will attempt to extend these findings to scoring method, and attempt to demonstrate that the criterion validity will be affected differently depending on the criteria of interest; typical performance or maximum performance.

Scoring Methods

Another important choice faced by SJT developers is the method of awarding points to participants and what data should be gathered from participants to assign these points (e.g. McHenry & Schmitt, 1994). In contrast to keying methods, scoring methods concerns the assignment of points from the answer key to participants, not the development of the key itself. Published research directly addressing the scoring methodology of SJTs is quite scarce (Weekley, Ployhart, & Holtz, 2006).

One of the most prevalent methods of scoring SJTs is variations on an approach first described by Motowidlo and colleagues (1990) (Muros, 2008). Respondents are asked to choose a 'best/most likely' and a 'worst/least likely' response from the options presented. If their responses are individually aligned with the keyed best/worst responses, they are awarded one point for each. If either of their responses are not so aligned (i.e., they choose responses keyed as neither best nor worst), they receive zero points for each. Finally, if either of their responses is directly contrary to the keyed response (e.g., they identify the best keyed response as worst/least likely), they lose a point for each. This approach allows a range of-2 to +2 points to be awarded for each SJT item. A frequent variation of this approach is to simply use the 'best/most likely

response' and score the item as correct (1) or incorrect (0). Alternately, sometimes the expert rating for the item the participants selected as the worst or least effective is subtracted from the expert rating for the item the participants selected as the best or most effective (e.g. Motowidlo and Beier, 2010).

Alternatively, some SJTs have employed Likert-type ratings of each response option instead of forced-choice ratings (Weekley, Ployhart, & Holtz, 2006). In such approaches, examinees are asked to rate each response option on a continuous scale, typically for effectiveness or the likelihood they would enact the response. These types of ratings can be aggregated in a variety of ways. Sternberg and colleagues (Sternberg and the Rainbow Project Collaborators, 2002) have used these ratings as part of a "distance-measure" approach, where scores are determined based on the sum of the squared deviations of respondents' scores from the keyed mean values. In this approach, lower scores are considered better because they indicate less distance from the "truth." An advantage of this approach is that it uses much more information (ratings for each situation), and thus may provide more variance and greater reliability across respondents. However, one disadvantage may be in its susceptibility to response distortion. Cullen, Sackett, and Lievens (2006) found that respondents could employ a simple strategy of avoiding the extreme ends of a scale to decrease their distance from the keyed mean values, thus artificially improving their SJT scores. Another approach might be to score the item by simply awarding the respondent the rating value they assigned to the keyed best response, which would mitigate the response distortion issue discovered by Cullen, Sackett, and Lievens (2006). Thus the key would be dichotomous, with only the 'best/not best' response options, and the only value assigned to the participant would be the rating they provided to the best response.

Only one published study has compared scoring approaches (Muros, 2008). Ployhart and Ehrhart (2003) compared three different scoring approaches: one using a forced choice rating of the most effective response only (most only), one using a forced choice rating of the most and least effective response (most/least), and another using a 1-5 rating of effectiveness for each response option (effectiveness ratings; only ratings of the keyed correct and incorrect responses were used - they were summed together after the rating for the keyed incorrect response was reverse-scored). However, it is important to note that this research was designed to compare the effects of response instructions, not necessarily scoring methods. Their results indicated moderate convergence of the SJT scores resulting from these three scoring approaches. Specifically, the best choice scores correlated with the 'best/worst choice' scores at r = .38, whereas the best choice scores correlated with the effectiveness ratings scores moderately at r = .32, and the 'best/worst choice' scores correlated moderately with the effectiveness ratings scores at r = .37. These inter-correlations were moderated by the type of instructions used (i.e. "should do" versus "would do"). These differences were found even in a within subjects design. Reliability was higher for the effectiveness ratings (α = .67) compared to the 'best/worst choice' (α = .36) and best choice ratings (α = .52). Validities for predicting peer-ratings of performance varied for the different scoring approaches as well. Corrected for attenuation due to unreliability in both measures, the best choice and 'best/worst choice' approaches predicted better than the effectiveness ratings for a within-subjects sample where the subjects completed the SJT using all three scoring approaches (r = .29 and .27 versus .18, respectively). Comparatively, the effectiveness ratings and most only approach predicted performance better than the most/least approach in a between-subjects sample where the subjects each completed the SJT using only one of the three scoring approaches (r = .37 and .33 vs. -.01 respectively; not corrected for

attenuation). In sum, Ployhart and Ehrhart's (2003) study demonstrates that scoring approaches can substantively impact the reliability and validity of an SJT, even to the point of having drastic effects on criterion of interest.

To our knowledge, no previous study has directly compared best choice scoring techniques to distance scoring techniques designed to capture ITPs in order to demonstrate the differences and the potential benefits this new scoring method. In the current study, the best choice scoring method employed is similar to the method employed in a previous study comparing scoring techniques (Ployhart & Ehrhart, 2003). Specifically:

- The distance score (the distance between the item responses rated highest and lowest on effectiveness)
- Best choice score (dichotomously scored correct or incorrect best choice response)

Face Validity

An often-touted benefit of SJTs is the routinely positive reactions applicants tend to have toward SJTs (Chan & Schmitt, 1997; Richman-Hirsch et al., 2000). People respond positively to SJTs because it is explicit that SJT content is related to the target jobs for which they are applying (Kluger & Rothstein, 1993; Ployhart & Ryan, 1998). Applicant reactions and face-validity are important to consider, as perceptions of job-relatedness are likely to prevent challenges to selection systems as well as prevent rulings against them (Smither et al., 1993).

Even beyond typical paper-and-pencil SJTs, Richman-Hirsch et al. (2000) established that a multimedia SJT was judged by applicants as significantly more face valid and more enjoyable than the same SJT in a written format. Recently, Kanning et al. (2006) examined job applicant perceptions of SJT items that varied along interactivity, stimulus fidelity, and response

fidelity. The results demonstrated that interactive or branching SJT items which used videos in the stimulus and response component received the highest applicant ratings.

Related to the above findings Chan and Schmitt (1997) compared face validity perceptions of a written and video-based SJT. While both were judged to be job-related by participants, the video-based SJTs were rated significantly higher on this dimension. Additionally, their data suggested Black participants may experience reduced test-taking motivation when confronted with lower face-valid perceptions. The authors suggested that this may play a role in reducing test performance for Blacks and, as a result, inadvertently cause increased adverse impact. Other researchers have found support for the concept that higher-fidelity SJTs will result in more positive reactions from applicants (e.g. Olson-Buchanan & Drasgow, 2006). Additionally, researchers have suggested that such detailed and immersive simulations may serve as realistic job previews for job candidates (Olson-Buchanan & Drasgow, 2006; Dalessio, 1994).

Other researchers had attempted to determine the rationale behind applicant's positive reactions to SJTs. For example Bauer and Truxillo (2006) applied Gilliland's (1993) procedural justice rules to applicant perceptions of SJTs. The authors believe that SJTs are more positively evaluated by applicants due to perceptions of relevance to the job and consistency of administration and scoring relative to alternate selection measures (e.g. unstructured interviews). The authors also believe that SJTs are more positively evaluated by applicants due to the applicant's appreciation for an opportunity to demonstrate job-relevant skills, and opportunity to receive immediate feedback. There would be great benefit to using SJTs to measure personality through ITP scores, as applicants would perceive the measure was relevant to the job and have a positive reaction to the measure. As the measure is tapping personality implicitly, the measure

would have a high level of face value while measuring personality and maintaining item subtlety. Item subtlety would be conceptualized as the lack of an obvious substantive link between test item content and its underlying construct (Holden & Jackson, 1979).

Criterion Validity

One of the most touted advantages to SJTs is their ability to predict relevant work outcomes. Large-scale studies have shown that SJTs have significant criterion-related validity (e.g. McDaniel & Nguyen, 2001; McDaniel et al., 2007). Further, SJTs possess incremental validity over and above GMA and personality tests in predicting relevant criteria (Chan & Schmitt, 2002; Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001). Across a multitude of studies, SJTs have been shown to predict multiple criteria, including both task and contextual job performance (Chan & Schmitt, 2002; Clevenger et al., 2001; O'Connell et al., 2001,2007; Ployhart & Ehrhart, 2003; Pulakos & Schmitt, 1996; Weekley & Jones, 1997, 1999; Weekley & Ployhart, 2005; Weekley, Ployhart, & Harold, 2004), student performance (e.g., GPA, course-specific grades, absenteeism, and study skills) (Lievens, Buyse, & Sackett, 2005; Lievens & Coetsier, 2002; Lievens & Sackett, 2006; Oswald et al., 2004; Peeters & Lievens, 2005), customer service ability (Jones & DeCotiis, 1986), turnover (Dalessio, 1994), and on-the-job accidents (Hunter, 2003; Legree et al., 2003). However, there are no studies that have been previously published regarding the ability of SJTs to predict typical or maximum performance.

When examining the relationship between SJT scores and criterion in a large scale study, McDaniel and Nguyen (2001) conducted a seminal meta-analysis of the criterion-related validities of SJTs in employment settings. After analyzing 102 individual validity coefficients of 10,640 participants, the corrected correlation between SJTs and job performance was $\rho = .34$. The researchers found that there was substantial variability in criterion-related validity

coefficients across studies, which suggested the presence of moderators. Analyses demonstrated that a key moderator of the validity of SJTs concerned whether a job analysis was used to develop the test. SJTs based on a job analysis demonstrated higher validities than those developed without basis on a job analysis (ρ = 0.38 versus ρ = 0.29). The authors also tested other moderators, such as the level of detail within the question, the g loading of the SJT, and predictive versus concurrent study design. The results from these moderation analyses are inconclusive and difficult to decipher; further exploration and more data is necessary to draw more concrete conclusions.

More recently, McDaniel and colleagues (2007) conducted a meta-analysis reporting the relationship of written SJTs to job performance. Using published and unpublished research, they aggregated 118 validity correlations from 24,756 participants. After correcting for sampling error for each study and measurement error in the criterion, they found a relationship of ρ = .26. This discrepancy in the findings between the two meta-analyses is likely due to the inclusion of additional studies in this more recent analysis, which were largely unpublished and that, on average, have lower criterion related validities than those included in the previous meta-analysis.

Upon examining an SJT designed to assess multidimensional student performance, Oswald and colleagues (2004) found statistically significant correlations with GPA (r = .16), absenteeism (r = -.27), and self- and peer-rated student performance (r = .53 and r = .16, respectively). These results demonstrate the versatility of SJTs in predicting a multitude of varied criterion. Additionally, with regard to predicting supervisor ratings, Jones and DeCotiis (1986) reported results from studies indicating that their customer service-oriented SJT was predictive of customer service ratings by supervisors.

In addition to being used to predict work-related criteria, SJTs are increasingly being used themselves to serve as criteria for training programs, or to assess training needs. For example, Fritzsche, Stagl, Salas, and Burke (2006) discuss how SJTs could be used to assess training needs, assist in teaching course content, or to evaluate the outcomes of training. Their review demonstrated that this use of SJTs is relatively recent, but SJTs demonstrate potential for being used in for such purposes in training.

Previous studies on SJT criterion validity have not examined the differential results in predicting typical and maximum performance. This is an important factor to consider, as SJTs that capture typical performance are likely to predict long-term on-the-job performance, and are likely to have less adverse impact. This research utilizes a theoretical model to illustrate how SJT scoring method can affect the prediction of typical or maximum performance. Understanding this differential prediction and accounting for the prediction of either maximum or typical performance can improve our understanding of how SJTs predict different criteria. This understanding will aid in an improved prediction of criteria by SJT scores, and a more accurate assessment of criterion validity.

Adverse Impact

The term ‘adverse impact’ refers to subgroup differences in the outcome of an employment decision (Collins & Morris, 2008). Adverse impact is important to consider in the development of any measure, as Title VII of the Civil Rights Act of 1964 clearly states that it is against the law for companies to base any hiring, retention, or promotion decisions based on race, sex, or national origin. Title VII is important to applicant screening because employers must ensure that any tests used are not biased against minorities or any other protected class of people. The enactment of this law is of particular relevance to Industrial/Organizational

Psychologists. This legislation forced organizations to take a closer look at the ways people were selected for jobs and particular attention was given to evaluating fairness in employment tests.

One reason SJTs are often used as an additional selection procedure would be to mitigate the effects of other measures that may demonstrate a greater amount of adverse impact. Research has demonstrated that SJTs have less adverse impact on racial minorities than traditional GMA tests (Clevenger et al., 2001; Harold & Ployhart, 2001; Jenson, 1998; McDaniel & Nguyen, 2001; Motowidlo et al., 1990; Oswald et al., 2004; Pulakos & Schmitt, 1996; Weekley & Jones, 1999), although some researchers have suggested that this may be partially the result of the lower reliability of SJTs (Chan & Schmitt, 1997). A recent meta-analysis by Nguyen et al. (2005) demonstrates that while SJTs reduce some subgroup differences, not all subgroup differences are mitigated. Specifically, Asians often score lower than Whites on SJTs, highlighting the issue that while SJTs tend to generally help reduce adverse impact, it depends on the subgroup being examined. However, these differences are often less pronounced than those found with GMA tests. A recent meta-analysis found a difference in mean SJT scores between Whites and Blacks of 0.38 standard deviations in favor of White participants (Nguyen et al., 2005). A key factor in determining the level of adverse impact of SJTs is the correlation of SJTs with GMA. Of particular importance to this research, the level of adverse impact of a test is considerably reduced if the SJT captures primarily non-cognitive aspects of job performance (Lievens, Peeters, & Schollaert, 2008). Thus the reduction of adverse impact seems dependent on the cognitive loading of the particular SJT in question, such that SJTs with a higher cognitive loading will have more adverse impact (Nguyen et al., 2005). An additional factor to consider is fidelity; video-based SJTs seem to result in less adverse impact than written SJTs because video-based SJTs are less cognitively loaded (Chan & Schmitt, 1997). Finally, SJTs with behavioral

tendency instructions (measures of typical performance) showed lower adverse impact than SJTs with knowledge instructions (Nguyen et al., 2005). One can infer that this reduction in adverse impact is due to the increased GMA necessary to answer questions that pertain to knowledge instructions. Cognitive factors play a large role in the level of adverse impact of an SJT.

When considering gender-related adverse impact, Male-Female group differences tend to favor females when differences are found (McDaniel & Nguyen, 2001; Motowidlo et al., 1990; Motowidlo & Tippins, 1993; Lievens & Coetsier, 2002; Oswald et al., 2004; Weekley & Jones, 1999). In their meta-analysis, Nguyen et al. (2005) found a difference in mean scores between females and males of 0.10 standard deviations, with females performing better than males. Some have suggested that this gender bias might be due to gender differences in terms of the personality traits triggered by the SJT situations. Often in several customer-service or team-based SJTs, these scenarios are interpersonal. In general, females tend to score higher on traits such as Agreeableness or Sociability (specifically warmth and openness to feelings) (Costa et al., 2001), which would explain these gender differences. These results demonstrate that SJTs may help organizations compensate for other methods that may inherently discriminate towards females.

The previous research on the adverse impact of SJTs suggests that SJTs that capture aspects of personality may help reduce both racial and gender adverse impact. A scoring technique that maximizes the ability of an SJT to capture personality and minimizes the correlation with GMA will hypothetically further reduce adverse impact. This research will explore how to reduce the cognitive loading of SJT scoring through the use of different scoring methods.

Reliability

Reliability, or the consistency of a measurement instrument (Meyer, 2010), is often a difficulty for SJTs, as the domain-sampling and multidimensional nature of SJT development has led multiple authors to report findings of low internal consistency (Chan & Schmitt, 2005). However, alternative forms of reliability, such as parallel forms reliability, alternate-forms reliability and test-retest reliability, have yielded higher, more conventionally acceptable reliabilities (Clause et al., 1998, Lievens & Sackett, 2007; Potosky and Bobko, 2004). Clause et al. (1998) showed that acceptable levels of parallel forms reliability can be achieved for SJTs using systematic processes for matching specific items. The authors describe a procedure of "item-cloning," in which items were repeatedly reviewed by SME panels to ensure they were parallel to the original SJT items. This process resulted in statistically equivalent alternate forms of the SJT with similar means, variances, and factor structures. These alternate forms exhibited substantially higher (.70) parallel forms reliabilities than the internal consistency reliabilities typically observed for SJTs (.30 - .60). Related to this Potosky and Bobko (2004) reported a relatively high correlation of $r = .84$ between a paper-and-pencil SJT and a web-based SJT administered in a within-subjects design, which can serve as an indicator of their SJT's test-retest reliability.

McDaniel and Nguyen (2001) conducted a meta-analysis that aggregated the results of several studies and found that the internal consistency coefficients ranged from 0.43 to 0.94. Research has recognized several moderating variables that affect the variability in internal consistency reliability. First, the length of the SJT was a moderating variable, with longer SJTs demonstrating higher internal consistency. Second, Ployhart and Ehrhart (2003) found that response instructions influenced the internal consistency. Response instructions asking

participants “to rate the effectiveness of each response” resulted in the highest internal consistency (0.73). Alternately, response instructions that asked candidates to choose two response alternatives (e.g. “pick the best and worst response”) resulted in slightly lower internal consistency (0.60). Finally, response instructions that asked participants to select only one response (e.g. “select the best response”) had the lowest internal consistency (0.24).

Other researchers have argued that internal consistency is only valid for uni-dimensional tests. If this is the case, research has demonstrated adequate test-retest reliability, as Ployhart et al. (2004) reported a test-retest reliability of 0.84. Additionally Bruce and Learner (1958) and Richardson, Bellows, Henry, and Co. (1981) found test-retest reliabilities that ranged from 0.77 to 0.89 for two SJTs, the “Supervisory Practices Test” and for the “Supervisory Profile Record.” Thus these results demonstrate that the test-retest reliability of SJTs is satisfactory.

Construct Validity

Construct validity refers to the degree to which a scale measures or correlates with the theorized psychological construct that it purports to measure (Pennington, 2003). Construct validity is a difficult issue to address with regard to SJTs, as often the development process causes differential relationships between SJTs and constructs of interest. The development process often includes critical incidents which may be unique to the position in question. As scoring methods hypothetically have an effect on the correlations of SJT scores with constructs of interest, then not controlling for these effects will only further confuse these relationships. Research has demonstrated some difficulty with finding consistent relationships between constructs of interest and SJT scores. Schmitt and Chan (2006) review the evidence of construct validity, taking into account SJTs' typical subgroup differences, factor structures, internal consistency, stability over time, and inter-correlations with other constructs. These other

constructs include GMA, personality, job knowledge, and occupational interests. The authors conclude that SJTs as methods of measurement can be designed to measure a variety of constructs. They draw analogies to assessment centers and interviews, both of which can be developed to assess a broad array of constructs but also tend to correlate consistently with certain constructs. Specific to SJTs, written formats of tests and SJTs employing knowledge-based instructions tend to correlate with GMA (Chan & Schmitt, 1997; McDaniel et al., 2007, McDaniel & Nguyen 2001). Those focusing on more interpersonally oriented constructs (as many SJTs do) as well as those employing behavioral tendency instructions tend to correlate with elements of personality (McDaniel et al., 2007; Motowidlo & Beier, 2010).

Considering some of these constructs with which SJT tend to have consistent relationships, McDaniel and colleagues (McDaniel & Nguyen, 2001; McDaniel et al., 2007) have aggregated the data across a variety of studies and domains to determine estimated mean population correlations between SJTs and pertinent constructs. It is important to mention that these results are subject to a large amount of variability as often happens when aggregating data from SJTs which tend to vary largely from domain to domain (McDaniel et al., 2006). Despite this, the authors reported the following meta-analytic correlations between written SJTs and: Agreeableness, $\rho = .25$, Conscientiousness, $\rho = .27$, Emotional Stability, $\rho = .22$, Extraversion, $\rho = .14$, Openness to Experience, $\rho = .13$, GMA, $\rho = .32$, and job experience, $\rho = .05$ (McDaniel et al., 2007).

Again, these results must be interpreted carefully due to plethora of domains in which SJTs are developed. Factor analytic SJT research has often found that SJTs are composed of an abundance of complicated and difficult to comprehend factors (Schmitt & Chan, 2006). This is to be expected as SJTs are measurement methods which assess a variety of work-related

knowledge, skills, and abilities (KSAs) (McDaniel & Nguyen, 2001; McDaniel & Whetzel, 2005; Weekley & Jones, 1999). For instance, SJTs were recently developed to capture domains as diverse as aviation pilot judgment (Hunter, 2003), teamwork knowledge (McClough & Rogelberg, 2003; Morgeson et al., 2005; Stevens & Campion, 1999), employee integrity (Becker, 2005), call-center performance (Konradt et al., 2003), academic performance (Oswald et al., 2004), and implicit aggression (Miller et al., 2010).

Additional research into the construct validity of SJTs is still warranted, as the unique constructs underlying SJT scores tend to have inconsistent relationships. Additionally, the mechanisms through which SJT scores consistently correlate with particular variables and achieve construct validity remain open to some debate in the literature (McDaniel et al., 2006; Ployhart & Weekley, 2006; Schmitt & Chan, 2006). The presence of moderating factors is likely. It is important for researchers to understand and control for these factors to aid in understanding the relationships SJT scores have to other predictors of work performance and aid in the prediction of typical performance. For the purpose of this research, it is proposed that scoring method can be demonstrated to cause significant differences in the relationship SJT scores have with personality and GMA. Once the moderating factor of scoring method is understood, accounting for this variance should give us a clearer picture of the factor structure of SJTs. Additionally, it would aid in optimizing the construct validity of an SJT during the development and validation process. Hypothetically, a scoring method could be used based on the ability of that particular method to capture either personality or GMA.

SJT Score Correlates

The multidimensional nature of SJTs has frequently resulted in low internal consistency reliabilities (Chan and Schmitt, 2005). This has complicated efforts to design SJTs to measure

specific constructs and construct dimensions (McDaniel et al., 2006; Schmitt & Chan, 2006). Coherent factor structures have been elusive and difficult to obtain, leading to the consistent use of overall SJT scores instead of scale scores. For example Oswald and colleagues (2004) used a domain-sampling approach to develop 12 dimensions of student performance. They arrived at these 12 dimensions by sampling a variety of college and university websites for explicit educational objectives or mission statements, which were later categorized into 12 dimensions. The researchers then developed a 57-item SJT to reflect these 12 dimensions, with three to six items per dimension. The coefficient alphas for each scale or dimension were quite low (α = .20s to .40s), the inter-correlations between the scales approached unity and lacked any evidence of discriminant validity. An exploratory factor analysis of the SJT data revealed a large general factor accounting for three times the variance of the second factor. In short, these results did not support the use of the 12 previously developed dimensions. Instead, the authors created a composite score across all the dimensions to reflect "judgment across a variety of situations relevant to college life," which resulted in an internal consistency reliability of $\alpha = .85$. However, it is important to remember that while there are some difficulties in finding dimensions within SJTs (McDaniel, 2006), there are some traits and constructs that tend to consistently correlate with SJT scores.

Despite some difficulty with differential prediction and discriminant validity within SJT scores, certain authors have overcome this shortcoming and have developed SJTs that can capture different dimensions. For example, Motowidlo Hooper, and Jackson (2006) have created an SJT that is able to measure both participant's Agreeableness and Extraversion within the same SJT. In their study, the authors created scenarios to tap these personality traits and used the Implicit Trait Policy scoring method. Their test of undergraduate students found that the ITP

scores from their SJT for Agreeableness were correlated with self-report trait Agreeableness (r = .31). Additionally, the ITP scores for Extraversion were correlated with trait Extraversion (r = .37) (Motowidlo, Hooper & Jackson, 2006). This is directly relevant to the current research, as the correlations that SJT scores have with these variables is a major theme in the theoretical model. Understanding these relationships and the moderating variables of these relationships is going to aid in the optimal prediction of performance. To move beyond the specific examples listed above, let us examine some of the relatively consistent correlates of SJTs.

Personality Traits

Different personality traits will often correlate with the scores obtained through the SJT method. This correlation tends to vary based on the domain-specificity of the SJT, in many cases there are correlations with GMA and the Big Five personality measures to varying degrees (McDaniel et al., 2007; McDaniel & Nguyen, 2001). Weekly and Ployhart (2005) determined through path analysis that the effects of personality on performance were partially mediated by SJT scores.

More specifically, within the Big Five framework, Conscientiousness, Agreeableness, and Emotional Stability have been found to be related to SJTs (Clevenger et al., 2001; McDaniel & Nguyen, 2001; Mullins & Schmitt, 1998; Smith & McDaniel, 1998; Weekley & Jones, 1999). McDaniel & Nguyen (2001) found meta-analytic results that demonstrated SJT scores relationship with Emotional Stability (ρ = .31), Conscientiousness (ρ = .26), and Agreeableness (ρ = .25). However, there were no significant relationships found with Openness to Experience and Extraversion. Interestingly, Emotional Stability, Conscientiousness, and Agreeableness are the same three personality-constructs that research has demonstrated to account for the validity of measures of customer service (Frei, 1997; Hogan, Hogan, & Busch, 1984; Ones &

Viswesvaran, 1996) and integrity (Ones, Viswesvaran, & Schmidt, 1993). SJTs may be valid predictors of performance because they are to some extent capturing these relevant personality constructs. This is reflected in the degree personality is reflected by the judgments applicants make regarding which course of action is most effective (Clevenger et al., 2001).

Another study by Oswald et al. (2005) demonstrated that the primary personality correlates were Agreeableness (r = .38), Conscientiousness (r = .28), and Openness (r = .21). These differential correlations may be caused by the domains specific nature of the SJTs, or the format in which the SJTs were administered. Research has demonstrated that high fidelity video-based SJTs may be better correlated with Openness to Experience relative to paper-and-pencil tests (Lievens & Coester, 2002). There may be several other moderating factors that will alter the correlations between the Big Five personality factors and SJT scores.

McDaniel and colleagues (McDaniel & Nguyen, 2001; McDaniel et al., 2007) have aggregated the data across a variety of studies and domains to determine estimated mean population correlations between SJTs and pertinent constructs. It is important to mention that these results are subject to a large amount of variability, as is typical when aggregating data from SJTs that vary largely across different domains (McDaniel et al., 2006). Despite this, the authors reported the following meta-analytic correlations between written SJTs and: Agreeableness (ρ = .25), Conscientiousness (ρ = .27), Emotional Stability (ρ = .22), Extraversion (ρ = .14) Openness to Experience (ρ = .13), GMA (ρ = .32) and job experience (ρ = .05). Another related study created an SJT for undergraduate college students and administered the test along with a 50-item personality test to assess the Big Five personality dimensions (Oswald et al., 2005). The results reflected that the primary personality correlates were Agreeableness (r = .38), Conscientiousness (r = .28), and Openness (r = .21). O'Connell, Hartman, McDaniel, Grubb, and Lawrence (2007)

evaluated seven different SJTs from various manufacturing companies. The authors found relationships between situational judgment scores and internal locus of control (r = .65), Conscientiousness (r = .33), Agreeableness (r = .31), positive affectivity (r = .26), and attention to detail (r = .33).

Nelson (2009) found that job type may moderate the relationship between personality traits and SJT scores. For example, the maintenance technicians who scored high in an SJT also scored high in social confidence and being outgoing but also scored low in being controlling and achievement-oriented. Conversely, the leasing agents who scored high in situational judgment also scored high in being democratic and affiliative but scored low in being decisive, competitive, and innovative. Thus, there were different personality traits that were predictive of success in each particular SJT. It is likely that these results would replicate when comparing SJTs between job domains.

There are several conclusions that can be drawn from these studies. First, while SJT scores tend to be related to the Big Five in meta-analyses, these relationships are often inconsistent in smaller-scale studies. The presence of relationships between SJT scores and the Big Five demonstrate that it may be plausible for SJTs to capture these personality constructs. This is directly important this research as this is the foundation of the theoretical model. The relationship with personality and SJT scores is necessary to establish before demonstrating the presence of moderating factors. Should scoring method prove to be one of these moderating factors, a particular scoring method could be employed in order to best capture personality traits. Determining the steps necessary to develop an SJT that optimally capture personality would be beneficial from a theoretical as well as a practical standpoint.

General Mental Ability

General Mental Ability (GMA) can be defined as "any measure that combines two, three, or more specific aptitudes, or any measure that includes a variety of items measuring specific abilities (e.g., verbal, numerical, spatial) (Salgado, Anderson, Moscoso, Beruta, de Fruyt & Rolland, 2003). GMA is a trait often considered when determining employment testing. This is due to the relationship between GMA and job performance and also the potential for adverse impact inherent in using highly cognitively loaded tests. GMA is positively correlated to beneficial work outcomes such as performance and role breadth (Morgeson et al., 2005). However, many measures demonstrate adverse impact against protected populations, with black-white differences as large as one standard deviation often found (Hunter & Hunter, 1984). In a recent meta-analysis, McDaniel et al. (2007) found that there is an average correlation of $r = .32$ between GMA and SJT scores. This correlation should be considered with the understanding that there is a large amount of variation, as different domain-specific SJTs are likely to have unique and varied correlations with GMA. This relationship between GMA and SJT scores was moderated by response instructions such that instructions asking participants to respond as they 'should do' in that particular situation, or knowledge-based instructions, resulted in a strong correlation between SJT scores and GMA. Relatively speaking, response instructions which asked the participant what they 'would do' in that situation, or behavior-based instructions resulted in lower correlations ($\rho = .35$ versus $\rho = .19$). Other research has replicated this relationship between GMA and SJT scores. Weekly and Ployhart (2005) have found a correlation between SJT scores and GMA ($r = .36$) and GPA ($r = .21$). In another meta-analysis McDaniel and Nguyen (2001) examined 79 study correlations and found that SJTs show a correlation of $\rho = 0.46$ with GMA. However, again there was substantial variability around this estimate due to the vast differences in SJTs and several moderating variables. Specifically, McDaniel and

Nguyen (2001) found that SJTs based on a job analysis were more highly related to GMA (ρ = .50) than SJTs not developed using a job analysis (ρ = .38). Additionally, there are other variables that cause variation in the relationship between GMA and SJT scores. For example, video-based SJTs tend to have lower correlations with GMA relative to written SJTs (Weekley & Jones, 1997).

Relevant to SJT research, Sternberg (2000) has introduced the theory of practical or tacit intelligence, which is intelligence that individuals use to determine the best fit between themselves and their environment. This intelligence is procedural rather than factual, it is usually learned without the help of others or explicit instruction, and is knowledge about issues that are personally important to the learner. Sternberg (2000) argues that this intelligence is a separate construct from general intelligence and will predict work outcomes more accurately. Some researchers found support for the theory that SJTs capture this construct (e.g. Stemler and Sternber, 2006), while others argued that there was no support for this assertion or even the construct of practical intelligence itself (e.g. McDaniel & Whetzel, 2004). Thus while this theory has been introduced to explain how and why SJTs have criterion validity, the research has yet to demonstrate conclusive support for this justification. The measurement of ITPs may be in part based on the ability of ITP scores to capture this practical intelligence.

Research has found support for the idea that the complexity of an SJT (i.e., its length, complexity, verbal comprehensibility, and use of layered items or responses) may influence its observed validity through its association with reading ability (Chan & Schmitt, 1997; McDaniel & Nguyen, 2001; Motowidlo, Hanson, & Crafts, 1997; Olson-Buchanan & Drasgow, 2006). This demonstrates that there are tangible aspects of SJTs that can be manipulated to alter the relationships between SJTs and pertinent dimensions. If the complexity of the SJT content has a

relationship with GMA, it also seems likely that the complexity of the scoring method will have an effect on the relationship between SJT scores and GMA.

Weekley and Jones (1999) have found differing results with respect to SJTs mediating the relationship between GMA and performance, with some studies demonstrating full mediation, and others demonstrating partial mediation. Although GMA is one variable related to SJT performance, their research has supported the theory that the validity of SJTs is apparently not solely a function of this relationship. As previously stated, many studies that have examined the incremental validity of SJTs have shown SJTs to be incrementally predictive of performance beyond GMA (e.g., Clevenger et al., 2001; McDaniel & Nguyen, 2001; Weekley & Jones, 1997, 1999).

The relationship of GMA to SJT scores is particularly important to the present study, as the item responses are expected to have cognitive loadings. A strong relationship with GMA should predict maximum performance better than typical performance, and may cause more adverse impact. Previous research has demonstrated that SJTs will have more adverse impact if they have higher cognitive loadings (Nguyen et al., 2005). If a scoring method can be employed that reduces the cognitive loadings of SJT scores, there are implications for the further reduction of adverse impact and the prediction of typical versus maximum performance.

Previous Experience

Another construct which is often measured in relation to SJT scores and which has direct relevance to construct validity is previous experience. In fact, one of the proposed mechanisms through which SJTs function is the relationship SJT scores tend to have with previous experience. Previous experience is vital to examine as it has been shown to be a valid predictor of work performance (Motowidlo, 1990). McDaniel, Schmidt, and Hunter (1988) defined work

experience as length of service in a given occupation; their meta-analysis of 947 validity coefficients yielded a mean of .21. Previous work experience has often been used as a predictor in personnel selection in the past (e.g., Olney, 1982; Wingrove, Glendinning, & Herriot, 1984).

Motowidlo et al. (1990) believed that previous work experience was one of the antecedents of SJT scores. This is based on the principle that the best predictor of future behavior is past behavior, or the 'behavioral consistency principle' (Wernimont & Campbell, 1968). Other researchers (e.g. Weekley & Jones, 1997; 1999; Weekley & Ployhart, 2005) have also found support for the relationship between SJTs and previous experience. Specifically Weekly and Ployhart (2005) found a correlation of $r = .13$ between job tenure and SJT scores (in a sample of 271 employees), $r = .21$ with general work experience, and $r = .21$ with training experience.

However, often there is substantial variability found in the relationship between previous work experience and SJT scores. Several researchers found inconsistent relationships between previous experience and SJT scores (Smith & McDaniel, 1998; Weekley & Jones, 1997, 1999; for an exception, see Clevenger et al., 2001). Several explanations exist for these inconsistencies. For example, often previous research utilized uni-dimensional measures of work experience, despite the fact that experience is a multidimensional construct with a multitude of individual difference and contextual influences (Quinones, Ford, & Teachout, 1995; Tesluk & Jacobs, 1998). For example, Weekley and Jones (1999) reported a correlation of $r = .23$ between SJT performance and a general measure of general work experience. The authors also found a correlation of $r = .02$ between organizational tenure and the same SJT measure. Ignoring such differences in experience (e.g., job experience versus organizational tenure) can result in inconsistent or misleading relations with other variables. Relationships between experience and SJTs consistently exhibit inconsistent findings (see Clevenger et al., 2001; Weekley & Jones,

1997, 1999). These inconsistent results only further demonstrate that there are likely other variables at play, such as GMA and personality. While previous experience may affect procedural and declarative knowledge and increase SJT scores, it is apparent that other factors influence the scores. Specifically, the explanation for these inconsistent findings may include the moderating effects of keying methods (as found by Motowidlo and Beier [2010]) or scoring methods, as the current study posits. Previous experience is an important variable to measure as it is a variable included in several models that demonstrate the predictive validity of SJT scores.

Maximum/Typical Performance

Maximum and typical performance criteria are very important to consider when understanding the predictive validity of selection tests. Campbell (1990) constructed a model of performance which posits that performance is a function of declarative knowledge, procedural knowledge, and motivation. Variation on the level of motivation of an individual has direct impact on the continuum from typical to maximum performance. Motivation tends to vary during typical situations, while motivation tends to be consistently high during maximum performance situations (Sackett, Zedeck, & Fogli, 1988). Related to this, Klehe and Anderson (2007) found that motivation, computer self-efficacy, and persistence played an important role in predicting typical performance, while measures of declarative knowledge were a stronger predictor of maximum performance. Motivation was found to contribute to the variation in typical performance. Thus, there would be less variation in maximum performance measures, as the variation would largely be caused by ability, while in typical performance measures the variation could be attributed to both ability and motivation (Sackett, 2007).

This relationship between typical and maximum performance is important to consider for the design of selection instruments. When examining the typical and maximum performance

predictors, it is important to note that different predictors will correlate more strongly to either maximum or typical performance. When considering these constructs in real life settings, DuBois et al. (1993) found GMA to be a better predictor of maximum performance than of the typical performance with regard to the speed with which supermarket cashiers processed goods. Marcus, Goffin, Johnston and Rothstein (2007) collected supervisory ratings as measures of typical performance, assessment center ratings as measures of maximum performance, and personality and GMA measures. A confirmatory factor analysis supported the authors' hypothesis that typical performance would be more strongly associated with personality predictors, while maximum performance would be more strongly associated with GMA. Witt and Spitzmuller (2005) found in a field study that GMA was more strongly related to maximum performance, while perceived organizational support was more strongly related to typical performance. Should scoring methods have an effect on the relationship between SJT scores and predictor variables such as personality and GMA, it is then likely the scores will also differently predict a participant's level of either maximum or typical performance.

Incremental Validity

While there are different predictors of work performance and behavior, SJTs have demonstrated incremental validity, or a level of predictive validity beyond the variance accounted for by many other predictors typically used to explain job performance. For example McDaniel and colleagues (2007) collected meta-analytic data and conducted hierarchical linear regression to determine the incremental validity of SJT scores over GMA, the Big Five, and a composite of each. They found SJTs provided incremental validity over GMA ranging from .03 to .05, over the Big Five ranging from .06 to .07, and over a composite ranging from .01 to .02. Across several studies, Weekley and Jones (1997, 1999) demonstrated that when combining

predictors of work performance such as GMA, previous experience, and a video-based SJT, regression analyses demonstrated that the SJT demonstrated a statistically significant increment in the proportion of variance accounted for by the SJT, or ΔR^2 ranging from .111 to .096 (Weekley & Jones, 1999). Further, Lievens and his colleagues (2005) have explored the incremental validity of a predictor set that included scores on several cognitive and factual tests, and a video-based SJT. They reported consistent, statistically significant increases in the SJT's incremental predictive validity for scores in school courses over four years of data collection (ΔR^2 = .01 (ns), .02, .06, .07 in each year, respectively). Although these increments seem relatively small, the authors observed that few predictors offer incremental validity beyond a composite of GMA and the Big Five. It should be noted that these results were in a condition in which there was an interpersonal content to the curriculum. In an alternate condition where there was no interpersonal aspect to the curriculum, the SJT scores explained no additional variance over cognitive tests (Lievens, Buyse, & Sackett, 2005). Clevenger and colleagues (2001) assessed the incremental validity of an SJT beyond GMA, conscientiousness, job experience, and job knowledge across three independent samples. Two of the three SJTs accounted for a statistically significant amount of additional variance in the criteria, which was supervisor ratings in this case (ΔR^2 = .026, ΔR^2 = .017, ΔR^2 = .016 respectively). The authors believed that the reason for the non-significant finding in the third study was due to the conservative nature of the incremental validity test, as the finding was nearly significant and relative in magnitude to previous findings (e.g. Weekley & Jones, 1997, 1999).

To examine additional findings Chan and Schmitt (2002) assessed the incremental validity of an SJT over multiple criteria in comparison to GMA, the Big Five personality traits, and previous experience. The findings of this research demonstrated that the SJT accounted for

significant incremental variance across four performance criteria: task performance (ΔR^2 = .05), motivational contextual performance (ΔR^2 = .08), interpersonal contextual performance *(*ΔR^2 = .03), and overall job performance (ΔR^2 = .04). Finally, examining multiple variables, Weekley and Ployhart (2005) reported on the incremental validity of an SJT beyond GMA, GPA, Big Five personality, and a multidimensional measure of previous experience including training experience, general work experience, and job tenure dimensions. The findings indicated that the SJT accounted for significant incremental variance in predicting managerial performance (ΔR^2 = .02).

The incremental validity of SJT scores is important for the current research. Specifically, SJT scores should have increased incremental validity over measures of constructs that they are less correlated with, as they are accounting for variance which is unaccounted for by that construct. Thus, if a scoring method better captures a particular construct, we would expect it to have less incremental validity over that construct relative to an alternate scoring method. If scoring method effects incremental validity, the understanding of these moderating effects would help to put incremental validity scores in perspective.

Summary of Chapter Two

Although research has established the criterion validity, reliability, and face validity of SJTs, there are several factors that could impact these estimates. Additionally, research has established the correlations between SJT scores and other variables of interest (e.g. personality scores, and GMA). However, these results often have a large amount of variability, and with SJTs having a wide variety of characteristics and domains, there are several moderating factors to examine.

The introduction of the Implicit Trait Policy scoring method has demonstrated how personality traits can better be captured by SJTs when compared to traditional methods. With the introduction of this unique scoring technique, the moderating effect of scoring techniques on correlations between SJT scores and other relevant variables is highlighted. This paper will attempt to demonstrate the moderating effect that scoring techniques have on the underlying dimensions of SJTs and in the prediction of criteria. Specifically, differential relationships will be explored with the predictors of personality and GMA. Additionally, differential relationships will be explored with the criteria of typical and maximum performance.

CHAPTER THREE: RESEARCH PROPOSALS

The multidimensional and domain-specific nature of SJTs has frequently resulted in low internal consistency reliabilities. This has presented difficulty when attempting to design SJTs to measure specific constructs and attempting to determine construct validity (Schmitt & Chan, 2006). Attempts to determine factor structure have been inconclusive likely due to the large number of potential differences/moderating variables between SJTs. This has resulted in the practice of using the overall SJT scores as opposed to scoring individual dimensions. For example, Oswald and colleagues (2004) used a domain-sampling approach to develop 12 dimensions of student performance. The authors found that the inter-correlations between the scales were exceedingly high, and the alphas for each scale were relatively low. An exploratory factor analysis revealed a large general factor in this particular SJT. The authors opted to use a composite of the scales as opposed to attempting to measure different dimensions, and this composite resulted in an internal consistency reliability of $\alpha = .85$. This research supports the idea that SJTs may have difficulty capturing individual dimensions because there are so many underlying correlates inherent in every scenario presented in SJTs (e.g. previous work experience, GMA, and personality traits such as Extraversion and Agreeableness).

Construct validity is important to consider when implementing an SJT for selection or training purposes, as the differential relationships with constructs of interest will affect the level of adverse impact, predictive validity, and the specific relationships with criteria of an SJT. The issue of SJT construct validity is very complex, as several factors can have an influence on the construct validity of an SJT. These factors include the development method, SJT instructions, or keying method of the test (Motowidlo & Beier, 2010; Chan & Schmitt, 1997). When considering the different constructs that SJTs measure, factor structure and inter-correlations can be difficult

to determine (Schmitt & Chan, 2006). When exploring these differential relationships, the understanding of any moderating factors can help researchers account for additional variation and obtain a clearer picture of the constructs with which SJTs have consistent relationships and how these relationships can best be captured.

As described in the previous sections of this dissertation, for instance, Motowidlo and Beier (2010) found that the keying method can influence the relationship SJT scores have with both declarative knowledge and personality, such that those keys developed by job incumbents will have a stronger relationship with knowledge (specifically, job-relevant knowledge). However, individuals not familiar with the job will develop keys that are more strongly correlated with personality. As keying method can have profound effects on the relationships between SJT scores and variables of interest, it would logically follow that scoring method may also have similar effects. As previously explained, keying method involves what scores are assigned to which event item responses (e.g. having SMEs rate which of the developed item responses would be the best and worst answers). The scoring method determines what is asked of the individual taking the test (e.g. choose the most effective response versus rate the effectiveness of each response) and how this data is interpreted (e.g. given one point for correctly identifying the best response versus calculating a distance score from the most effective to the least effective response). Due to the theoretical foundations of Implicit Trait Policies (Accentuation Effects, Contrast Effects, and Assimilation Effects) it is likely that certain methods will be more strongly related to particular traits because they capture these effects. This research will attempt to explore the moderating effect of scoring method, determine which scoring method will be suitable depending upon which constructs the test is attempting to capture, and which outcome (maximum or typical performance) the test is attempting to predict.

Personality

It has been demonstrated that individuals higher in a particular personality trait perform better on dimensions related to those traits in assessment situations in which they must determine the targeted dimensions themselves but not when they are told what dimensions are being targeted (Smith-Jentsch, 2007). This finding supports the notion that individuals higher on a particular trait are more likely to correctly guess when behavior associated with that trait is being targeted. This should lead them to also identify when levels of that trait are being varied in SJT response options. However, higher levels of these particular constructs may also be subject to Assimilation and Contrast Effects. These decision-making biases may make accuracy more difficult in determining the best and worst SJT item responses (Mascaro, 1969). Additionally, the decision-making bias of Accentuation Effects of those individuals high in the targeted constructs may cause larger distance scores between the highest and lowest rated item audio responses.

Because ITP theory is built on the theoretical foundation of Accentuation Effects, which state that scores will be exaggerated, this exaggeration is not captured by simply scoring the accuracy of best/worst judgment. Based on the theoretical tenants of Accentuation Effects, we would expect that when evaluating the effectiveness of SJT event item responses that vary in their levels of a particular personality trait. Greater discrepancies should exist between the effectiveness rating assigned to a participant's chosen 'best response' and the effectiveness rating assigned to their chosen 'worst response' the higher that participant is on that particular personality trait. For an illustration of this, figure 2 demonstrates the intended relationship between effectiveness levels and traits, while figure 4 illustrates the results of an Accentuation Effect on the data resulting from an individual who is high in a particular trait. This is due to the functions of the Attenuation Effect that causes exaggerations in effectiveness ratings and would

make it most viable to have participants rate all of the potential response items and measure the distance score to capture this effect. Motowidlo, Hooper, and Jackson (2006b) have found support for the effects of Accentuation on ratings of effectiveness for SJT items. Specifically, results demonstrated that those high in a personality trait demonstrated more exaggerated rating responses when ITP scores were calculated as distance scores between the most and least effective items. The Accentuation Effects cause an exaggeration of the effective/ineffective score that is influenced by value judgments that in turn result in an increase in the distance score. For this study, ITP scores will be captured using a distance score which will capture the distance from the item response rated as most effective to the item response rated as least effective to capture the Accentuation Effects. ITP / distance scores will hereafter be referred to as distance scores.

Additionally, when considering the theoretical foundations of Implicit Trait Policies, contrast and Assimilation Effects will also have a profound effect. One may expect that individuals who possess a high level of a trait may be better able to identify SJT items in which levels of that trait are being systematically varied; however they should also be less able to distinguish the event item responses similar to their position or opposite from their position due to the biasing effects of the 'latitude of acceptance' and the 'latitude of rejection' (Hovland & Sherif, 1952). The 'latitude of acceptance' refers to the breadth at which one will accept opinions or positions similar to their own. 'Latitude of rejection' refers to the breadth at which one will reject opinions or positions dissimilar to their own. For example, if an individual is high on Agreeableness, he or she is likely to have larger latitudes of acceptance and rejection for Agreeableness. Thus they may be expected to make simpler or more global distinctions between event item responses that are consistent or inconsistent with the targeted trait. However, their

ratings of item responses within the group of scripted item responses inconsistent with that trait should show less differentiation due to Contrast Effects. Similarly ratings of event item responses generally consistent with the targeted trait should show less differentiation due to Assimilation Effects. Hovland and Sherif (1952) hypothesized as much in their initial description of social judgment:

> From the present results there emerges an interesting possibility for developing a behavioral, "projective" method of attitude measurement through study of the way an individual sorts (judges) statements on an issue. If the tendencies found in the present experiment for individuals with extreme positions to bunch up the statements at the extremes are found for other issues, it may be possible to assess the attitude of an individual without ever asking him his opinion but by relying entirely on the way he distributes his judgments. Individuals with more or less neutral attitudes would be expected to space their judgments rather evenly over the entire range, those at the pro end would tend to reject neutral items and hence pile them up at the anti end, and those with anti attitudes would place them at the opposite end of the scale (p. 831).

Others have found support for Hovland and Sherif's contentions. Mascaro (1969) studied 87 participants and by measuring attitude extremity and latitudes of acceptance, rejection, and non-commitment. Results demonstrated relationships such that those with the most extreme positions tended to have the largest latitudes of acceptance and rejection while having the smallest latitude of non-commitment. These results suggest that the more extreme a position is, the wider the latitudes of acceptance and rejection will be, which may result in participants making less of a distinction in their judgments of effectiveness of positions similar to theirs. This means that they would not necessarily be better able to single out the best or the worst response

among those item responses generally high or low in effectiveness. For an illustration of this, figure 4 demonstrates the bunching that may occur due to assimilation and Contrast Effects, with the best choice answers highlighted in a larger data point, while figure 1 demonstrates the intended relationship between effectiveness ratings and intended trait levels. Thus it is hypothesized:

H1: Self-reported personality measures will be more strongly related to distance scores derived from an SJT than to "best choice" scores.

General mental ability

General mental ability (GMA) has been related to the ability to identify targeted dimensions in assessment situations, regardless of which dimension is being measured (e.g. Melchers, Klehe, Richter, Kleinmann, Konig, & Lievens, 2009; König, Melchers, Kleinmann, Richter, & Klehe, 2007). GMA is also related to the accuracy with which an individual can rate levels of personality traits and performance effectiveness (e.g. Harris, Vernon, & Jang, 1999; Lippa & Dietz, 2000). For example, Lippa and Dietz (2000) found that individuals higher in GMA are more accurate when assessing Extraversion, Neuroticism, and masculinity/femininity. Harris, Vernon, and Jang (1999) found that intelligence was related to the accuracy with which one twin could answer a personality inventory rating the other twin. Hauenstien and Alexander (1991) found that intelligence was positively related to an individual's accuracy at rating performance of lecturers in taped presentations. Finally, Smither and Reilly (1987) found that intelligent individuals were better able to accurately rate the performance of individuals after watching them at work, and that these ratings actually predicted objective performance. Thus, research has supported the link between intelligence and accuracy at rating another individual. Research has also demonstrated that intelligent assessors tend to make less rating errors (Davis,

2000) and may thus be less prone to the biases that may influence others scores such as stringency, leniency, exaggeration, assimilation, contrast, and accentuation. See figure 3 for an example of the relationship between effectiveness ratings and intended trait levels as determined by an individual high in GMA, with the item responses the participants rated to be the best and worst highlighted in bold.

In sum, research has supported the notion that GMA enables individuals to identify targeted dimensions of performance, to more accurately identify best and worst event item audio responses with respect to those dimensions, and to avoid bias in their ratings. Thus it is hypothesized:

H2: GMA will be more strongly related to best choice scores derived from an SJT than to Distance scores.

When considering typical and maximum performance, Smith-Jentsch (2007) found that transparency would reduce the relationship between conceptually matched typical performance predictors and dimension ratings in an assessment center. Related to this, one could infer that as a best choice scoring method is more simplistic, it will in turn be more transparent. In other words, when a participant is asked to choose the best and worst answers, it is quite easy for them to determine that their answers will be scored as either being correct or incorrect, and they can thus make an intelligent choice to determine which answer they should mark as the best and worst. However, when a participant is asked to rate the effectiveness of each SJT item response, and a distance score is derived from this response, it may not be clear to the participant how their effectiveness ratings will be evaluated. It would thus be likely that the best choice scoring method would result in lower correlations between personality-based SJT scores (a typical performance predictor) and peer ratings of typical behavior, just as transparency would result in

a lower correlation between typical performance predictors and dimension ratings (Smith-Jentsch, 2007). A measurement of typical performance is often obtained through peer ratings, as peers are often in sustained contact with the individual and can judge the individual's "will do" job performance (e.g. Ployhart, Lim & Chan, 2001; Sackett, Zedeck, & Fogli, 1988). Thus as this one of this study's criterion is peer ratings, it can be expected that this criteria will be a rating of typical performance. As previously stated, it has been demonstrated that personality is often most strongly associated with typical performance, while GMA is often most strongly associated with maximum performance (Sackett, Zedeck, & Fogli, 1988). As it is hypothesized that Distance scores will be most closely related to personality, it can then be hypothesized that:

H3: Distance scores derived from an SJT will have a stronger relationship with typical performance than will best choice scores.

Conversely, when examining the prediction of maximum performance, it is important to remember that GMA is a strong predictor of maximum performance (Marcus, Foggin, Johnston, & Rothstien, 2007). Due to the expected relationship with GMA and best choice scoring methods, one can logically infer a relationship between maximum performance and the best choice scoring methods. Beyond that, the definition of maximum performance states that under maximum performance scenarios, performers are aware their performance is being observed and evaluated, are instructed to perform their best, and have a mean performance which is judged from a brief period (Sackett, Zedeck, & Fogli, 1988). It should be noted that asking an individual to choose the best response will require significantly less time than asking the participant to rate the effectiveness of five response items and may result in maximum performance relative to the ITP scores as the best choice scores prevent fatigue. As previous research has demonstrated that those who possess the ability to fake or to answer more accurately will take advantage of the

opportunity to do so (Levashina, 2009), it can be inferred that those with higher GMA will be more accurate in a transparent / simplistic situation and this accuracy will inflate their scores and thus increase the correlation between SJT scores and maximum performance ratings. Additionally, as GMA is considered a predictor of maximum performance and personality is a predictor of typical performance (Sackett, Zedeck, and Fogli, 1988) we can expect that personality indicators such as Distance scores will be more strongly associated with typical performance rating. Thus it can be hypothesized:

H4: Best choice scores derived from an SJT will have a stronger relationship with maximum performance than will distance scores.

SJT scores previously have been demonstrated to partially mediate the relationship between personality and criterion of procedural knowledge (Motowido, Hooper, & Jackson, 2006). Additionally, procedural knowledge has been demonstrated to predict work performance and to be related to personality traits (Motowidlo, Crook, Kell, & Naemi, 2009). As personality is to some extent heritable (Veselka, Schermer, Petrides, & Vernon, 2009), we can expect it to influence the trait of procedural knowledge (which is more variable over time) represented by SJT scores, which will in turn influence on-the-job behavior and performance. The SJT scores can be expected to transmit knowledge. It has been postulated that SJTs capture some aspect of procedural knowledge. Previous research has supported this, as Weekly and Ployhart (2005) determined through path analysis that the effects of personality on performance were partially mediated by SJT scores. Based on the previously established relationships, and the different focus of SJT testing from classical personality and GMA tests, which will result in less than full mediation, it can be hypothesized:

H5: Distance scores derived from an SJT will partially mediate the relationship between personality and typical performance.

It has been previously hypothesized that ITP scores will be more strongly associated with personality, and best choice scores will be more strongly associated with GMA. Based on these relationships, incremental validity can be inferred for these variables. Specifically, due to the ability of an SJT to capture procedural knowledge (Motowidlo & Beier, 2010), the SJT score should capture variance beyond the predictor variable. SJT scores are multifaceted and capture additional information beyond GMA and personality, such as multiple work-related KSAs (McDaniel & Nguyen, 2001; McDaniel & Whetzel, 2005; Weekley & Jones, 1999), teamwork knowledge (McClough & Rogelberg, 2003; Morgeson et al., 2005; Stevens & Campion, 1999), employee integrity (Becker, 2005). Previous research has supported this incremental validity. For example Weekley and Jones (1999) found SJTs had incremental validity above GMA and work experience to the level of ΔR^2 ranging from .111 to .096 in predicting work performance. Lievens and colleagues (2005) found SJTs had incremental validity above cognitive and factual tests in predicting school performance over four years, with changes to the magnitude of ΔR^2 = .01 *(ns),* .02, .06, .07 in each year, respectively. Finally, Chan and Schmitt (2002) found SJTs had incremental validity to the level of ΔR^2 = .05 over GMA and the Big Five personality dimensions in predicting task performance. With regard to the hypothesized effects of scoring technique, it can be expected that those scoring techniques which capture variance that is less related to the predictor variable will have additional incremental validity over those variables. Thus it is hypothesized:

H7: Best choice SJT scores will have greater incremental validity over personality measures in explaining variance in typical performance than will Distance scores.

H8: When traditional personality measures are not included, best choice SJT scores and distance scores will each contribute incrementally to the prediction of typical performance.

These hypotheses form a theoretical model (See Figure 1) that demonstrates the effects of scoring method on the relationships SJT scores have with predictor and criteria variables. As scoring methods can be relatively easily manipulated, it is important to explore these moderating effects. The differential prediction of maximum and typical performance is of direct relevance to selection testing. Additionally, exploring the potential to implicitly measure personality is of great importance to the application of Industrial/Organizational Psychology to the field. The exploration of the effects of scoring techniques has the potential to allow researchers to tailor SJTs to capture particular traits and predict specific criteria.

Now that the theoretical framework has been established, the next step is to discuss the development of the SJT, the methodology of the study utilized to support these hypotheses, and the analyses and results. In the next chapter, the development of the SJT based on established practices and scientific principles will be discussed, in addition to the results of pilot tests.

(*see* Figure 1).

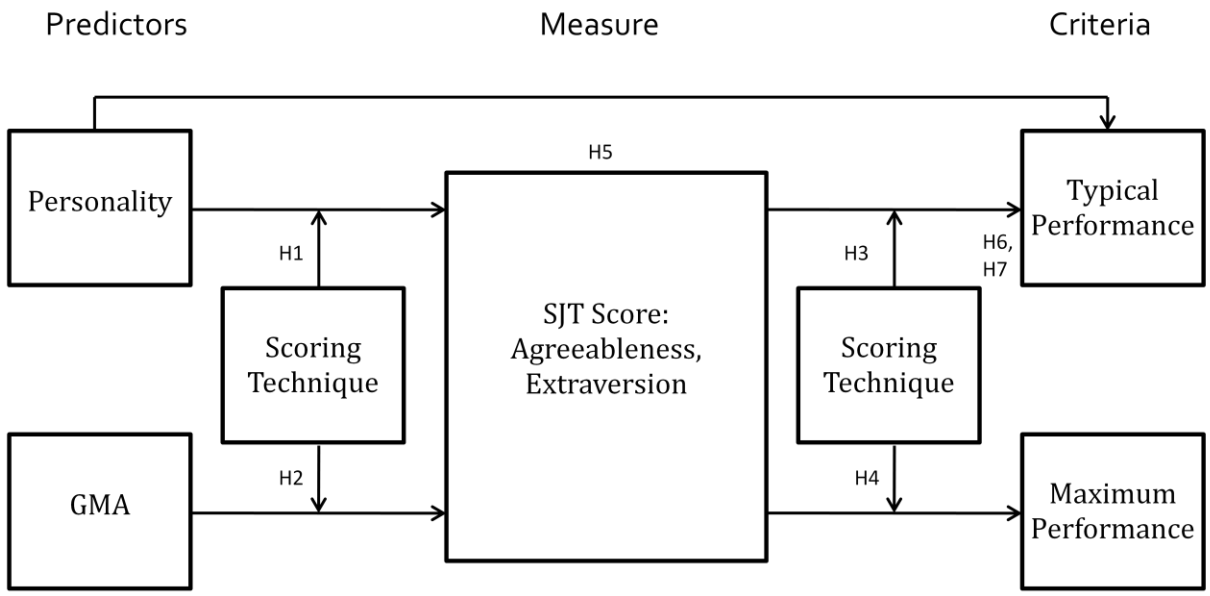


Figure 1. Hypothesized model of scoring technique interactions

(*See* Figure 2)

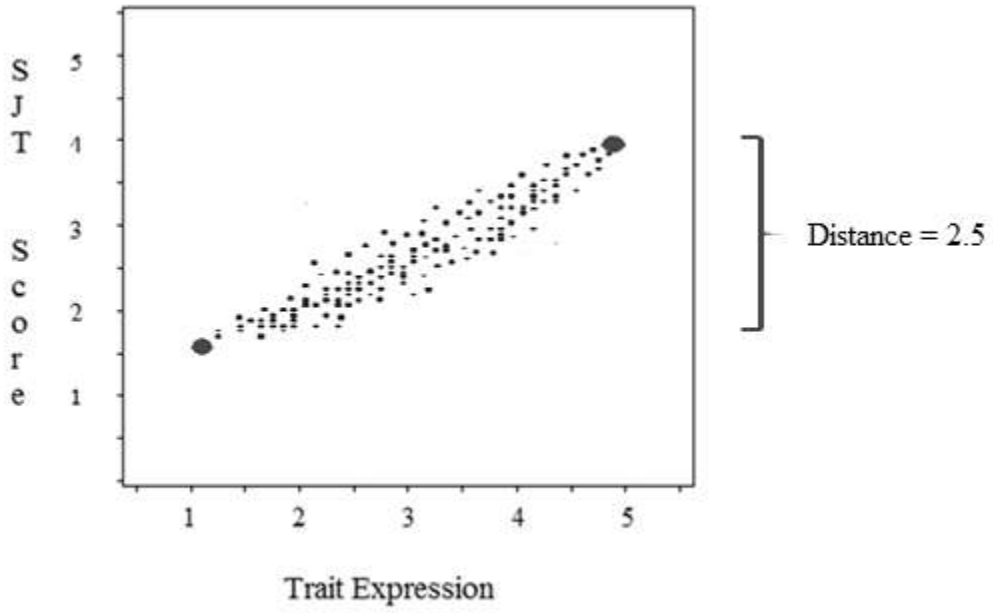


Figure 2. Graphical depiction of the intended level of trait expression to the effectiveness SJT score

(*See* Figure 3)

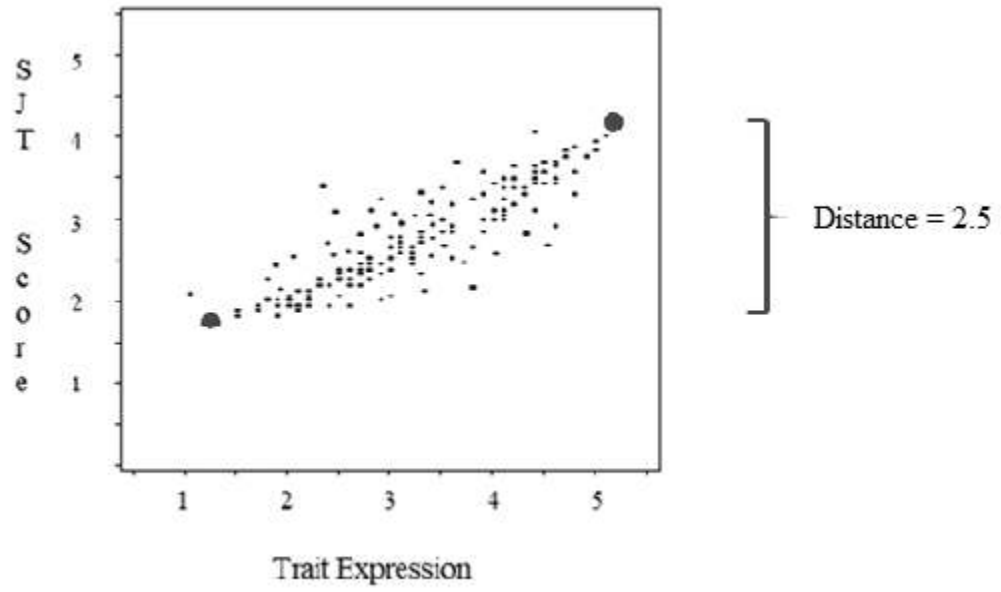


Figure 3. Graphical depiction of the relationship between SJT Scores and trait expression for an individual high in intelligence

(*See* Figure 4)

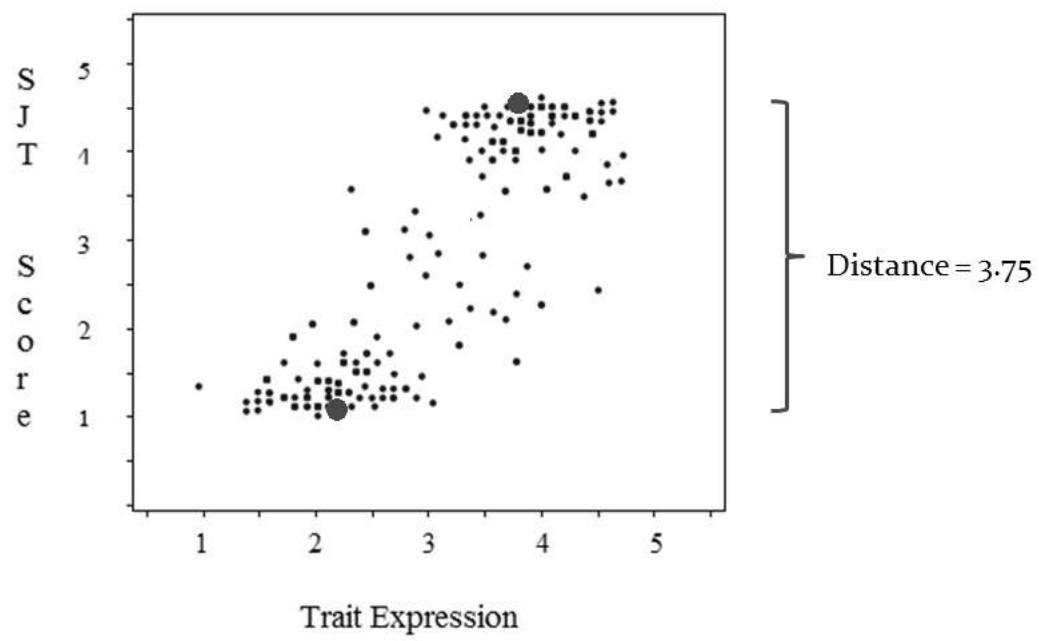


Figure 4. Graphical depiction of Contrast, Assimilation, and Accentuation Effects on data, as would result from an individual who is high on the targeted dimension

CHAPTER FOUR: DEVELOPMENT OF CUSTOMER SERVICE SITUATIONAL JUDGMENT TEST

The first step in demonstrating support for the above hypotheses was to properly develop a customer-service based SJT containing events with the potential to activate personality traits of interest to a customer service job. Specifically, these traits of interest are; Agreeableness, Conscientiousness, Emotional Stability and Extraversion. Openness to Experience was not included due to the fact that it has not been strongly related to performance in customer service positions (Frei and, 1997; Hogan, Hogan, & Busch, 1984; Ones & Viswesvaran, 1996) or more generally to SJT scores in meta-analyses (McDaniel et al., 2006). Further, Openness to Experience was not included as item responses on an SJT could not effectively be varied on this trait. The SJT was developed in a multi-media video-based format, as is the current trend in SJT research, and which should result in a higher fidelity simulation relative to text-based SJTs. Response options are to be provided as audio clips. This should provide more criterion-validity and face-validity to the SJT, in addition to more positive applicant reactions.

The development of this SJT was funded by an agency concerned with the training and certification of individuals making the transition from being on welfare to an employment setting (Workforce Central Florida, 2006). As such the target job was a customer service representative, as many of these individuals were seeking employment in such positions. The setting chosen was an emergency room waiting area. This high stress environment requires employees to exhibit Agreeableness, Conscientiousness, Emotional Stability, and Extraversion to resolve difficult situations. The events were based on critical incidents gathered from subject matter experts (SMEs), interviews with SMEs, and direct observation of on-the-job performance. This methodology coincides with the currently accepted methodology of SJT development, of which the first step is to gather critical incident reports from a pool of SMEs. The events were scripted

in such a way so as to include interaction with customers and coworkers, as well as interactions with different genders and ethnicities. Additionally, these events were scripted to capture particular dimensions of customer service behavior that can be conceptually mapped to the traits in the Five Factor Model (Frei, 1997; Hogan, Hogan, & Busch, 1984; Ones & Viswesvaran, 1996).

In order to correctly determine which events should be included in the SJT, and to script the item responses for the SJT, two pilot studies were conducted. These pilot studies continued with the accepted practices for SJT development and validation – practices such as using SMEs to script actual item responses, basing scripts on actual participant data, and having SMEs rate the effectiveness of each response.

Pilot Study One: Trait Activation Potential of Stimulus Events

Objective

An initial investigation was conducted to determine which of the stimulus events had the trait-activation potential for the relevant traits. Trait-activation potential refers to the capacity to observe differences in trait-related behaviors within a given situation (Tett & Guterman, 2000). Additionally, this study was conducted in order to establish which of the Big Five traits would be determined by participant data to be activated by enough events in order to justify inclusion in the final SJT. This was done by determining which events expressed traits in such a way that individuals high in those particular traits would be able to identify that trait as important to resolve the situation, while individuals low in that trait may not be able to make such a distinction. Although the events were already designed to capture several interpersonal constructs, it was important to determine where there was trait-related variance in the ability to

identify the targeted construct. In other words, the study attempted to determine for which events individuals high in the trait in question identified that trait as being necessary to resolve the situation. In order to accomplish this, a group of participants viewed each event and indicated the traits they believed the event was designed to assess. Thus, participants respond as to which traits they believed were being assessed in each event, and these responses were correlated back to the actual self-report trait levels of these participants for each event. This procedure will help determine which events best elicit trait-related variance, as only those individuals high on the trait in question should be able to determine that the event is assessing the construct in question if the event is good at distinguishing high-trait from low-trait individuals. For some events, the trait which is being measured may be too obvious to elicit any variation in responses from participants. In other words, the events may be transparent, such that individuals who are either high or low in the trait would both be able to determine equally well what trait the event is attempting to measure. This would result in minimal variation, similar to a ceiling effect. Conversely, some events may be excessively difficult even for those high on the relevant trait to determine the correct trait that the event was attempting to capture. Again, this would result in minimal variation. For this reason, the correlation between self-reported NEO scores and number of relevant adjectives listed per event was calculated for Agreeableness, Conscientiousness, Emotional Stability, and Extraversion. The relevancy of adjectives for particular traits was determined by having subject matter experts sort the adjectives into the four traits of interest. After correlations were calculated, the number of events which had significant correlations for each trait was determined, and conclusions were drawn as to which of the traits would be included in the final SJT.

Methodology

Participants were presented with 29 events scripted and filmed based on our job analysis of the emergency room customer service representative. After viewing a particular event, participants were asked to identify the correct skills or traits necessary to resolve the situation presented. These participant responses were provided in blanks, with the participants receiving minimal instruction regarding the specific types of traits or skills being requested. Participants were allowed to list as many written self-report responses as they believed necessary in order to resolve the situation. Initially, there were hundreds of unique skill names generated across participants and events. Upon examining these responses from the participants, it became clear that the response adjectives could be collapsed into a smaller number of categories based upon the similarities between some of the responses provided by participants. For example, the skill traits "being truthful" and "not lying" could both be sorted into an "honesty" category. Additionally, "handling many things at once" and "dividing attention between tasks" could both be sorted into a "multitasking ability" category. Coders examined these responses from the participants and using a card-sorting method determined that 36 unique categories could be extrapolated from the written open-ended trait responses. The research assistants used a card sort task, sorting and resorting adjectives into categories until final definitive categories were established based on consensus. Specifically, the categories established were; consideration, getting along with others, willingness to compromise, teamwork, ability to pacify, conflict resolution, conforming, empathy/compassion, loyalty, communication, confidence, persuasiveness, leadership, composure, ability to work under pressure, coping skills, rational behavior, calm confrontation, patience, prioritizing, honesty, organizational skills, responsibility, credibility, following rules, time management, knowledge of rules, ethics/morals, multitasking

ability, efficiency, professional behavior, fairness, objectiveness, and taking charge. Later, graduate students familiar with the Big Five personality traits assigned these 36 trait adjectives into the broader personality traits of Agreeableness, Conscientiousness, Emotional Stability, and Extraversion. See figure 5 for an illustration of the results of coding these participant's responses into the overarching categories. Thus, both bottom-up and top-down methods were employed in to ensure participant's responses were coded into the correct personality traits.

When considering the dimension ratings of these adjectives sorted into the traits of Agreeableness, Conscientiousness, Emotional Stability and Emotional Stability, the two original raters' reliability was estimated. In other words, the 36 trait adjectives were coded into the four broad dimensions. Then for each event the agreement between the two raters regarding how many of the participant's responses were sorted in these four broad dimensions was calculated. Average correlations indicated interrater reliability was (r = .70) for Extraversion, (r = .85) for Agreeableness, (r = .68) for Conscientiousness, and (r = .65) for Emotional Stability.

After assigning the adjectives into the different personality traits, correlations were calculated between participant scores on the four traits within the NEO and the numbers of adjectives they had listed for each event that were coded as being conceptually-related to those traits. These overarching open-ended dimensions of Agreeableness, Conscientiousness, Emotional Stability and Extraversion were calculated for each participant for every event based on how many written responses relevant to the particular trait they had listed as helpful for resolving the situation at hand. The participant's NEO scores for the corresponding traits were the correlated to the number of these relevant written responses for each event. Ideally, good events should have the trait-activation potential that would allow individuals who are high in the relevant trait to correctly identify the importance of utilizing that trait in order to resolve the

problem. Thus the stronger the correlation for an event between the participants NEO scores and the open-ended participant written responses, the stronger the trait activation for that particular event.

Results

Table 1 demonstrates the relationship of the participants' self-report traits to the number of relevant trait terms used for each event. Specifically, the correlations are the correlations between the participant's NEO score and the number of relevant trait terms identified for the specific trait in question for that particular event. To be included in the SJT, the threshold of achieving a correlation at or above $r = .15$ (or $p < .1$, one-tailed) was set. Thus for Agreeableness the events selected that met this criteria, had positive correlations, and were not more strongly correlated with another trait were 2, 4, 6, 7, 13, 21 and 27. Event 13 was subsequently discarded as the event could not logically be scripted for Agreeableness. Cronbach's alpha for this scale, meaning the intercorrelation of these events with each other event using the measurement of the number of open-ended participant adjective responses coded as Agreeableness as the dependent variable, was .643. To tap the construct of Extraversion, the events selected were 1, 5, 9, 12, 14 and 24. Cronbach's alpha for this scale, the combination these events with each event measuring the number of open-ended participant adjective responses coded as Extraversion, was .649. The events that were correlated with Emotional Stability were 3 and 17. Cronbach's alpha for these two events, the combination of these events with each event measuring the number of open-ended participant adjective responses coded as Emotional Stability, was .384. There were no significant positive correlation between Conscientiousness and any of the number of open-ended participant adjective responses for any of the events.

Table 1. Inter-correlations between traits and events

Scale Dimension	Agreeableness	Conscientiousness	Extraversion	Emotional stability
1. Event 1	.15*	-.17*	.30*	.01
2. Event 2	.19*	-.07	-.17	.06
3. Event 3	.04	-.05	.11	.25*
4. Event 4	.22*	.10	.14	-.05
5. Event 5	.11	-.10	.15*	-.14
6. Event 6	.26*	.00	.06	.12
7. Event 7	.25*	.01	.13	-.10
8. Event 8	.06	-.06	.09	.04
9. Event 9	.00	.01	.17*	.01
10. Event 10	-.14	.08	.09	.02
11. Event 11	-.12	.00	.10	.06
12. Event 12	.13	-.02	.23*	.01
13. Event 13	.16*	.04	-.04	-.01
14. Event 14	-.09	.08	.19*	-.09
15. Event 15	.01	.06	.05	-.14
16. Event 16	.10	-.09	-.04	-.08
17. Event 17	.11	-.16	.15	.16*
18. Event 18	.01	-.02	-.14	- .07
19. Event 19	.10	.11	-.03	-.25*
20. Event 20	.07	.04	-.12	.00
21. Event 21	.17*	-.10	-.13	-.19*
22. Event 22	-.12	-.06	.03	-.20*
23. Event 23	.11	-.01	.02	.07
24. Event 24	.11	-.07	.18*	.03
25. Event 25	.02	-.11	-.07	-.08
26. Event 26	.11	-.02	-.05	-.20*
27. Event 27	.17*	.16	-.01	.11
28. Event 28	.07	.13	.05	- .01

* Correlation is significant at the 0.1 level (1-tailed).

Discussion

Results of the pilot study indicated that those individuals higher on particular traits as determined by a self-report NEO were better able to identify those traits as being necessary in certain situations, as determined by the significant relationships between NEO scores and adjective responses to events. These results also aided in determining which traits would be included in the final SJT product, as well as what events would comprise these traits. After analyzing the results, it was decided that Conscientiousness would not included in the final SJT product as there were no significant relationships between self-report data and the participant adjective responses. Emotional Stability will also not be included as there were only two significant relationships, which would mean that this trait would not be composed of enough events to elicit any significant results within an SJT. This study was a necessary step in the development of a well-validated SJT instrument. This step was necessary to ensure events were accurately assigned to dimensions. The next step in the development process is the scripting of item responses that vary in the expression of traits for the events determined to be relevant.

Pilot Test 2: Trait Variation Within Item Responses

An important step in the development of any SJT is the development and validation of item responses. For each event, five item responses were scripted that varied on the amount of the trait expressed, with each of the five item responses hypothetically expressing a different level of the relevant trait. The scripted responses were carefully constructed to prevent overlap (in other words, while the item responses from one event may vary on Agreeableness, the level of Extraversion of those item responses was held constant). Additionally, during the scripting process the reading level of the item responses was considered and held constant to prevent any confounding effects from participants considering events that use more advanced vernacular as

more effective. When scripting events to capture Implicit Trait Policies it was important that there was a correlation between effectiveness and trait expression, such that the choice that was scripted to express the highest level of a trait also was the most effective response from a logical standpoint (Motowidlo, Hooper, & Jackson, 2006a). See figure 2 for an example of the intended relationship between SJT effectiveness scores and the expressed level of the targeted trait. During scripting, several aspects were considered. First, I considered the adjectives participants in the pilot study one used to describe the dimensions targeted. For Agreeableness, these trait terms were respect, composure, getting along with others, consideration, calm confrontation, empathy/compassion/sympathy, conforming, willingness to compromise, patience. For Extraversion these trait terms were communication, confidence, persuasiveness, leadership, and taking charge. The scripted item responses for the traits attempted to reflect varying levels of these terms. Second, I considered theory and prior research on the manner in which individuals express the two traits behaviorally (e.g. Penley & Tomaka, 2002; Trouvain, Schmidt, Schröder, Schmitz, & Barry, 2006). Previous research has demonstrated that individuals high on particular traits will use certain mannerisms and vernacular when speaking. For example, for Extraversion the answers were varied on length and energy, and for Agreeableness the answers were varied on empathy and ability to provide a win/win solution. Additionally, literature provided guidelines for the verbal expression of particular traits. For example, extraverts are likely to speak with more assertiveness, be quicker to respond, and speak loudly and rapidly (Markel, Phillis, Vargas & Howard, 1972; McCroskey, Heisel, & Richmond, 2001). Agreeable individuals are more likely to speak less frequently, more slowly, and have softer communication patterns (Markel, Phillis, Vargas & Howard, 1972; McCroskey, Heisel, & Richmond, 2001). Third, I reviewed audio-taped responses from a previous study in which participants verbally responded to the

twelve events. Participants who were the lowest and highest scoring individuals on Agreeableness and Extraversion were identified based on their NEO scores for Agreeableness and Extraversion. The audio responses to these items were then listened to and the script based on their actual responses. This was done to ensure SJT item responses reflected what individuals would actually say in response to the particular situation. Additionally, these item responses were examined to determine the verbal speaking patterns of individuals high or low in particular traits and to aid in the verbal audio recording of item responses.

After the scripts were developed, a group of eight PhD students read the scripts and suggested changes to ensure the item responses conveyed the appropriate level of the specified trait. After the scripts were finalized, item responses were recorded using individuals that had been trained on the expression of the particular traits based on current research. These individuals were also trained by having them listen to the audio results from the previous simulation study. Audio events were recorded by a male and a female reader. The male audio recordings were conducted first, with the female recorder attempting to replicate the tone and pace of the male audio exactly. Both male and female audio was recorded for each of the five item responses for all 12 events. It was decided that in the final SJT participants will be presented with either five male or five female responses, depending upon the gender of the participant in order to prevent any interaction between the gender of the participant and the gender of the individual reading the presented verbal item response. The five male and five female responses will be identical. With the recordings complete, the audio was brought before a panel of eight PhD students who listened and provided feedback. Specifically, items had been scripted to express five levels of variation of the particular trait in question, and the panel of

students was used to determine whether the responses reflected the intended trait levels. On the basis of feedback from these students, items were re-recorded as necessary.

A separate PhD student who was not familiar with the development process then provided assessments for each stimulus event as to which of the two targeted traits (i.e., Agreeableness, Extraversion) was being varied in the associated item responses. This student identified the correct targeted trait for 91.6% of the events. Next, this student rated each item response as to the level of the targeted trait that was expressed on a five-point likert scale. The correlation between the student's ratings of the trait levels of each item response and the intended trait level of the scripted item response was $r = .83$. In order to determine whether the male and female versions of the identical item responses were perceived to reflect the same trait levels, a correlation was computed between the student's ratings of trait level for the male and female responses. This correlation was $r = .99$.

Current Study

The final SJT product resulting from this development process required participants to view and listen to audiovisual clips of team-based workplace vignettes from the previously mentioned simulation. Participants were presented with a series of audio clips of potential responses to the situations that vary on the targeted dimensions as determined from the first pilot study (Agreeableness and Extraversion). Only those events which had a significant level of variation that relates to self-report NEO scores were used.

These SJT item responses were developed based on previous studies, SME input and current research, as discussed in the second pilot study. The item responses were scripted after examining the responses of participants who actually responded to the events in an initial administration of the simulation. These responses have also been scripted with consideration of

current research which gave insight into verbal nuances to include. Finally, the results of the card sort by SMEs of adjectives into different categories were considered during scripting.

As such, the development of both the events and the item responses was founded in scientific principles and was based on the results of previous guidelines laid out for SJT development (e.g. Motowidlo, Dunnette & Carter, 1990). The final SJT product is a high-fidelity SJT which should validly predict customer service performance and will hypothetically be related to personality or intelligence depending on the scoring method employed.

(See Figure 5)

Agreeableness:	**Conscientiousness:**	**Extraversion:**	**Neuroticism:**
Consideration	Prioritizing	Communication	Composure
Getting Along with Others	Honesty	Confidence	Ability to Work under Pressure
Willingness to Compromise	Organizational Skills	Persuasive	Coping Skills
Teamwork	Responsibility	Leadership	Rational Behavior
Ability to Pacify	Credibility	Taking Charge	Calm Confrontation
Conflict Resolution	Following Rules	(Mean = 3.74, SD = .41)	Patience
Conforming	Time Management		(Mean = 2.64, SD = .65)
Empathy / Compassion	Knowledge of Rules		
Loyalty	Ethics / Morals		
(Mean = 3.75, SD = .48)	Multitasking Ability		
	Efficiency		
	Professional Behavior		
	Fairness		
	Objectiveness		
	(Mean = 3.59, SD = .48)		

Figure 5. Results of SME card sort of participant written open-ended responses into overarching trait terms of Neuroticism, Extraversion, Agreeableness, and Conscientiousness

CHAPTER FIVE: METHODOLOGY

Participants

The study utilized 116 undergraduate students from a large southeastern university. The mean age of participants was 20.97 with a standard deviation of 5.27. The study participants were 34% male and 66% female. Regarding racial demographics, the participants were 55% Caucasian, 11% African American, 9% Asian, 3% American Indian, 3% Pacific Islander, and 19% Hispanic. These students participated in the research for course credit. A power analysis conducted using G*power statistical analysis program version 3.1.2 determined that 64 participants would be necessary to detect a medium effect size (roughly 0.3, based on previous research demonstrating correlations between SJT scores and variables of interest, e.g. McDaniel, Hartman, Whetzel & Grubb, 2007). This power analysis was also based on using a one tailed test with an α of .05 and 1-ß = .80. Additional participants were run through the SJT due to the large number and complexity of hypotheses.

Instruments

General Mental Ability

General Mental Ability (GMA)was assessed using the Wonderlic WPT-Q test, a shortened 30-item version of the Wonderlic which required eight minutes for the participants to complete. An example Wonderlic test item is "An instrument store gives a 10% discount to all students off the original cost of an instrument. During a back to school sale an additional 15% is taken off the discounted price. Julie, a student at the local high school, purchases a flute for $306. How much did it originally cost?" The participant would then be provided five multiple choice answers and be required to select the correct choice (see Appendix B).

Personality Inventory

Personality was assessed by the NEO Five-Factor Inventory (NEO-FFI), a 60-item shortened version of the NEO PI-R (Costa & McCrae, 1992). There were 12 items per personality trait. An example item for Agreeableness is "I try to be courteous to everyone I meet." An example item for Extraversion is "I like to have a lot of people around me." Participants were asked to respond on a scale of 1 to 5 with regard to how strongly they agreed with the statements presented. For Agreeableness, the scale had a reliability level of $\alpha = .73$, for Extraversion, the scale had a reliability level of $\alpha = .81$.

Control Variables

Demographic/Customer Service Information

Demographic information was collected regarding the participant's race or ethnic background. This information was collected by having participants indicate each race they associate with. The participants were given the choice of "White (Non-Hispanic)," "Black or African American," "Asian," "American Indian or Alaska Native," "Native Hawaiian or Other Pacific Islander," "Hispanic or Latino," and "Other: (Specify) _____." Additionally, if they select more than one racial group, the participant was asked to indicate which group the associate most strongly with. The participants were asked to indicate their age in an open-ended question, and indicate if they are male or female.

Customer Service Experience

Participants were asked if they have had previous customer service experience, and were asked to indicate the number of months he or she has maintained different customer service jobs.

The number of months that the participant had then spent in customer service was summed to a final customer service score. Results demonstrated that 72% of the participants had previous customer service experience. The mean length of customer service experience was 24.9 months, with a standard deviation of 44.6 months.

Situational Judgment Test

As described in the preceding sections, the SJT used in the present study consisted of 12 workplace events in a multi-media format. For each event, the participants listened to five potential item responses. Male participants listened to audio recordings of male-read item responses, while female participants listened to female-recorded item responses. The item responses were identical for male and female responses, and had been recorded to reflect the same inflection and tone. The participants were read instructions before taking the SJT that informed them that they have two tasks. For the first task, after viewing an event and listening to every potential response to the situation, the participant was instructed to mark on their answer sheet which response they believed was the best response for that situation and which response was the worst possible option. For the second task, after listening to each potential response to a situation, participants were instructed to rate each response item with regard to how effective they believe that response would be for that particular situation on a scale of 1-10. Participants were explicitly informed that they were not to rank the items; instead the same rating can be used for multiple item responses within the same event. Thus, in task one, respondents were asked to choose the best and worst response options while in task two they were asked to provide Likert-scale ratings of effectiveness for all possible response options. These two tasks were completed for each of the 12 events. When scoring the SJT, scores were calculated utilizing two scoring methods: distance scores and best choice scores.

Distance Scores

Distance scores were calculated by determining the distance between the effectiveness ratings participants assigned to the item responses participants identified as being the best and the item responses they identified as being worst for a particular event. Thus, if for one event the participant rated the most effective item response as a "seven" on a scale of one to ten, and rated the least effective item response as a "two" on a scale of one to ten, the participant would receive a distance score of six, as that is the number of scale points they are utilizing from two to seven, inclusive. For distance scores of Agreeableness, the reliability across 6 SJT events was $\alpha = .82$. For distance scores of Extraversion, the reliability across 6 SJT events was $\alpha = .81$. This is comparable to previous findings regarding distance scores, which demonstrated internal consistency (alpha) reliability estimates to be roughly .80 (Motowidlo, Hooper, & Jackson, 2006b). A total score was aggregated by taking the mean of the distance scores for the six events which measured Agreeableness and the six events which measured Extraversion.

Best Choice Scores

Best choice scores were calculated dichotomously for each event, meaning that the participant received one point for correctly identifying the best response and zero points for incorrectly identifying the best response for each event. These points were summed for an overall best choice score for each participant. For best choice scores of Agreeableness, the reliability across the 6 SJT events was $\alpha = .38$. For best choice scores of Extraversion, the reliability across the 6 SJT events was $\alpha = .22$. Reliability was considerably lower when this indexing method was used as compared to the distance scoring method in part due to the fact that these items were dichotomous in nature. This is also a lower reliability estimate relative to previous findings regarding best choice scores, which demonstrated internal consistency (alpha)

reliability estimates to be .50 (Ployhart & Ehrhart, 2007). Final scores were aggregated by calculating the mean of the number of items participants had correctly identified as the best response. In order to further improve scales, reliability analyses were run on every scale in order to determine if there were any scale items that did not relate well to other items in the scale. For the best choice scale for Extraversion, event 7 was removed as the removal of this item increased the cronbach's alpha from .22 to .29. For the best choice scale for Agreeableness, event 5 was removed as removal of this item increased the scale's cronbach's alpha from .38 to .45.

Maximum Performance Measure

As a measure of maximum performance, a multi-media based customer service simulation (Workforce Central Florida, 2006) was used. This simulation can be considered a measure of maximum performance because it meets the criteria defined by Sackett, Zedeck, and Fogli (1988). First, the participants were aware that they are being evaluated and what specifically they were being evaluated on. Second, the simulation was brief enough (approximately 40 minutes) to reasonably enable them to persist in their efforts. Third, the participants were given explicit instructions to perform their best. This simulation required participants to respond verbally and by typing email responses to a series of 32 events woven together into a seamless 40-minute work scenario. Twelve of the 32 events were the same events utilized in the SJT. The simulation required participants to respond to voicemails, emails, and interact directly with characters on the screen. Participants responded in written and verbal formats, and each participant's responses were audio-taped. These responses were then coded on the two targeted trait dimensions by trained raters. A coding scheme was used whereby raters assessed the audio responses from the simulation using scripted item responses from the SJT as scale anchors and assigning that trait level score to the simulation audio response. This score will

be referred to as the maximum performance score. The trained raters rated 12 simulation items, 6 items for each trait. After rating all simulation responses, the raters had an inter-rater reliability regarding maximum performance Agreeableness ratings of $r = .91$. For Extraversion, the raters had a level of interrater reliability of $r = .87$ thus ratings from the two raters were averaged for each simulation event. For ratings of Agreeableness, the reliability across 6 items was $\alpha = .49$. For peer ratings of Extraversion, the reliability across 6 items was $\alpha = .54$. The mean of the simulation ratings was calculated for Agreeableness and Extraversion to form two trait maximum performance scores.

Peer-Rated Typical Behavior

Typical performance was captured by having participants bring in a partner who was familiar with them. The partner was also awarded experimental credit. First, the peer raters' familiarity with the participants for whom they rated typical behavior was measured to ensure peers had an adequate degree of familiarity with their partners. Familiarity was measured using a knowledge-based scale, which asked questions of the partners to ensure that they had sufficient personal knowledge of the participant to accurately assess their partner's behaviors. For instance, the partner was asked to provide answers concerning the participants' middle name, favorite television shows, birthday, etc. Peer familiarity answers were compared with answers the participant had provided about the same questions. Peer ratings were only utilized from participants whose partners were able to correctly answer at least two questions. For the partners who were able to correctly answer two questions, the mean length of time the partners reported having been acquainted was 37.6 months ($SD = 57.2$). For partners who were unable to answer the two questions correctly, the mean length of time they had been acquainted was 9 months ($SD = 8.54$).

A t-test demonstrated that there was a significant difference between the two groups ($t = 4.35$, $p < .01$)

Peer raters determined to be familiar enough with their partners to provide typical performance ratings as judged by the criteria described above were presented with the identical 12 stimulus events presented in the SJT. The events were grouped by trait to make the rating exercise easier for the partner. Peers rated their partners regarding what they expected their partner would normally do in the situations depicted in simulation events using the same behaviorally anchored rating scale utilized by raters of the simulation. The peers rated the participants on a scale of one to five in order to indicate how agreeable or how extraverted their partner/participant would typically behave in such a situation. Thus, by calculating an average of the trait-based ratings provided by the peer raters for six Agreeableness events and the six Extraversion events, two overall typical performance scores were formed. The two traits were explained in detail at the beginning of the rating exercise, and reminders were presented during each event rating. For example, participants were provided the following text: "Agreeableness is a tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others. Agreeable people tend to be considerate, willing to compromise, conforming, empathetic, and loyal. They tend to be very good at getting along with others, pacifying individuals, and resolving conflict." For Agreeableness, the reliability of partner ratings assigned to the 6 events was $\alpha = .39$, and for Extraversion the reliability of partner ratings assigned to the 6 events was $\alpha = .54$. It is important to note that these reliabilities may be lower than other types of personality assessment (e.g. NEO scores) due to the fact that peer raters were asked to provide ratings of situation specific trait expression. For the peer rating scale of Agreeableness, events 5 and 6 were removed, as removal of these items increased the cronbach's alpha from .386 to .487. Peer ratings were averaged for Agreeableness and Extraversion to form two typical performance scores.

Procedure

Upon arrival in the study setting, participants first completed the Wonderlic WPT-Q in order to prevent fatigue on this cognitive measure. Participants then received a questionnaire packet that included the following measures detailed above: the demographic information questionnaire, the personality inventory, the familiarity manipulation check, and customer service experience measures (see Appendix A). The participant was then read the SJT instructions from a script, and was provided the answer sheets for the SJT. After receiving instructions and an answer sheet, participants completed the video-based SJT. While the participant completed these tasks, their partner completed the familiarity check and the typical behavior ratings. Following these exercises, the partner was dismissed and the participant completed the maximum performance test.

CHAPTER SIX: RESULTS

In order to analyze the data, correlations were calculated between the variables of interest, namely: demographic variables, GMA, personality (Extraversion and Agreeableness), SJT scores (scored using the best choice and distance score scoring techniques), typical performance, and maximum performance. A correlation table of Pearson Product-Moment Correlations was computed for these variables of interest. These correlations were initially examined in order to determine the convergent and discriminant validity of the different scoring types. The strength of the relationships between same-trait measures of different methods was then compared to determine if there were significant differences based on scoring method in the hypothesized directions. Fisher r-to-z transformations were calculated to determine if there were significant differences between the correlations to demonstrate differential validity. This was done to determine if there were significant effects of scoring method (best choice vs. distance scores) on the relationships between variables of interest and SJT scores in support of hypotheses one, two, three, and four. Regression analyses were also used to test hypotheses one, two, three, and four in order to allow for the inclusion of control variables and to test for the unique contribution of the predictors when considered as a set. Mediation for hypothesis five was tested by using correlation and regression analyses through the Baron and Kenney (1986) method. To test hypotheses six and seven, incremental validity was calculated using hierarchical-regression analyses and calculating change in r-squared to determine if there was incremental validity, or additional variance explained in typical performance, by the SJT scores.

Initial Correlation Findings

In Table 2, the means, standard deviations, and inter-correlations among study variables are displayed. First, to examine the correlations between different traits using the same

assessment method, it was found that the correlation between self-reported Agreeableness and self-reported Extraversion was $r = .20$, $p < .05$, which was similar to the findings from previous studies ($\rho = .17$) (Judge, Jackson, Shaw, Scott, & Rich, 2007). When examining the SJT distance scores for the two traits, it was found that there was a strong correlation between the Agreeableness and Extraversion distance scores ($r = .78$, $p < .01$). This demonstrates that there is a large amount of shared variance between the traits using this indexing method. It is possible that there is a general factor causing higher distance scores in both traits, and that the variance captured using this assessment method may not be exclusively related to the traits of Extraversion and Agreeableness. Consistent with this notion, both distance scores were significantly and negatively associated with cognitive ability (Agreeableness $r = -.35$, $p < .05$ Extraversion $r = -.36$, $p < .05$). When examining best choice scores, there was no relationship found between best choice scores for Agreeableness and best choice scores for Extraversion ($r = .04$, $p > .05$). When examining criteria, a weaker correlation was found between typical Agreeableness and Extraversion ratings of typical behavior ($r = .39$, $p < .01$) than of Agreeableness and Extraversion ratings of maximum performance ($r = .58$, $p < .01$). This is consistent with the notion that one's maximum performance is affected more by ability and less by personality traits – thus it is more consistent – whereas typical performance is affected by ability, personality, and differences in situational reinforcers.

Some interesting correlations were found for gender as well. For example, females tended to have high higher distance scores than males for both Agreeableness and Extraversion ($r = -.23$ and $r = -.21$, respectively). Unexpectedly, those with prior customer service experience had lower best choice scores for Extraversion ($r = -.17$, $p < .05$). This finding was interesting as it would be more typical to expect the two variables to have some degree of convergence. Another

Table 2. Means, Standard Deviations, and Inter-correlations between Study Variables for the Full Sample

Variable	M	SD	1	2	3	4	5	6	7	8
Demographic										
1. Age	20.87	5.12	1							
2. Gender	.34	.48	-.08	1						
3. C.S. Experience	24.94	44.68	.55**	-.17*	1					
Antecedents										
4. GMA	23.40	3.59	-.24**	.39**	-.05	1				
5. Self-report Agreeableness	3.62	.51	.04	-.09	.14	-.05	[.73]			
6. Self-report Extraversion	3.58	.58	-.17*	-.01	.00	.09	.20*	[.81]		
SJT scores										
7. Distance Agree	6.24	1.50	.18*	-.23**	.00	-.23*	.12	-.02	[.82]	
8. Distance Extra	6.47	1.56	.07	-.21*	-.06	-.21*	.17*	.02	.78**	[.81]
9. Best Agree	.50	.38	-.03	.12	.00	.20*	-.03	.07	.05	.03
10. Best Extra	.43	.26	-.24**	.10	-.20*	.09	-.01	.17*	-.07	.02
Criteria										
11. Maximum Agree	2.90	.72	.04	-.05	.03	-.05	-.17	-.14	.05	.04
12. Maximum Extra	2.89	.75	-.07	-.15	.14	-.06	.04	.04	.21*	.17
13. Typical Agree	3.77	.66	-.09	-.06	.14	-.09	.24*	.25*	.10	.06
14. Typical Extra	3.45	.68	-.03	-.17	.16	-.29**	.30**	.35**	.14	.15

Table 2 (continued). Means, Standard Deviations, and Inter-correlations between Study Variables for the Full Sample

Variable	9	10	11	12	13	14
Demographic						
1. Age						
2. Gender						
3. C.S. Experience						
Antecedents						
4. GMA						
5. Self-report Agreeableness						
6. Self-report Extraversion						
SJT scores						
7. Distance Agree						
8. Distance Extra						
9. Best Agree	[.45]					
10. Best Extra	.04	[.29]				
Criteria						
11. Maximum Agree	-.13	-.05	[.87]			
12. Maximum Extra	.11	.02	.58**	[.94]		
13. Typical Agree	.22*	.03	.02	.08	[.49]	
14. Typical Extra	.22*	-.04	.16	.07	.39**	[.54]

**p <.05, ** p < .0*

**Note: Gender (0=female, 1=male), Customer service experience = number of months employed in a customer service position*

Table 3. Means, Standard Deviations, and Inter-correlations between Study Variables for Caucasian participants

Variable	M	SD	1	2	3	4	5	6	7	8
Demographics										
1. Age	20.78	5.77	1							
2. Gender	.36	.49	-.08	1						
3. C.S. Experience	29.84	54.37	.62**	-.23*	1					
Antecedents										
4. GMA	24.29	3.34	-.20	.54**	-.07	1				
5. Self-report Agreeableness	3.63	.55	.01	-.10	.17	-.05	[.78]			
6. Self-report Extraversion	3.54	.56	-.25*	.12	-.03	.17	.12	[.81]		
SJT Scores										
7. Distance Agree	5.97	1.37	.13	-.31**	.04	-.35*	.18	-.01	[.79]	
8. Distance Extra	6.26	1.47	.02	-.30**	-.06	-.36*	.18	-.11	.74**	[.77]
9. Best Agree	.50	.25	-.06	-.04	.06	.05	.08	.21	.11	.07
10. Best Extra	.45	.27	-.27**	.06	-.27*	.03	.00	.13	-.11	-.01
Criteria										
11. Maximum Agree	2.92	.75	.21*	-.04	-.01	-.07	-.21	-.21	-.08	-.08
12. Maximum Extra	2.83	.69	.14	-.11	.15	-.13	.11	-.03	.12	.04
13. Typical Agree	3.62	.75	-.08	-.06	.20	.05	.26*	.36**	.09	.02
14. Typical	3.45	.70	-.06	-.20	.17	-.20	.27*	.34**	.30*	.36*

Table 3 (continued). Means, Standard Deviations, and Inter-correlations between Study Variables for Caucasian participants

Variable	9	10	11	12	13	14
Demographics						
1. Age						
2. Gender						
3. C.S. Experience						
Antecedents						
4. GMA						
5. Self-report Agreeableness						
6. Self-report Extraversion						
SJT Scores						
7. Distance Agree						
8. Distance Extra						
9. Best Agree	[.47]					
10. Best Extra	.22*	[.31]				
Criteria						
11. Maximum Agree	-.12	-.14	[.89]			
12. Maximum Extra	.00	-.07	.44**	[.89]		
13. Typical Agree	.43**	.03	.07	.13	[.58]	
14. Typical Extra	.40**	-.09	.23	.14	.50**	[.59]

*p <.05, ** p < .01

**Note: Gender (0=female, 1=male), Customer service experience = number of months employed in a customer service position*

Table 4. Means, Standard Deviations, and Inter-correlations between Study Variables for minority participants

Variable	M	SD	1	2	3	4	5	6	7	8
Demographics										
1. Age	21	3.88	1							
2. Gender	.31	.47	-.10	1						
3. C.S. Experience	20.64	24.22	.23	.00	1					
Antecedents										
4. GMA	22.29	3.42	-.27*	.20	-.16	1				
5. Self-report Agreeableness	3.60	.45	.13	-.07	.00	-.16	[.63]			
6. Self-report Extraversion	3.64	.60	.00	-.20	.14	.02	.36**	[.81]		
SJT Scores										
7. Distance Agree	6.64	1.60	.30*	-.08	.01	.01	.02	-.09	[.86]	
8. Distance Extra	6.79	1.66	.20	-.04	.00	.08	.17	.15	.81**	[.87]
9. Best Agree	.50	.52	.01	.27	-.11	.36*	-.15	-.02	.01	.01
10. Best Extra	.38	.25	-.16	.13	-.02	.10	-.02	.26	.05	.13
Criteria										
11. Maximum Agree	2.93	.68	-.34*	-.05	.03	-.11	-.07	-.00	.27	.25
12. Maximum Extra	3.03	.81	-.44*	-.23	.11	-.05	-.05	.14	.27	.30*
13. Typical Agree	3.99	.44	.02	-.05	.03	-.16	.26	-.02	-.02	.00
14. Typical Extra	3.44	.66	.13	-.12	.18	-.46**	.37*	.37*	-.08	-.13
15. Race / Af. Amer.	.29	.46	-.03	-.01	-.04	.07	-.07	-.02	.11	0
16. Race / Hispanic	.42	.50	-.18	-.19	.14	-.26	.05	.17	-.25	-.18

Table 4 (continued). Means, Standard Deviations, and Inter-correlations between Study Variables for minority participants

Variable	9	10	11	12	13	14	15	16
Demographics								
1. Age								
2. Gender								
3. C.S. Experience								
Antecedents								
4. GMA								
5. Self-report Agreeableness								
6. Self-report Extraversion								
SJT scores								
7. Distance Agree								
8. Distance Extra								
9. Best Agree	[.29]							
10. Best Extra	-.09	[.21]						
Criteria								
11. Maximum Agree	-.16	.13	[.88]					
12. Maximum Extra	-.20	.21	.75**	[.97]				
13. Typical Agree	-.12	.16	-.21	-.01	[-.11]			
14. Typical Extra	-.06	.05	-.14	.01	.18	[.45]		
15. Race / Af. Amer.	.23	-.04	.02	.17	.09	.11	1	
16. Race / Hispanic	-.11	.01	-.04	.20	-.03	-.02	0	1

*p <.05, ** p < .01

**Note: Gender (0=female, 1=male), Race/ Af. Amer. (1 = member of race, 0 = non-member of race), Race/ Hispanic (1 = member of race, 0 = non-member of race). Customer service experience = number of months employed in a customer service position*

unexpected finding was a negative correlation between typical performance Extraversion and GMA (r = -.29, p < .01). It is also important to note that when considering typical performance ratings, there was no significant relationship between level of familiarity between partners and typical performance ratings (Agreeableness r = -.07, Extraversion = .11) demonstrating that there is no confound, and that familiar partners do not necessarily inflate partner ratings.

Differences for Race and Gender

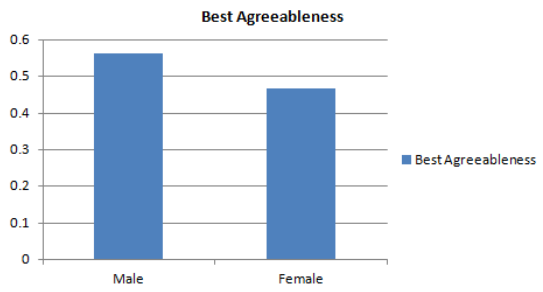
Mean Differences

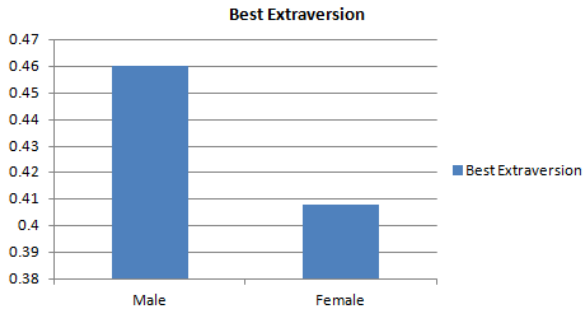
Mean differences were explored for gender and race with regard to GMA, self-reported personality, SJT scores and criteria. First, t-tests were conducted to determine if there were mean differences for gender. Results demonstrated a mean difference between males and female with regard to distance scores of Agreeableness favoring females (females *mean* = 6.47, males *mean* = 5.64, t = 2.37, p > .05) (see figure 6). A t-test was conducted to determine if there were differences in typical performance ratings received by Caucasian and minority participants. Results demonstrated that minorities tended to have higher typical performance ratings of Agreeableness (Caucasian *mean* = 3.63, minority *mean* = 3.99, t = -2.55, p < .01). However, there was no difference found for typical performance Extraversion (Caucasian *mean* = 3.45, minority *mean* = 3.44, t = .03, p > .05). Additionally, no differences were found between Caucasians and minorities in maximum performance scores (Agreeableness: Caucasian *mean* = 2.92, minority *mean* = 2.93, t = -.11, p > .05; Extraversion: Caucasian *mean* = 2.83, minority *mean* = 3.03, t = -1.31, p > .05). ANOVAs were calculated to determine if there were any differences in the typical performance ratings assigned as a function of participant race and gender. Results indicated gender and race interacted in the prediction of typical performance Extraversion (f = 4.15, p < .05). However, there was not a significant interaction for typical

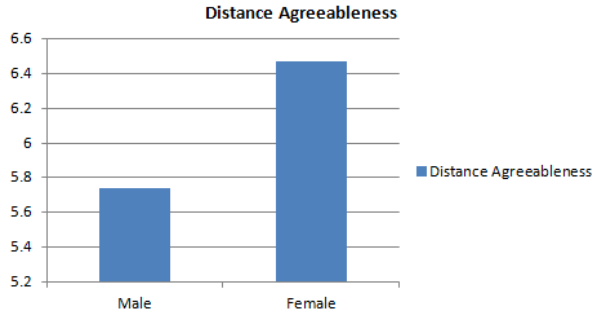
performance Agreeableness (f = .16, p > .05). A summary of these results is presented in figure 6.

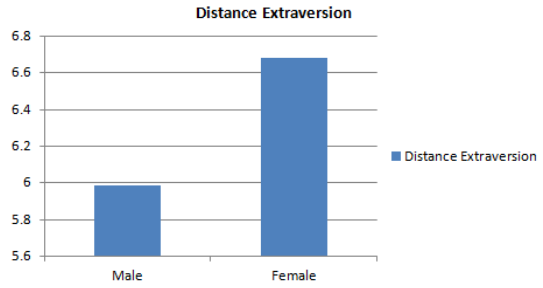
Finally, ANOVAs were conducted to determine if there were any mean differences in the typical performance ratings provided to participants by same- or different-race partners (see table 5). Data was coded as to reflect if the partner was being rated by an individual of the same race or an individual of a difference race. Results demonstrated that there were no significant differences in typical performance ratings received by raters of the same or a different race on Agreeableness (Caucasian n = 54, African American n = 11, Hispanic n = 19; Caucasian f = .43, p > .05, African Americans f = 4.09, p > .05, Hispanic participants f = 1.38, p > .05). Additionally, no significant differences were found in typical performance ratings of Extraversion assigned by same or different race peers (Caucasian, f = 1.23, p > .05; African American f = .46, p > .05, Hispanic f = .71, p > .05). T-tests were conducted to determine if there were any mean differences in ratings of typical performance for dyads in which the participant and the partner were the same gender or different genders. Results indicated there were no significant differences (Agreeableness t = .39, p > .05; Extraversion t = .46, p > .05). In sum, these results suggest that mean differences for typical performance ratings varied as a function of gender and race. However, no significant differences were found in the level of typical performance ratings as a function of respondent and peer-rater similarity with respect to gender or race.

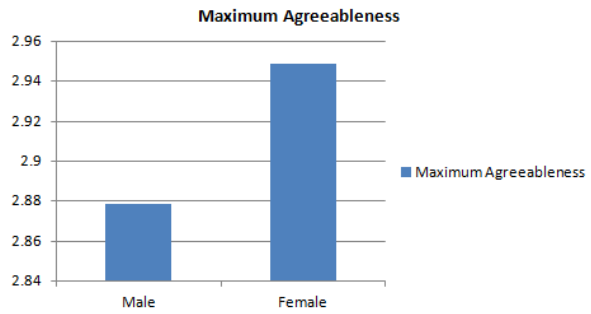
(See Figure 6)

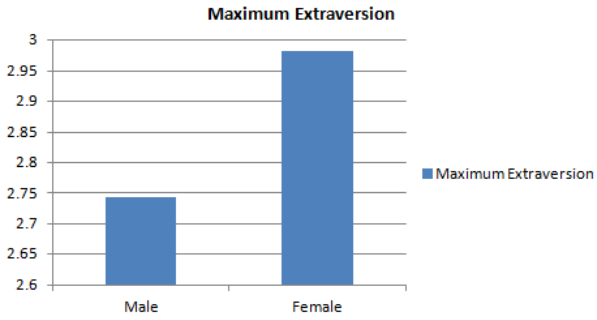

Best Agreeableness
0.6
0.5
0.4
0.3
0.2
0.1
0
Male
Female
Best Agreeableness

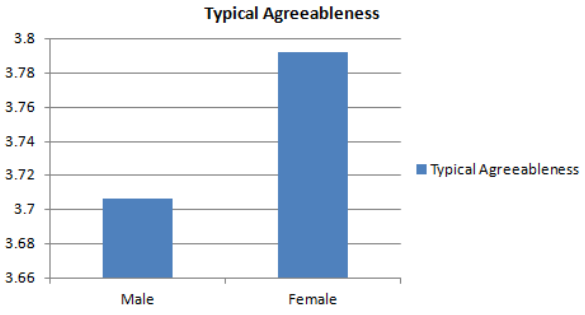

Best Extraversion
0.47
0.46
0.45
0.44
0.43
0.42
0.41
0.4
0.39
0.38
Male
Female
Best Extraversion


Distance Agreeableness
6.6
6.4
6.2
6
5.8
5.6
5.4
5.2
Male
Female
Distance Agreeableness


Distance Extraversion
6.8
6.6
6.4
6.2
6
5.8
5.6
Male
Female
Distance Extraversion


Maximum Agreeableness
2.96
2.94
2.92
2.9
2.88
2.86
2.84
Male
Female
Maximum Agreeableness


Maximum Extraversion
3
2.95
2.9
2.85
2.8
2.75
2.7
2.65
2.6
Male
Female
Maximum Extraversion


Typical Agreeableness
3.8
3.78
3.76
3.74
3.72
3.7
3.68
3.66
Male
Female
Typical Agreeableness

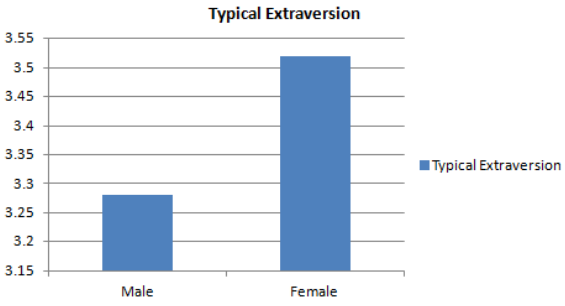

Typical Extraversion
3.55
3.5
3.45
3.4
3.35
3.3
3.25
3.2
3.15
Male
Female
Typical Extraversion

Figure 6 (continued)

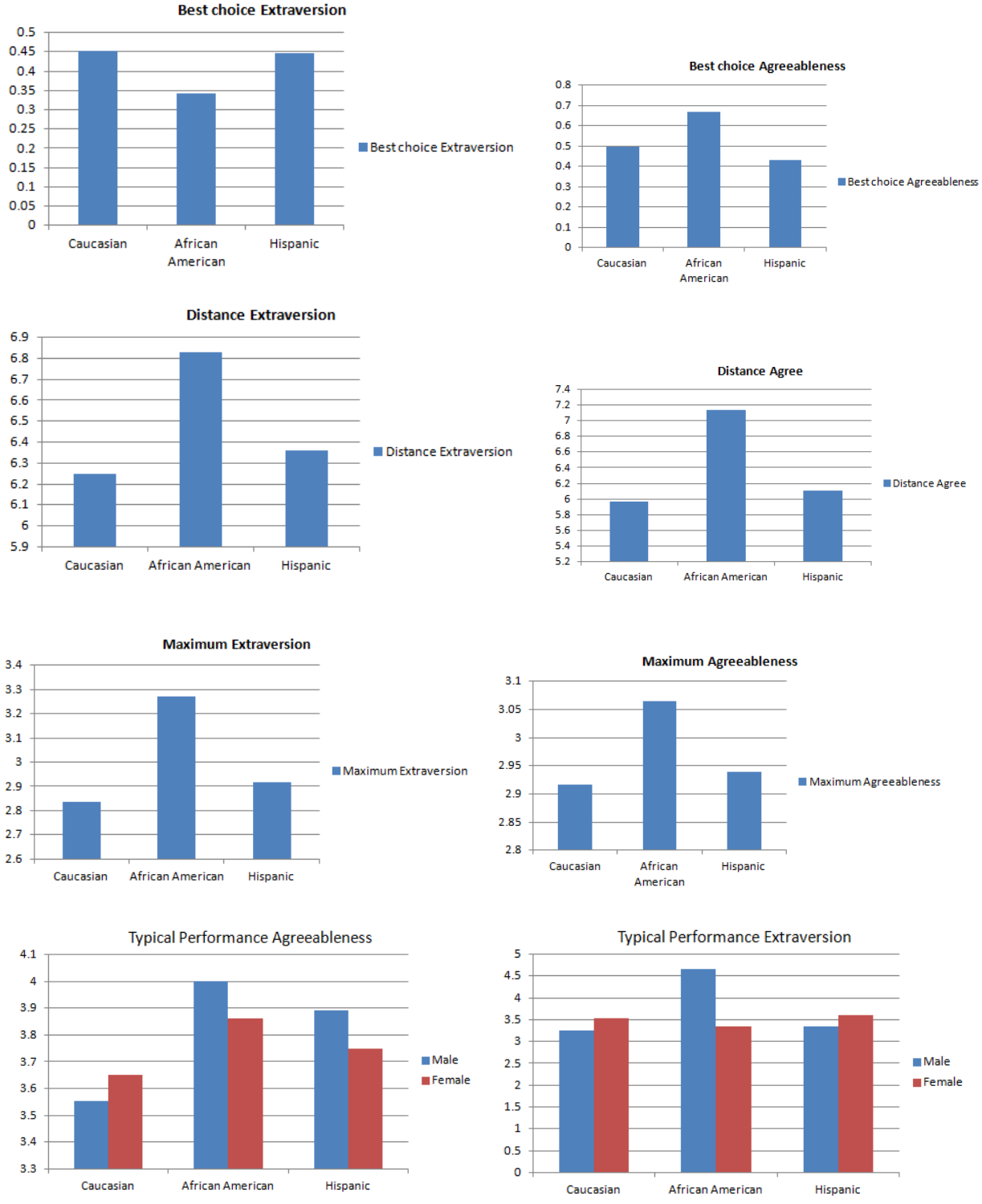


Figure 6 (continued)

Figure 6. Bar graphs demonstrating mean differences for race and gender on SJT scores and criteria.

Table 5. Tables demonstrating effects of racial similarity of partner on typical behavior ratings

ANOVA Best choice Agreeableness

	df	*F*	*Ƞ*	*p*
Corrected Model	2	1.56	.24	.22
Intercept	1	114.59	17.36	0
Race	2	1.56	.24	.22

Note: Race coded (1=Caucasian, 2=African American, 3=Hispanics)

ANOVA Best choice Extraversion

	df	*F*	*Ƞ*	*p*
Corrected Model	2	.54	.04	.59
Intercept	1	154.35	10.46	.00
Race	2	.54	.34	.59

Note: Race coded (1=Caucasian, 2=African American, 3=Hispanics)

ANOVA Distance Agreeableness

	df	*F*	*Ƞ*	*p*
Corrected Model	2	3.873	7.27	.02
Intercept	1	1338.79	2513.87	.00
Race	2	3.87	7.27	.02

Note: Race coded (1=Caucasian, 2=African American, 3=Hispanics)

Table 5 (continued). Tables demonstrating effects of racial similarity of partner on typical behavior ratings

ANOVA Distance Extraversion

	df	*F*	*η*	*p*
Corrected Model	2	.48	1.14	.62
Intercept	1	1087.82	2574.44	.00
Race	2	.48	1.14	.62

Note: Race coded (1=Caucasian, 2=African American, 3=Hispanics)

ANOVA Maximum performance Agreeableness

	df	*F*	*Ƞ*	*p*
Corrected Model	2	2.95	.157	.746
Intercept	1	835.34	446.47	.00
Race	2	.295	.157	.746

Note: Race coded (1=Caucasian, 2=African American, 3=Hispanics)

ANOVA Maximum performance Extraversion

	df	*F*	*η*	*p*
Corrected Model	2	2.30	1.10	.11
Intercept	1	983.75	469.78	.00
Race	2	2.30	1.10	.11

Note: Race coded (1=Caucasian, 2=African American, 3=Hispanics)

Table 5 (continued). Tables demonstrating effects of racial similarity of partner on typical behavior ratings

ANOVA Typical performance Agreeableness

	df	*F*	*Ƞ*	*p*
Corrected Model	2	1.27	.58	.29
Intercept	1	1617.41	736.51	.00
Race	2	1.27	.58	.29

Note: Race coded (1=Caucasian, 2=African American, 3=Hispanics)

ANOVA Typical performance Extraversion

	df	*F*	*η*	*p*
Corrected Model	2	.647	.31	.53
Intercept	1	1341.19	641.14	.00
Race	2	.65	.31	.53

Note: Race coded (1=Caucasian, 2=African American, 3=Hispanics)

ANOVA Typical performance Agreeableness Race and Gender

	df	*F*	*η*	*p*
Corrected Model	5	.61	.29	.69
Intercept	1	1142.93	540.90	.00
Race	2	1.456	.69	.24
Gender	1	.02	.01	.90
Race * Gender	2	.160	.08	.85

Note: Gender coded (0 = female, 1 = male), Race coded (1=Caucasian, 2=African American, 3=Hispanics)

Table 5 (continued). Tables demonstrating effects of racial similarity of partner on typical behavior ratings

ANOVA Typical performance Extraversion Race and Gender

	df	*F*	*η*	*p*
Corrected Model	5	2.13	.95	.07
Intercept	1	1111.34	494.90	.00
Race	2	3.01	1.34	.06
Gender	1	1.29	.57	.26
Gender * Race	2	4.15	1.85	.02

Note: Gender coded (0 = female, 1 = male), Race coded (1=Caucasian, 2=African American, 3=Hispanics)

ANOVA Typical performance Agreeableness (Matched/Unmatched Race)

	df	*F*	*η*	*p*
Caucasian	53	.43	.09	.95
African American	10	4.09	.41	.07
Hispanic	18	1.38	.30	.30

Note: Race coded (1=matched, 0=unmatched)

ANOVA Typical performance Extraversion (Matched/Unmatched Race)

	df	*F*	*η*	*p*
Caucasian	53	1.23	.21	.30
African American	10	.46	.46	.83
Hispanic	18	.71	.71	.72

Note: Race coded (1=matched, 0=unmatched)

Differential Validity

As adverse impact is of particular importance in assessment, additional tests were conducted to determine if there were any interactions between study variables and race. Differential validity, or differences in validity coefficients between a predictor and a criteria between subgroups (Linn,1978), was calculated. Differential validity is related to another concept called differential prediction. Differential predition focuses on differences between regression slopes and intercepts relating the test and criterion across subgroups (Berry, Clark & McClure, 2011). Differential predictions provides more information than differential validity, such as slope and intercept scores. However, differential validity is still important to examine, as he examination of both differential prediction and differential validity provide unique information (Linn, 1978). In order to determine whether there were significant differential results for race, preliminary examinations were conducted by computing additional correlation tables to compare Caucasian participants and minority participants. Correlation tables were computed for the Caucasian subset and the minority subset. Results demonstrated several significant differences in validity coefficients for Caucasians and minorities (see tables 3 and 4). First, to compare the correlations between constructs of interest separately for participants of different races, see table 6. These significant differences were tested with Fisher's r-to-z transformations. For example, when examining differential relationships between gender SJT score, it was found that distance scores were larger for Caucasian females (Agreeableness r = -.31, Extraversion r = -.30) but not for minority females (Agreeableness r = -.08, Extraversion r = -.04) (Agreeableness *z-score* = 1.67, $p < .05$; Extraversion *z-score* 1.88, $p < .05$). Interestingly, when examining the relationships between gender and GMA, it was found that there was a very strong relationship favoring males for the Caucasian population (r = .54, $p < .01$) and a much weaker relationship between gender and GMA for the

minority population (r = .20, p > .05) (z-*score* = 2.8, p < .01). Also, it was found that there was a moderate negative relationship between GMA and distance scores for Caucasians (Agreeableness distance score r = -.35, p < .05; Extraversion distance score r = -.36, p < .05) but very small positive relationship between GMA and distance scores for minorities (Agreeableness distance score r = .01, Extraversion distance score r = .08) (Agreeableness z-*score* = 2.61, p > .01, Extraversion z-*score* = 3.18, p > .01). Conversely, there was a strong negative relationship between GMA scores and partner rated Extraversion for minorities (r = -.46, p < .01). However, this relationship was not significant for Caucasians (r = -.20, p > .05) (z-*score* = 4.88, p < .01). Also for Caucasians it was found that there was a moderate significant relationship between distance scores and typical Extraversion (r = .36, p < .05). However, for minorities a small negative relationship was found between distance scores for Extraversion and typical Extraversion (r = -.08, p > .05) (z-*score* = 3.18, p < .01). Finally, for Caucasians a significant relationship was found between best choice scores for Agreeableness and typical performance ratings of Agreeableness (r = .43, p < .01). However, for minorities, a small negative relationship was found between best choice Agreeableness scores and typical performance ratings of Agreeableness (r = -.12, p > .05) (z-*score* = 4.04, p < .01). These differences highlight that the race of an individual can drastically affect the validity of an SJT scoring method.

Table 6. Tables Demonstrating Differential Correlations and Validity Coefficients

Correlations between member/nonmember of race and variables

Variable	N	GMA	Ag	Ex	Distance Agree	Distance Ex	Best Agree	Best Ex	Typical Agree	Typical Extra	Max. Agree	Max. Extra
White	64	.28*	.04	-.08	-.22*	-.17*	-.01	.12	-.26*	.00	-.01	-.13
African American	13	-.14	-.04	-.01	.23*	.09	.17*	-.12	.07	.08	.07	.17
Hispanic	22	-.19*	.06	.15	-.05	-.04	-.09	.04	.02	.06	.01	.01

**Relationships significant (two tailed) p<.05*
***Relationships significant (two tailed) p<.01*
Note: races coded (1=member of race, 0=not a member of the race)

Validity Coefficients for best choice Agreeableness

Variable	N	GMA	Ag	Ex	Typical Agree	Typical Extra	Max. Agree	Max. Extra
White	64	.05	.08	.21	.43*	.41*	-.21	0
African American	13	.56*	-.03	.06	-.14	-.21	-.37	-.61*
Hispanic	22	-.21	-.11	-.14	-.20	-.02	0	.02

**Relationships significant (two tailed) p<.05*
***Relationships significant (two tailed) p<.01*

Table 6 (continued). Tables Demonstrating Differential Correlations and Validity Coefficients

Validity Coefficients for best choice Extraversion

Variable	N	GMA	Ag	Ex	Typical Agree	Typical Extra	Max. Agree	Max. Extra
White	64	.03	0	.13	.03	-.09	-.14	-.07
African American	13	-.07	-.02	.33	.39	.32	-.36	-.29
Hispanic	22	.24	.08	.05	-.02	-.08	-.06	-.04

**Relationships significant (two tailed) p<.05*

***Relationships significant (two tailed) p<.01*

Validity Coefficients for distance Agreeableness

Variable	N	GMA	Ag	Ex	Typical Agree	Typical Extra	Max. Agree	Max. Extra
White	64	-.35	.18	-.01	.09	.30*	-.08	.12
African American	13	-.32	.01	-.25	-.35	-.11	.16	.58*
Hispanic	22	-.26	-.24	-.03	-.25	-.39	.29	.65**

**Relationships significant (two tailed) p<.05*

***Relationships significant (two tailed) p<.01*

Validity Coefficients for distance Extraversion

Variable	N	GMA	Ag	Ex	Typical Agree	Typical Extra	Max. Agree	Max. Extra
White	64	-.36**	.18	-.11	.02	.36**	-.08	.04
African American	13	-.20	.33	.45	-.31	.08	.16	.25
Hispanic	22	-.13	-.21	.11	-.04	-.34	.23	.61

**Relationships significant (two tailed) p<.05*

***Relationships significant (two tailed) p<.01*

Differential Prediction

Due to the number of differences found between the validity coefficients for Caucasians and minorities, further analyses were conducted in order to determine if differential prediction existed for race. As already discussed, there were differences between Caucasians and minorities with regard to relationships between GMA and SJT scores, and also relationships between SJT scores and maximum and typical performance criteria. Table 5 summarizes some differential prediction between African Americans and Hispanics. However, as the numbers of African Americans and Hispanics are inadequate to explore differential prediction separately, analyses were conducted to compare Caucasians to all other minorities.

Regression analyses were conducted to determine if there was differential prediction by race for antecedents predicting SJT scores. These results are summarized in table 7. Results indicate that there was a significant interaction between race and GMA in predicting best choice Agreeableness scores (β = -1.70, p > .05). Specifically, the results demonstrated that there was a negative relationship between GMA and best choice scores for Caucasians, but there was no relationship for minorities (see figure 7). Similar results were found for predicting typical performance Agreeableness (β = -1.70, p > .05) (see figure 8).

Regression analyses were conducted to determine if differential prediction existed for the subgroups of race or gender. The results of these regression analyses are summarized in table 8. A significant interaction was found for best choice scores of Extraversion and race (β = -.57, p < .01). For these scores, it was found that best choice scores were positively related to maximum performance of extraversion for minorities, but the regression coefficient for Caucasians was negative (see figure 9). Also, for maximum performance Extraversion, a significant interaction term was found between distance scores for Extraversion and race (β = -1.82, p < .01). Again, it

was found that distance scores were positively related to maximum performance for minorities, but the regression coefficient for Caucasians was negative (see figure 10).

A significant interaction term was found for best choice Agreeableness and race ($\beta = .63$, $p < .05$) as predictors of typical performance Agreeableness scores. Specifically, it was found that best choice Agreeableness scores were better predictors of typical performance Agreeableness scores for Caucasians than for minority participants (see figure 11).

A significant interaction was found for distance scores of Extraversion and race ($\beta = 2.66$, $p < .05$) as predictors of typical Extraversion. Specifically, it was found that distance scores of Extraversion were positively related to typical performance for Caucasians, and in fact the regression coefficient for the minority group was negative (see figure 12).

In sum, SJT scores were better predictors of maximum performance for minorities than for Caucasians. However, SJT scores were better predictors of typical performance for Caucasians relative to minorities. Additionally, GMA was a stronger predictor of best choice scores for Caucasians relative to minorities. These results shall be explored later in this paper.

Table 7. Regression Analyses demonstrating Race and Gender Interactions for antecedents

Predictors of best choice Agreeableness

Variable	B	β	95% C	
Constant	-.33		[-1.77,	1.12]
Race / Caucasian	.50	.62	[-1.38,	2.38]
GMA	.06	.50*	[.018,	.10]
Self-report Agreeableness	-.10	-.13	[-.41,	.21]
Self-report Extraversion	-.02	-.02	[-.24,	.20]
Inter. Race and GMA	-.06	-1.70*	[-.11,	-.01]
Inter. Race and Agreeableness	.08	.39	[-.28,	.45]
Inter Race and Extraversion	.11	.51	[-.18,	.40]
R^2	0.11			
F	1.48			

Notes. N=90. CI = Confidence

*p <.05, ** p < .01

Predictors of best choice Extraversion

Variable	B	β	95% CI	
Constant	.06		[-.94,	1.06]
Race / Caucasian	.14	.26	[-1.13,	1.41]
GMA	.01	.08	[-.02,	.03]
Self-report Agreeableness	-.05	-.09	[-.26,	.17]
Extraversion	.10	.23	[-.05,	.26]
Inter. Race and GMA	-.01	-.25	[-.04,	.03]
Inter. Race and Agreeableness	.03	.24	[-.22,	.29]
Inter Race and Extraversion	-.02	-.14	[-.22,	.18]
R^2	0.06			
F	0.71			

Notes. N=90. CI = Confidence Interval.

*p <.05, ** p < .01

Table 7 (continued). Regression Analyses demonstrating Race and Gender Interactions for antecedents

Predictors of distance Agreeableness

Variable	B	β	95% CI	
Constant	7.18		[1.88,	12.48]
Race / Caucasian	.98	.33	[-5.94,	7.90]
GMA	.01	.02	[-.14,	.15]
Self-report Agreeableness	-.10	-.03	[-1.25,	1.05]
Self-report Extraversion	-.12	-.05	[-.93,	.69]
Inter. Race and GMA	-.15	-1.27	[-.34,	.04]
Inter. Race and Agreeableness	.32	.41	[-1.04,	1.69]
Inter Race and Extraversion	.27	.34	[-.81,	1.35]
R^2	0.11			
F	1.48			

Notes. N=91. CI = Confidence

*p <.05, ** p < .01

Predictors of distance Extraversion

Variable	B	β	95% CI	
Constant	3.39		[2.12,	8.90]
Race / Caucasian	5.18	1.68	[-1.95,	12.30]
GMA	.04	.08	[-.11,	.19]
Self-report Agreeableness	.34	.11	[-.86,	1.53]
Self-report Extraversion	.35	.13	[-.50,	1.19]
Inter. Race and GMA	-.18	-1.44*	[-.37,	.01]
Inter. Race and Agreeableness	.05	.06	[-1.37,	1.47]
Inter Race and Extraversion	-.44	-.52	[-1.55,	.68]
R^2	0.12			
F	1.68			

Notes. N=90. CI = Confidence Interval.

*p <.05, ** p < .01

Table 8. Regression Analyses demonstrating Race and Gender Interactions for criteria

Predictors of maximum performance Agreeableness

Variable	B	β	95% CI	
Constant	2.20		[.75,	3.65]
Race / Caucasian	1.42	.97	[-.63,	3.48]
Age	.01	.04	[-.04,	.05]
C.S. Experience	.00	.01	[-.01,	.01]
Gender	-.63	-.42	[-2.67,	1.41]
Distance Agree	.13	.26	[-.05,	.30]
Best Choice Agree	-.61	-.32	[-1.70,	.47]
Inter. Race and Distance Agree	-.22	-.97	[-.52,	.08]
Inter. Race and Best Choice Agree	.01	.02	[-1.11,	1.21]
Inter. Gender and Distance Agree	.06	.22	[-.28,	.39]
Inter. Gender and Best Choice Agree	.53	.30	[-.66,	1.73]
R^2	0.06			
F	.47			

Notes. N=79. CI = Confidence Interval. Race/Caucasian (1=Caucasian, 0=minority). Gender (0=female, 1=male)

*p <.05, ** p < .01

Predictors of typical performance Agreeableness

Variable	B	β	95% CI	
Constant	4.50		[3.41,	5.58]
Race / Caucasian	-1.24	-.94*	[-2.43,	.04]
Age	-.04	-.31*	[-.07,	0]
C.S. Experience	.01	.40*	[0,	.01]
Gender	.54	.39	[-.80,	1.87]
Distance Agree	.03	-.08	[-.18,	.11]
Best Choice Agree	-.115	-.05	[-1.02,	.79]
Inter. Race and Distance Agree	.04	.17	[-.15,	.22]
Inter. Race and Best Choice Agree	1.28	.63*	[.19,	2.36]
Inter. Gender and Distance Agree	-.07	-.28	[-.27,	.14]
Inter. Gender and Best Choice Agree	-.11	-.56	[-1.22,	1]
R^2	0.33			
F	3.36			

Notes. N=79. CI = Confidence Interval. Race/Caucasian (1=Caucasian, 0=minority). Gender (0=female, 1=male)

*p <.05, ** p < .01

Table 8 (continued). Regression Analyses demonstrating Race and Gender Interactions for criteria

Predictors of maximum performance Extraversion

Variable	B	β	95% CI	
Constant	2.28		[.89,	3.69]
Race / Caucasian	3.20	2.08*	[.86,	5.55]
Age	-.02	-.17	[-.06,	.01]
C.S. Experience	.00	.00	[.00,	.01]
Gender	-3.56	-2.27*	[-5.96,	-1.15]
Distance Extra	.15	.31*	[-.01,	.30]
Best Choice Extra	.55	.19	[-.53,	1.62]
Inter. Race and Distance Extra	-.41	.15*	[-.72,	-.11]
Inter. Race and Best Choice Extra	-1.40	-.57*	[-2.91,	.11]
Inter. Gender and Distance Extra	.48	1.84*	[.13,	.83]
Inter. Gender and Best Choice Extra	.95	.72	[-.48,	2.39]
R^2	0.21			
F	2.10			

Notes. N=88. CI = Confidence Interval. Race/Caucasian (1=Caucasian, 0=minority). Gender (0=female, 1=male)

*p <.05, ** p < .01

Predictors of typical performance Extraversion

Variable	B	β	95% CI	
Constant	4.99		[3.67,	6.31]
Race / Caucasian	-4.02	-2.88**	[-6.22,	-1.82]
Age	-.04	-.29*	[-.07,	.00]
C.S. Experience	.01	.47**	[.00,	.01]
Gender	2.78	1.96*	[.53,	5.03]
Distance Extra	-.10	.07	[-.24,	.05]
Best Choice Extra	-.24	-.09	[1.25,	.77]
Inter. Race and Distance Extra	.55	2.66**	[.26,	.83]
Inter. Race and Best Choice Extra	-78	.35	[.63,	2.20]
Inter. Gender and Distance Extra	-.42	-1.76*	[-.75,	-.08]
Inter. Gender and Best Choice Extra	-.66	-.25	[-2.00,	.68]
R^2	0.24			
F	2.26			

Notes. N=80. CI = Confidence Interval. Race/Caucasian (1=Caucasian, 0=minority). Gender (0=female, 1=male)

*p <.05, ** p < .01

Table 8 (continued). Regression Analyses demonstrating Race and Gender Interactions for criteria

Predictors of best choice scores Agreeableness

Variable	B	β	95% CI	
Constant	.01		[-1.53,	1.56]
Age	.01	.07	[-.02,	.03]
C.S. Experience	0	0	[0,	0]
Gender	-.52	-.65	[-2.74,	1.71]
GMA	0	.04	[-.03,	.05]
Self-Report Agreeableness	0	0	[-.21,	.22]
Self-Report Extraversion	.03	.05	[-.16,	.23]
Inter. Gender and GMA	.04	1.21	[-.02,	.10]
Inter. Race and GMA	0	-.09	[-.01,	.01]
Inter. Gender and Self-Reported Agree	-.04	-.17	[-.39,	.32]
Inter. Gender and Self-Reported Extra	-.06	-.29	[-.39,	.27]
R^2	0.08			
F	0.67			

Notes. N=89. CI = Confidence Interval. Race/Caucasian (1=Caucasian, 0=minority). Gender (0=female, 1=male)

*p <.05, ** p < .01

Predictors of best choice scores Extraversion

Variable	B	β	95% CI	
Constant	.20		[-.82,	1.21]
Age	-.01	-.11	[-.02,	.01]
C.S. Experience	0	-.13	[0,	0]
Gender	.48	.88	[-.98,	1.94]
GMA	0	.04	[-.02,	.03]
Self-Report Agreeableness	-.07	-.13	[-.21,	.08]
Self-Report Extraversion	.13	.29*	[.01,	.26]
Inter. Gender and GMA	-.02	-.83	[-.06,	.02]
Inter. Race and GMA	0	.19	[0,	.01]
Inter. Gender and Self-Reported Agree	.12	.81	[-.11,	.36]
Inter. Gender and Self-Reported Extra	-.13	-.86	[-.35,	.09]
R^2	0.14			
F	1.27			

Notes. N=80. CI = Confidence Interval. Race/Caucasian (1=Caucasian, 0=minority). Gender (0=female, 1=male)

*p <.05, ** p < .01

Table 8 (continued). Regression Analyses demonstrating Race and Gender Interactions for criteria

Predictors of distance scores Agreeableness

Variable	B	β	95% CI
Constant	1.635		[-4.85, 8.12]
Age	.10	.20	[-.01, .22]
C.S. Experience	-.01	-.21	[-.03, 0]
Gender	3.03	.98	[-4.22 10.29]
GMA	.09	.21	[-.09, .27]
Race / Caucasian	2.40	.82	[-2.21, 7.01]
Self-Report Agreeableness	.46	.16	[-.40, 1.31]
Self-Report Extraversion	-.15	-.06	[-.83, .53]
Inter. Gender and GMA	-.14	-1.18	[-.36, .07]
Inter. Race and GMA	-.12	-.99	[-.32, .08]
Inter. Gender and Self-Reported Agree	-.54	-.64	[-1.82, .74]
Inter. Gender and Self-Reported Extra	.52	.62	[-.60, 1.64]
R^2	0.17		
F	1.44		

Notes. N=89. CI = Confidence Interval. Race/Caucasian (1=Caucasian, 0=minority). Gender (0=female, 1=male)

*p <.05, ** p < .01

Predictors of distance scores Extraversion

Variable	B	β	95% CI
Constant	4.07		[-1.98, 10.13]
Age	.02	.05	[-.07, .10]
C.S. Experience	0	-.14	[-.01, 0]
Gender	1.90	.58	[-6.82 10.63]
GMA	.03	.06	[-.12, .17]
Self-Report Agreeableness	.65	.21	[-.21, 1.50]
Self-Report Extraversion	-.07	-.02	[-.82, .69]
Inter. Gender and GMA	-.11	-.85	[-.34, .12]
Inter. Race and GMA	-.02	-.19	[-.06, .01]
Inter. Gender and Self-Reported Agree	-.24	-.70	[-1.63, 1.15]
Inter. Gender and Self-Reported Extra	.29	.32	[-1.00, 1.58]
R^2	0.15		
F	1.42		

Notes. N=89. CI = Confidence Interval. Race/Caucasian (1=Caucasian, 0=minority). Gender (0=female, 1=male)

*p <.05, ** p < .01

(*see* Figure 7)

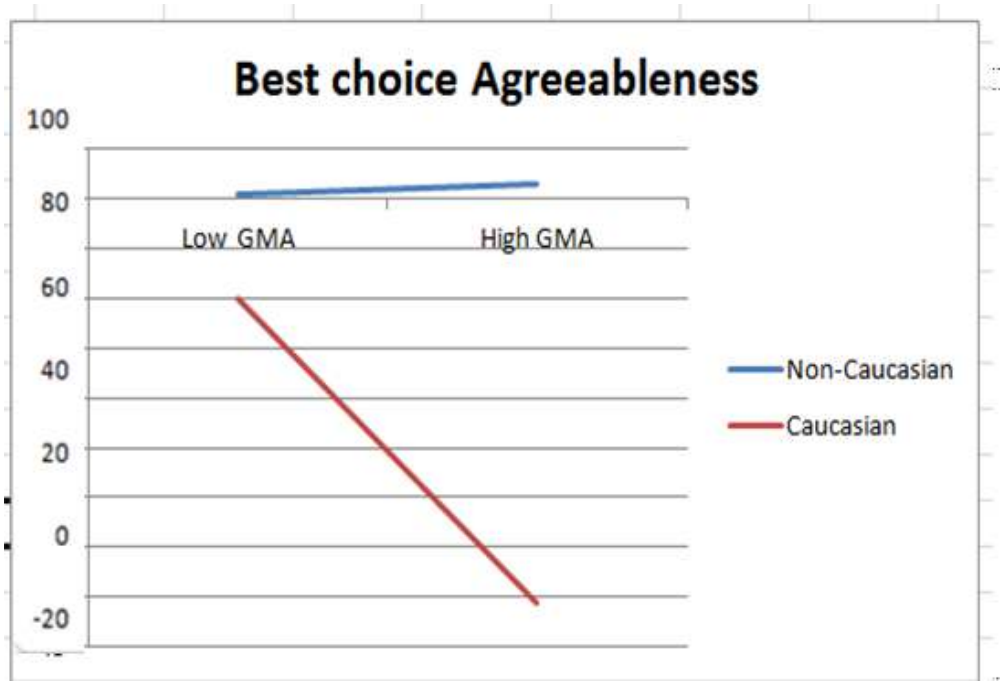


Figure 7. Graph of best choice Agreeableness regressed onto GMA moderated by race

(*see* Figure 8)

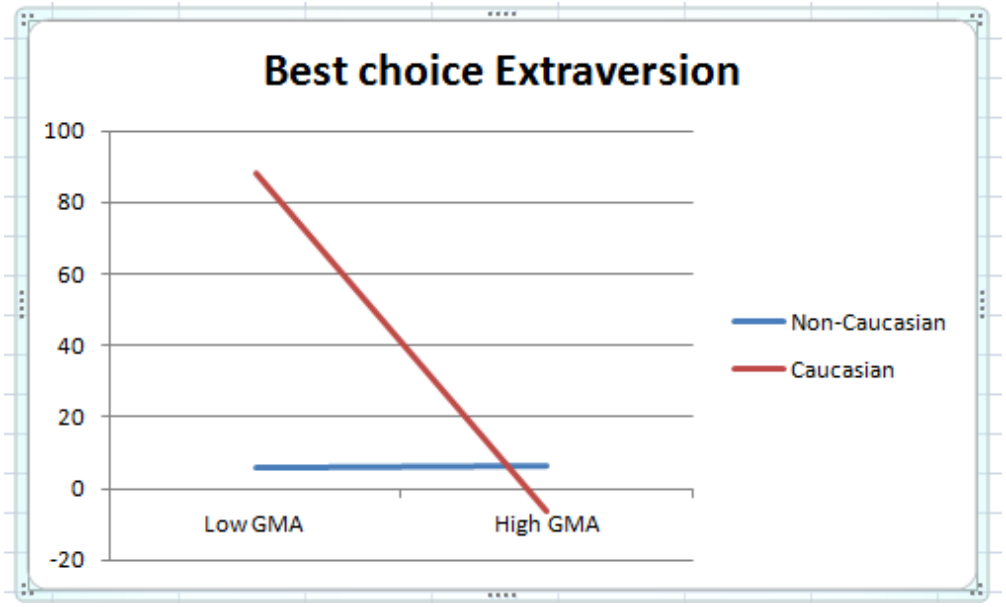


Figure 8. Graph of best choice Extraversion regressed onto GMA moderated by race

(see Figure 9)

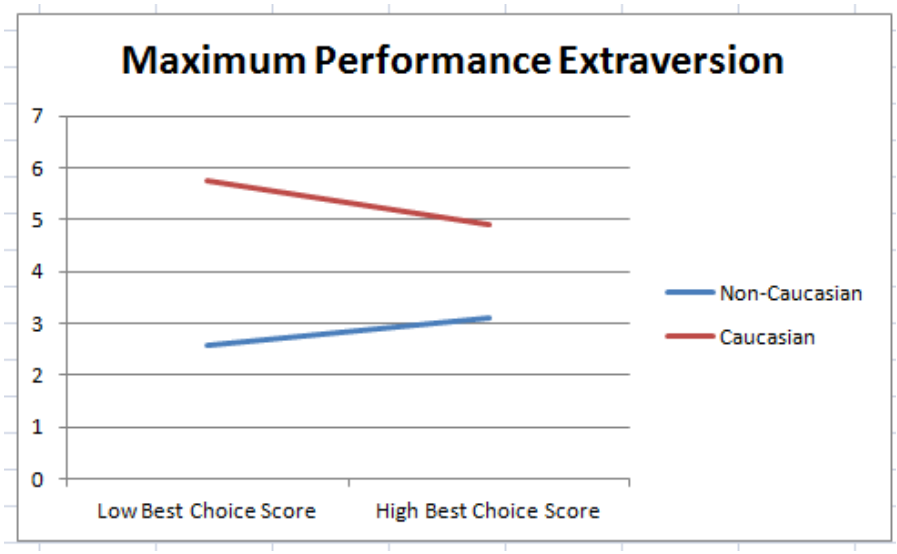


Figure 9. Graph of maximum performance Extraversion regressed onto best choice scores moderated by race

(*see* Figure 10)

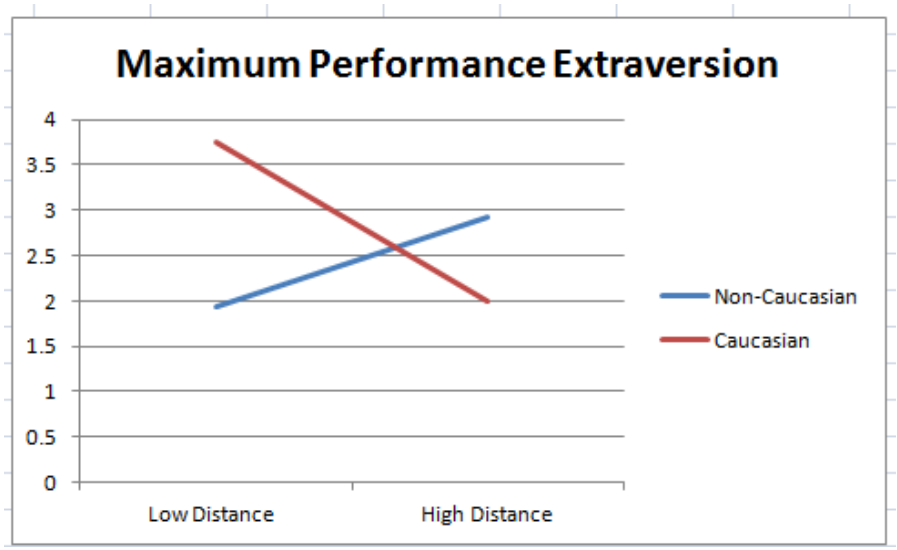


Figure 10. Graph of maximum performance Extraversion regressed onto distance scores moderated by race

(See Figure 11)

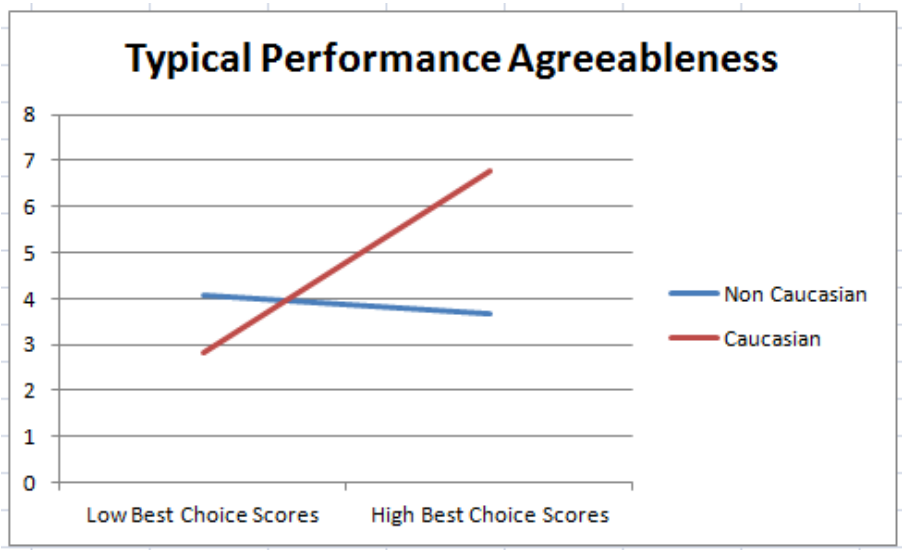


Figure 11. Graph of typical performance Agreeableness regressed onto best choice scores moderated by race

(*see* Figure 12)

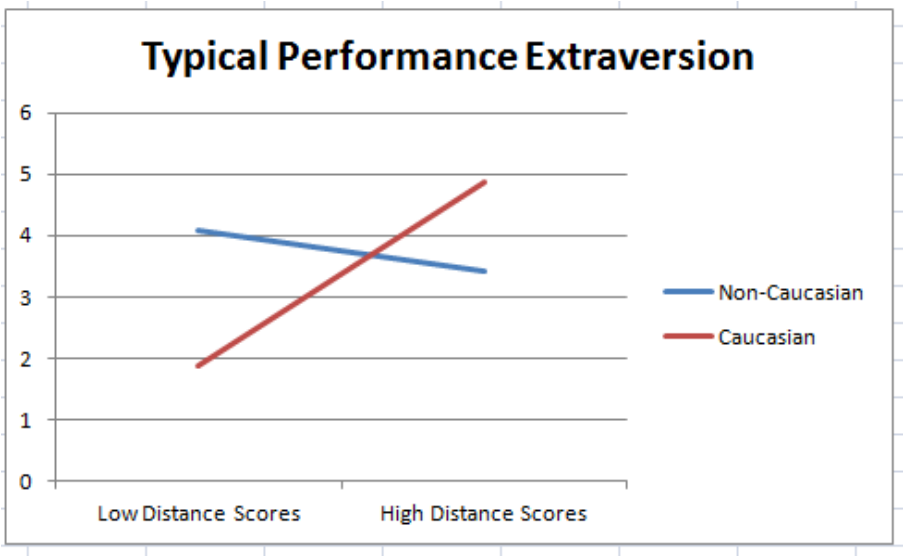


Figure 12. Graph of typical performance Extraversion regressed onto distance scores moderated by race

Hypothesis Testing

Due to the strong results found for differential prediction based on race, hypotheses were tested using the Caucasian population only. This was due to the fact that differential prediction caused stronger relationships for Caucasians for several of the indices. Additionally, there were differential effects found for African Americans and Hispanics; however, these results could not be explored separately due to the small number of participants in these subsets. Because of this, the data set for the Caucasian population only was used in all subsequent analyses. Hypothesis tests will be discussed sequentially - in the order in which they were proposed in the theoretical section.

Hypothesis One

Hypothesis one stated that self-reported personality measures would be more strongly related to distance scores derived from an SJT than to best choice scores. Thus, hypothesis one was tested by comparing the correlations in a matched-trait fashion (e.g. distance Agreeableness score's correlation with self-report Agreeableness to best choice Agreeableness score's correlation with self-report Agreeableness). Thus, Fisher's r-to-z transformations were calculated in order to determine if there were significant differences between the two correlations in the correct direction, demonstrating that one correlation was significantly stronger than the other. In examining the distance score's correlations with self-report trait scores, a non-significant positive correlation was found for Agreeableness ($r = .18$, $p > .05$) and a non-significant negative correlation was found for Extraversion ($r = -.11$, $p > .05$). Upon examining the best choice score's correlations with self-report trait scores, non-significant positive correlations were found for both Agreeableness ($r = .08$, $p > .05$) and Extraversion ($r = .13$, $p > .05$). When comparing the distance score and best choice score correlations, it was found that there were no significant differences between the correlations (Agreeableness, *z-score* = .69, $p > .05$; Extraversion, *z-score* = -1.64, $p > .05$). The z-score is nearly

significant in the opposite direction for Extraversion, meaning that best choice scores may tend to be more strongly related to self-report Extraversion than distance scores are.

Hypothesis Two

Hypothesis two stated that GMA would be more strongly related to best choice scores derived from an SJT than to distance scores. Upon examining the correlations between distance scores and GMA, moderately strong negative correlations were found (distance Agreeableness r = -.35, $p < .01$; distance Extraversion r = -.36, $p < .01$). Examining the correlations between best choice scores and GMA, weak positive correlations were found (best choice Agreeableness r = .05, $p > .05$; best choice Extraversion r = .03, $p > .05$). When compared in a matched-trait fashion (e.g. distance Agreeableness score's correlation with GMA compared to best choice Agreeableness score's correlation to GMA) it was found that there were significant differences between the correlations (Agreeableness, *z-score* = 2.82, $p < .01$; Extraversion, *z-score* = 2.76, $p < .01$). It is important to note that the reason for this significant difference in scores in this case is not due to the strong positive relationship between best choice scores and GMA as expected, but instead due to the strong negative relationship between distance scores and GMA. Thus, lower distance scores are associated with higher GMA. Distance scores are transmitting more variance due to cognitive ability in that individuals with lower cognitive ability may be more likely to be less accurate and be susceptible to Accentuation Effects.

Regression Analyses

Regression analyses were also conducted in order to determine the relative prediction of GMA and personality as predictors of SJT scores when considered simultaneously . Results for Agreeableness distance scores, Agreeableness best choice scores, Extraversion distance scores, and Extraversion best choice scores were explored (see table 9). First, when distance scores for Agreeableness were regressed onto GMA and Agreeableness, GMA was found to be a

significant negative predictor (β = -.34, p < .01). However, self-reported Agreeableness was not a significant predictor in the model (β = .16, p > .05). When regressing best choice Agreeableness onto GMA and Agreeableness, neither were found to be significant predictors (GMA β = .06, p > .05, Agreeableness β = .09, p > .05). When distance scores for Extraversion were regressed onto GMA and self-reported Extraversion, GMA was found to be a significant negative predictor (β = -.35, p < .01). However, Extraversion was not a significant predictor in the model (β = -.06, p > .05). When regressing best choice Extraversion onto GMA and Extraversion, neither were found to be significant predictors (GMA β = .01, p > .05, Extraversion β = .13, p > .05).

In sum, whether considered alone or with other variables, GMA was a significant predictor of distance scores while self-report personality was not. This was true for both Agreeableness and Extraversion. Neither GMA nor self-report personality was related to best-choice scores when considered individually or together. This was true for both traits.

In addition to calculating the regression analyses for simple effects, interactions were explored the between antecedents (GMA and personality) and customer service experience. Results demonstrated that when Agreeableness was dichotomized at the median (3.58), and customer service experience was dichotomized at the median (16 months), these two predictors interacted to relate to distance scores of Agreeableness (see table 10). Mean differences demonstrated that individuals with high levels of agreeableness and a large amount of customer service experience had the highest distance scores. These results demonstrate that Agreeableness is actually related to distance scores by interacting with customer service experience.

Table 9. Regression analyses demonstrating prediction of SJT scores

Predictors of Typical performance Agreeableness

Variable	B	β	95% CI	
Constant	7.88		[4.22,	11.54]
GMA	-.14	-.34**	[-.25,	-.03]
Self-reported Agreeableness	.41	.16	[-.26,	1.07]
R^2	0.15			
F	4.26			

Notes. N=53. CI = Confidence Interval.
*p <.05, ** p < .01

Predictors of Best Choice Agreeableness

Variable	B	β	95% CI	
Constant	0.25		[-.47,	.96]
GMA	0	.06	[-.02,	.03]
Self-reported Agreeableness	.04	.09	[-.09,	.17]
R^2	0.01			
F	0.26			

Notes. N=53. CI = Confidence Interval.
*p <.05, ** p < .01

Predictors of Typical performance Extraversion

Variable	B	β	95% CI	
Constant	10.49		[7.04,	13.95]
GMA	-.15	-.35**	[-.27,	-.04]
Self-report Extraversion	-.15	-.06	[-.84,	.55]
R^2	0.13			
F	3.82			

Notes. N=54. CI = Confidence Interval.
*p <.05, ** p < .01

Predictors of Best Choice Extraversion

Variable	B	β	95% CI	
Constant	0.22		[-.44,	.87]
GMA	0	.01	[-.02,	.02]
Self-report Extraversion	.06	.13	[-.07,	.19]
R^2	0.02			
F	0.47			

Notes. N=54. CI = Confidence Interval.
*p <.05, ** p < .01

Table 10. Tables highlighting interaction between Experience and self-report Agreeableness

ANOVA Distance scores of Agreeableness

	df	*F*	*η*	*p*
Corrected Model	5	4.04	6.22	.01
Intercept	1	48.46	74.63	.00
Experience	1	.05	.07	.83
Self-report Agree	1	.11	.17	.74
Exp*Agree	1	9.41	14.50	.00

Note: Experience dichotomized (>16 months = 0, < or = 16 months = 1)

Agreeableness dichotomized (>3.58 = 0, < or = 3.58 = 1)

Means for distance Agreeableness scores

	High self-report Agreeableness	Low self-report Agreeableness
High C.S. Exp (16 months +)	6.38	5.28
Low C.S. Exp	5.43	6.26

Note: Experience dichotomized (>16 months = 0, < or = 16 months = 1)

Agreeableness dichotomized (>3.58 = 0, < or = 3.58 = 1)

Hypothesis Three

Hypothesis three stated that distance scores derived from an SJT would have a stronger relationship with typical performance than would best choice scores. Hypothesis three was also tested using Fisher's r-to-z transformations, only this time the SJT scores were the antecedents. When compared in a matched-trait fashion (e.g. distance Agreeableness score's correlation with the typical performance Agreeableness to best choice Agreeableness score's correlation the typical performance Agreeableness) it was found that there were significant differences between the correlations (Agreeableness, $z\text{-}score = -2.47$, $p < .01$; Extraversion, $z\text{-}score = 3.12$, $p < .01$). Interestingly, these results demonstrate that distance scores more strongly predicted typical ratings of Extraversion, but best choice scores more strongly predicted typical ratings of Agreeableness. Regression analyses supported these results, demonstrating that when considered together only best choice scores predicted typical performance Agreeableness ($\beta = .39$) and only distance scores predicted typical performance Extraversion ($\beta = .38$). To additionally support the regression analyses, relative weights of predictor variables were calculated (see table 11 for a summary of the regression results). Relative weights are an often requested measure to determine the relative importance of the predictor variables in a multiple regression analysis.

Table 11. Regression analyses demonstrating prediction of criteria

Predictors of Typical Performance Agreeableness

Variable	B	B	95% CI	
Constant	3.39		[2.26,	4.52]
Age	-.04	.37*	[-.08,	0]
C.S. Experience	0	.49**	[0,	.01]
Gender	.12	.08	[-.32	.55]
Distance Agreeableness	.05	.10	[-.09,	.20]
Best Choice Agreeableness	1.10	.39**	[.34,	1.87]
R^2	0.26			
F	4.14			

Notes. N=45. CI = Confidence Interval. Gender (0=female, 1=male)
*p <.05, ** p < .01

Relative Contribution to Multiple R (reported as percentages)

	Typical Performance Agreeableness
Age	6.4
Gender	0.3
C.S. Experience	16.9
Distance Agree	1.1
Best Choice Agree	75.3

Table 11 (continued). Regression analyses demonstrating prediction of criteria.

Predictors of Typical Performance Extraversion

Variable	B	β	95% CI	
Constant	2.96		[1.69,	4.22]
Age	-.04	.39*	[-.08,	0]
C.S. Experience	0	.48**	[0,	.01]
Gender	-.07	-.04	[-.50	.37]
Distance Extraversion	.20	.38*	[.05,	.35]
Best Choice Extraversion	-.03	-.01	[-.76,	.69]
R^2	0.19			
F	3.05			

Notes. N=52. CI = Confidence Interval. Gender (0=female, 1=male)
*p <.05, ** p < .01

Relative Contribution to Multiple R (reported as percentages)

	Typical Performance Rated Extraversion
Age	13.9
Gender	6.8
C.S. Experience	26.7
Distance Extra	49.3
Best Choice Extra	3.4

Table 11 (continued). Regression analyses demonstrating prediction of criteria

Predictors of Maximum Performance Agreeableness

Variable	B	β	95% CI	
Constant	2.88		[1.57,	4.19]
Age	.04	.29*	[0,	.09]
C.S. Experience	0	-.02	[-.01,	.01]
Gender	-.27	-.17	[-.77	.22]
Distance Agreeableness	-.10	.08	[-.26,	.06]
Best Choice Agreeableness	-.27	-.09	[.34,	1.87]
R^2	0.11			
F	1.17			

Notes. N=52. CI = Confidence Interval. Gender (0=female, 1=male)
*p <.05, ** p < .01

Relative Contribution to Multiple R (reported as percentages)

	Maximum Performance Agreeableness
Age	50.9
Gender	12.9
C.S. Experience	1.1
Distance Agree	15.6
Best Choice Agree	19.5

Table 11 (continued). Regression analyses demonstrating prediction of criteria

Predictors of Maximum Performance Extraversion

Variable	B	β	95% CI	
Constant	2.08		[.62,	3.53]
Age	.02	.13	[-.02,	.06]
C.S. Experience	0	.25	[0,	.02]
Gender	-.11	-.07	[-.259	.37]
Distance Extraversion	.03	.06	[-.13,	.19]
Best Choice Extraversion	.07	.03	[-.72,	.87]
R^2	0.10			
F	1.01			

Notes. N=52. CI = Confidence Interval. Gender (0=female, 1=male)

*p <.05, ** p < .01

Relative Contribution to Multiple R (reported as percentages)

	Maximum performance Extraversion
Age	20.8
Gender	11.6
C.S. Experience	59.4
Distance Extra	4.1
Best Choice Extra	4.1

Relative weight is defined as the "proportionate contribution each predictor makes to r^2, considering both its unique contribution and its contribution when combined with other variables" (Johnson, 2000, p. 1). Relative weights were also calculated and are included below the regression analyses. Relative weights demonstrated that best choice scores for Agreeableness accounted for roughly 75 percent of the multiple R, while distance scores accounted for 1 percent of the multiple R for typical performance Agreeableness scores. Relative weights also demonstrated that distance scores for Extraversion accounted for roughly 45 percent of the multiple R, while best choice scores accounted for 3 percent of the multiple R for typical performance Extraversion scores.

In sum, the optimal SJT scoring method for predicting typical behavior depended on the trait being measured. Whether the two indices were considered alone or in combination, only best choice scores predicted typical performance Agreeableness, and only distance scores predicted typical performance Extraversion.

Hypothesis Four

Hypothesis four stated that best choice scores derived from an SJT will have a stronger relationship with maximum performance than will distance scores. This hypothesis was also tested using Fisher's r-to-z analysis. First, the correlations were compared in a match trait fashion (e.g. the correlation between best choice Extraversion scores to maximum performance Extraversion compared to the correlation between distance scores of Extraversion and maximum performance Extraversion). Upon conducting this analysis, it was found that there was none of the indices were significantly correlated with maximum performance ratings and none of the correlations were significantly different from one another. The results demonstrated no significant differences between maximum performance Agreeableness and best choice scores of Agreeableness ($r = -.12$, $p > .05$)

and maximum performance Agreeableness and distance scores of Agreeableness (r = -.08, p > .05) (z-*score* = -.02, p > .05). Additionally, there were also no significant differences found when comparing the correlations between best choice Extraversion and maximum performance Extraversion (r = -.07, p > .05) with the correlation between distance scores of Extraversion and maximum performance Extraversion (r =.04, p > .05) (z-*score* = -.55, p < .05). Thus, hypothesis 4 was not supported. Next, hypothesis four was tested further using regression analyses in which the SJT indices were considered together as shown in table 10. When maximum performance was regressed onto both distance scores and best choice scores simultaneously, neither of the indices were unique predictors (Agreeableness distance, β = .08, p > .05, Agreeableness best choice β = -.09, p > .05, Extraversion distance β = .06, p > .05, Extraversion best choice β = .03, p > .05). Additionally, relative weights demonstrated very small differences in contributions. Specifically, relative weights demonstrated that for maximum performance Agreeableness, distance scores accounted for roughly 16 percent of the multiple R, with best choice scores accounting for roughly 20 percent. For maximum performance Extraversion, both types of scores accounted for roughly 4 percent of the multiple R. Thus, in summary, none of the SJT indices were significant predictors of maximum performance whether considered alone or in combination. Note however that these findings differed for minorities, as described earlier in the results section.

Hypothesis Five

Hypothesis five stated that distance scores derived from an SJT would partially explain covariance between personality and typical performance measures. Mediation for hypothesis five was tested by using correlation and regression analyses through the Baron and Kenney (1986) method. There are four steps in establishing mediation. First it must be demonstrated that the independent variable is related to the outcome variable. Then, it must be shown that the independent variable is related to the mediator. Subsequently, it must be shown that the mediator

is related to the outcome. Finally, it must be demonstrated that the inclusion of the mediator in a regression analysis reduces (partial mediation) or eliminated (full mediation) the variance accounted for by the independent variable on the outcome variable (Baron and Kenney, 1986). See table 12 for a summary of the regressions for the mediation analyses. Results indicated that self-report Extraversion was not a significant predictor of distance scores for Extraversion (β = -.11, p > .05). Thus, given this was a necessary condition of mediation, no further steps were tested. When testing the mediation analysis for Agreeableness, it was found that self-report Agreeableness was not a significant predictor of distance scores for Agreeableness (β = .18, p < .05). Given this was a necessary precondition of mediation, no further steps were tested. In sum, hypothesis 5 was not supported given that self-reported personality was not a significant predictor of distance scores for either trait.

Table 12. Mediation Results

Predictors of Distance scores of Extraversion

Variable	B	β	95% CI
Constant	7.29		[4.97, 9.61]
Self-reported Extraversion	-.29	-.11	[-.94, .36]
R^2	.01		
F	.81		

Notes. N=64. CI = Confidence Interval.
*p <.05, ** p < .01

Predictors of Distance Agreeableness

Variable	B	β	95% CI
Constant	4.40		[2.13, 6.66]
Self-reported Agreeableness	.44	.17	[-.19, 1.06]
R^2	.03		
F	1.97		

Notes. N=63. CI = Confidence Interval.
*p <.05, ** p < .01

Hypothesis Six

Hypothesis six stated that best choice scores would have greater incremental validity over personality measures in explaining variance in typical performance than would distance scores. When testing hypothesis six, regression analyses demonstrated that this hypothesis was supported for Agreeableness. First, when examining a model that regressed typical performance Agreeableness onto self-report Agreeableness, best choice scores for Agreeableness, and distance scores of Agreeableness, it was found that best choice scores were a significant predictor. The beta for best choice scores was stronger than the beta for self-report Agreeableness and distance agreeableness in predicting typical performance Agreeableness (self-report Agreeableness $\beta = .31$, $p > .05$; best choice scores for Agreeableness $\beta = .41$, $p < .01$, distance scores for Agreeableness $\beta = .01$, $p > .05$). When considering relative contribution to multiple R, it was found that best choice Agreeableness contributed 87.7 percent, with distance scores for Agreeableness contributing only 0.5 percent to multiple R for typical Agreeableness. Thus, when running these regressions, it was demonstrated that best choice Agreeableness scores had greater incremental validity than distance scores for Agreeableness beyond that contributed by self-reported Agreeableness. In fact, when these three predictors were considered together, only the best choice index was a unique predictor. These results were not supported for Extraversion. When regressing typical Extraversion onto best choice Extraversion, self-report Extraversion, and distance scores of Extraversion, it was found that self-report Extraversion was a significant predictor ($\beta = .41$, $p < .01$). However, in this regression, best choice scores were not a significant predictor ($\beta = -.15$, $p > .05$). Distance scores of Extraversion were a significant predictor ($\beta = .40$, $p < .01$). When considering relative contribution to multiple R for typical performance Extraversion, self-report extraversion was the strongest contributor (62.4 percent), distance scores of Extraversion contributed 32.7 percent, and best choice Extraversion

contributed only 4.9 percent. Thus, the pattern of results demonstrated that distance scores had greater incremental validity than best choice scores when it came to predicting typical performance extraversion. Thus, six was not supported for Extraversion. See table 13 for regression tables demonstrating incremental validity.

In sum, consistent with findings regarding the validity of distance and best choice SJT scores as predictors of typical behavior, incremental validity was greater for best choice agreeableness. However, incremental validity was greater for distance scores of Extraversion.

Table 13. Incremental Validity

Predictors of Typical Performance Agreeableness

Variable	B	β	95% CI	
Constant	1.88		[.38,	3.38]
Self-reported Agreeableness	.31	.23	[-.06,	.68]
Best choice Agreeableness	1.23	.41**	[.43,	2.03]
Distance Agreeableness	0	.01	[-.14,	.15]
R^2	0.24			
F	4.52			

Notes. N=48. CI = Confidence Interval.
*p <.05, ** p < .01

Relative Contribution to Multiple R (reported as percentages)

Typical Performance Agreeableness

Agreeableness	11.9
Best choice Agree	87.7
Distance Agree	0.5

Table 13 (continued). Incremental Validity

Predictors of Typical Performance Extraversion

Variable	B	β	95% CI
Constant	0.63		[-.80, 2.08]
Self-reported Extraversion	.51	.41**	[.19, .83]
Best choice Extraversion	-.38	-.15	[-1.06, .29]
Distance Extraversion	.19	.40**	[.07, .31]
R^2	0.29		
F	6.23		

Notes. N=48. CI = Confidence Interval.

*p <.05, ** p < .01

$\Delta R^2 = .16$

Relative Contribution to Multiple R (reported as percentages)

	Typical Performance Extraversion
Extraversion	62.4
Best choice Extra	4.9
Distance Extra	32.7

Hypothesis Seven

Hypothesis seven stated that best choice scores and distance scores would each contribute incrementally to the prediction of typical performance. When testing hypothesis seven, regression analyses were run with typical performance being regressed in the same model onto distance scores and best choice scores (see table 14). For Extraversion, distance scores were a significant predictor in the model (β = .36, p < .05) but best choice scores were not (β = -.09, p > .05). For Agreeableness, the distance scores were not a significant predictor of typical performance (β = .05, p > .05), but best choice scores were significant (β = .43, p < .01). In sum, in both cases only one of the two indices contributed uniquely to the prediction of typical performance when considered together. Consistent with the analyses presented in hypothesis 3 and 6, only best choice scores predicted typical performance Agreeableness and only distance scores predicted typical performance Extraversion.

Table 14. Tables Demonstrating Incremental Validity and Interactions of Distance Scores and Best Choice Scores

Predictors of Typical Performance Extraversion

Variable	B	β	95% CI	
Model 1				
Constant	2.50		[1.59,	3.41]
Distance Extraversion	-.24	.36*	[-.97,	.50]
Best Choice Extraversion	.17	-.09	[.04,	.30]
R^2	0.14			
F	3.53			
Model 2				
Constant	1.07		[1.35,	.19]
Distance Extraversion	.40	.84**	[3.23,	0]
Best Choice Extraversion	3.14	1.23*	[1.97,	.06]
Inter. Best and Distance Extra	-.56	-1.43*	[-2.16,	0]
R^2	0.22			
F	4.10			
Model 3				
Constant	-.83		[-.92,	.36]
Distance Extraversion	.42	.89**	[3.82,	0]
Best Choice Extraversion	3.14	1.19*	[2.13,	.04]
Inter. Best and Distance Extra	-.56	-1.45**	[-2.45,	.02]
NEO Extraversion	.51	.41**	[3.40,	0]
R^2	0.38			
F	6.70			

Notes. N=47. CI = Confidence Interval.
*p <.05, ** p < .01
ΔR^2 (model 1 to model 2) = .08*
ΔR^2 (model 2 to model 3) = .17**

Table 15 (continued). Tables Demonstrating Incremental Validity and Interactions of Distance Scores and Best Choice Scores

Predictors of Maximum Performance Extraversion

Variable	B	β	95% CI	
Model 1				
Constant	2.78		[1.89,	3.68]
Distance Extraversion	-.17	.04	[-.11,	.15]
Best Choice Extraversion	.02	-.07	[-.89,	.55]
R^2	0.01			
F	0.16			
Model 2				
Constant	2.67		[1.03,	4.30]
Distance Extraversion	.04	.08	[-.22,	.29]
Best Choice Extraversion	.12	.05	[-3.26,	3.51]
Inter. Best and Distance Extra	-.05	-.12	[-.58,	.48]
R^2	0.01			
F	0.17			
Model 3				
Constant	2.76		[-.65,	4.87]
Distance Extraversion	.04	.08	[-.22,	.30]
Best Choice Extraversion	.13	.05	[-3.29,	3.55]
Inter. Best and Distance Extra	-.05	-.12	[-.58,	.49]
NEO Extraversion	-.03	-.02	[-.37,	.32]
R^2	0.01			
F	0.09			

Notes. N=56. CI = Confidence Interval.
*p <.05, ** p < .01
ΔR^2 (model 1 to model 2) = 0
ΔR^2 (model 2 to model 3) = 0

Table 14 (continued). Tables Demonstrating Incremental Validity and Interactions of Distance Scores and Best Choice Scores

Predictors of Typical Performance Agreeableness

Variable	B	β	95% CI	
Model 1				
Constant	2.85		[1.90,	3.81]
Distance Agreeableness	.02	.05	[-.12,	.17]
Best Choice Agreeableness	1.27	.43*	[.46,	2.08]
R^2	0.19			
F	5.16			
Model 2				
Constant	3.08		[.77,	5.40]
Distance Agreeableness	-.01	-.03	[-.40,	.37]
Best Choice Agreeableness	.76	.25	[-4.03,	5.55]
Inter. Best and Distance Agree	.09	.20	[-.69,	.86]
R^2	0.19			
F	3.38			
Model 3				
Constant	2.12		[-.44,	4.68]
Distance Agreeableness	-.04	-.07	[-.41,	.34]
Best Choice Agreeableness	.69	.23	[-4.01,	5.39]
Inter. Best and Distance Agree	.09	.20	[-.67,	-.85]
NEO Agreeableness	.31	.23	[-.07,	.68]
R^2	0.24			
F	3.33			

Notes. N=47. CI = Confidence Interval.
*p <.05, ** p < .01
ΔR^2 (model 1 to model 2) = 0
ΔR^2 (model 2 to model 3) = -.05

Table 14 (continued). Tables Demonstrating Incremental Validity and Interactions of Distance Scores and Best Choice Scores

Predictors of Maximum Performance Agreeableness

Variable	B	β	95% CI	
Model 1				
Constant	3.32		[2.34,	4.29]
Distance Agreeableness	-.04	-.07	[-.19,	.11]
'Best Choice' Agreeableness	-.34	-.11	[-1.17,	.49]
R^2	0.02			
F	0.52			
Model 2				
Constant	4.41		[2.08,	6.74]
Distance Agreeableness	-.22	-.41	[-.61,	.16]
'Best Choice' Agreeableness	-2.79	-.93	[-7.60,	2.03]
Inter. Best and Distance Agree	.40	-.93	[-.38,	1.19]
R^2	0.04			
F	.71			
Model 3				
Constant	5.23		[2.63,	7.83]
Distance Agreeableness	-.20	-.37	[-.58,	.18]
'Best Choice' Agreeableness	-2.73	-.91	[-7.50,	2.05]
Inter. Best and Distance Agree	.40	.91	[-.38,	1.18]
NEO Agreeableness	-.26	-.19	[-.65,	.12]
R^2	0.08			
F	1.02			

Notes. N=54. CI = Confidence Interval.
*p <.05, ** p < .01
ΔR^2 (model 1 to model 2) = .02
ΔR^2 (model 2 to model 3) = .03

Exploratory Analyses

Beyond the exploration of the hypothesized model, interactions between the scoring techniques were explored. First, an interaction was explored between best choice scores and distance scores in predicting typical performance ratings of Extraversion. Results demonstrated that there was a significant interaction (β = -1.43, $p < .05$). As shown in Figure 13, when best choice scores high, distance scores were not predictive. However, when best choice scores were low, distance scores were positively related to typical performance Extraversion. These results demonstrate that distance scores predict typical performance only when best choice scores are low. Similar regression analyses were run to examine whether best choice and distance scores interacted to predict typical performance Agreeableness, maximum performance Agreeableness, and maximum performance Extraversion. As shown in table 14, there were no significant interactions between the scoring techniques in predicting these criteria.

(see Figure 13)

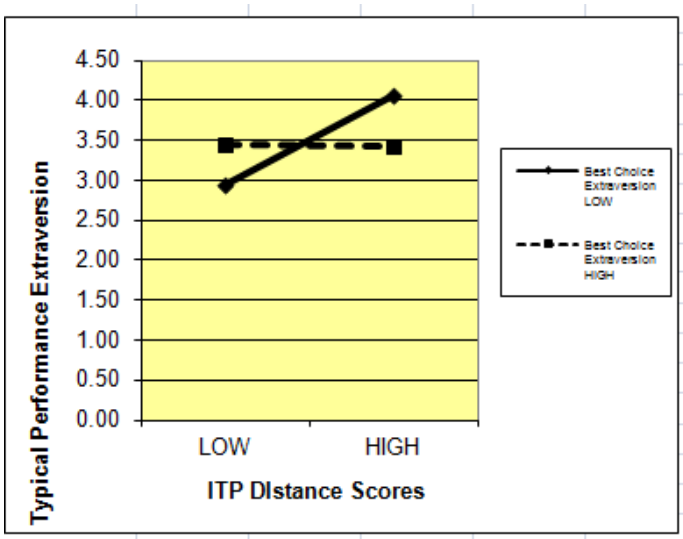


Figure 13. Graph of typical performance Extraversion regressed onto SJT distance scores moderated by SJT best choice scores

Hypothesis	Result
1. Hypothesis 1:	
Self-reported personality measures will be more strongly related to distance scores derived from an SJT than to best choice scores.	Partially supported, self-report Agreeableness and customer service experience interacted to relate to distance scores but not best choice scores
2. Hypothesis 2:	
GMA will be more strongly related to best choice scores derived from an SJT than to distance scores.	Partially supported, significant difference found with distance scores having a stronger significant but negative relationship
3. Hypothesis 3:	
Distance scores derived from an SJT will have a stronger relationship with typical performance than will best choice scores.	Partially Supported, best choice scores more strongly predicted typical performance Agreeableness, distance scores more strongly predicted typical performance Extraversion
4. Hypothesis 4:	
Best choice scores derived from an SJT will have a stronger relationship with maximum performance than will distance scores.	Not supported
5. Hypothesis 5:	
Distance scores derived from an SJT will partially explain covariance between personality and typical performance measures.	Not supported
6. Hypothesis 6:	
Best choice SJT scores will have greater incremental validity over personality measures in explaining variance in typical performance than will SJT distance scores.	Supported for Agreeableness, opposite pattern found for Extraversion

7. Hypothesis 7:	
Best choice SJT scores and SJT distance scores will each contribute incrementally to the prediction of typical performance.	Not Supported
8. Exploratory Analysis	
Do SJT scores relate differentially to different traits?	In predicting typical performance, best choice scores were more predictive of agreeableness, distance scores were more predictive of extraversion
9. Exploratory Analysis	
Do SJT distance scores and SJT best choice scores interact to predict criteria?	For Caucasian participants, when best choice scores are low distance scores are more positively associated with typical Extraversion. No interaction between SJT indices was found for Agreeableness.
10. Exploratory Analysis	
Do mean differences exist in assessment variables for gender or racial subgroups?	Caucasian females had higher distance scores than Caucasian males. Minorities had higher typical performance ratings of Agreeableness than Caucasians. Gender and race were shown to interact in the prediction of typical performance Extraversion.
11. Exploratory Analysis	
Do SJT scores demonstrate differential validity for race and gender subgroups?	Significant differences were found for gender and race in validity coefficients between GMA and distance scores, and in validity coefficients between best choice scores and typical performance ratings
12. Exploratory Analysis	
Do SJT scores demonstrate differential prediction for race and gender subgroups?	Significant interactions were found for race and GMA in predicting best choice scores, for race and both SJT scores in predicting maximum performance, and for race and both SJT scores in predicting typical performance

CHAPTER SEVEN: DISCUSSION

Summary of Results

The objective of the current study was to investigate whether significant differences existed between SJT scoring techniques with respect to construct, criterion-related, and incremental validity. Specifically, the study hypothesized a model for differential relationships between SJT scores and antecedents/criteria of interest. This model posited that distance scoring methods would be more strongly related to personality and, as a result, to typical performance, while best choice scoring methods would be more strongly related to GMA and, as a result, to maximum performance (see figure 1). While many of the specific relationships illustrated in my theoretical model were not supported, the general thesis of this dissertation, the fact that SJT scoring method can have a significant impact on relationships with variables of interest, was supported by my findings. There are three noteworthy themes that emerge when examining the data. First, SJT scores showed differential validity and prediction as a function of race. For instance, distance scores were more strongly related to GMA for Caucasians than for minorities. Second, the differential effectiveness of best choice and distance scores often varied (and reversed) depending on the dimension/trait being measured. Third, SJT indices were differentially related to maximum and typical performance criteria.

Relations with personality and GMA. Neither of the self-report personality measures correlated significantly with either of the SJT indices. Moreover, in contrast to hypothesis 1, a simple best choice scoring method resulted in a non-significant positive correlation between SJT scores and GMA (mean correlation across traits $r = .04$), while a more complex distance score actually resulted in a negative relationship between SJT scores and GMA (mean correlation across traits $r = -.36$). These results suggest that a distance scoring method may transmit variance in GMA by identifying those most prone to accentuation biases in their SJT scores.

Relations with maximum performance. SJT scores (both best choice and distance scores) were positively related to maximum performance ratings of Extraversion for minority participants. However, SJT scores were not, regardless of scoring method, positively related to maximum performance on either of the trait dimensions for Caucasian participants. In fact, the slope for distance scores of extraversion and best choice scores of agreeableness was actually negative when predicting maximum performance. It is interesting to note that maximum and typical performance scores were not significantly related. However, this is similar to previous findings (e.g., Klehe & Anderson, 2007; Dubois, Sackett, Zedeck, & Fogli, 1993) which have demonstrated that for certain dimensions there are no significant relationships or very small relationships between maximum and typical performance.

Relations with typical performance. While there were not significant correlations between SJT scores and self-report personality, SJT scores did predict typical performance ratings but only for Caucasian participants. The specific pattern found, however, differed as a function of the trait in question. Only best choice scores were unique predictors of typical Agreeableness and this remained true when self-reported Agreeableness was added to the equation. Distance scores interacted with best choice scores to predict typical Extraversion for Caucasian participants and this was true even when self-reported Extraversion was included as a predictor. Specifically, the positive slope for the relationship between SJT distance scores and typical performance ratings of extraversion was steeper for participants who scored low on the best choice index. In fact, for those scoring highest on the best choice index for Extraversion, the slope of the relationship between distance scores and typical performance was essentially flat. It is interesting to note that SJTs predict typical performance and not maximum performance for Caucasians. This may be because previous research has demonstrated that high-fidelity SJTs

such as the one utilized in this study tend to predict interpersonally-oriented criteria much more strongly than cognitive criteria (Lievens & Sackett, 2006). As maximum performance is related to cognitive ability, and typical performance was rated by partners who had interpersonally interacted with the participant, it can be expected that the video based SJT would better predict typical performance.

Theoretical Implications

The primary goal of this study was to investigate whether different methods of scoring SJTs would demonstrate differential relationships with antecedents (personality and GMA) and criteria (maximum and typical performance). Theoretically, the manipulation of scoring technique would have substantial effects on the correlations between SJT scores and variables of interest. While the proposed theoretical mechanisms may not have been wholly supported, it has been inarguably demonstrated that there are significant differences in what is predicted with different SJT scoring methods. As these scoring methods are all based on value judgments the participants have regarding the same response items to the same item stems, this variation truly demonstrates the importance of considering scoring technique when developing and administering an SJT.

SJT scores and Antecedents

When considering the results with regards to personality, while it was demonstrated that personality interacted with customer service to predict SJT scores, there were no direct significant relationships found. Other studies have demonstrated small correlations between SJT scores and personality (Agreeableness ρ = .25, Extraversion ρ = .14, McDaniel & Nguyen, 2001). Even SJTs designed to capture personality have found modest correlations (Agreeableness r = .34, Extraversion r = .32, Motowildo, Hooper & Jackson, 2006b). It is possible that capturing self-report personality through a situational judgment test is difficult due to the other constructs that correlate with SJT scores and thus cloud relationships with personality, such as procedural knowledge and job experience (Motowidlo, Hooper, & Jackson, 2006a).

When considering the results with regard to GMA, it is important to note that the significant differences in correlations between SJT scores and GMA for distance scores and best choice scores were due to the significant negative relationship between GMA and the distance scores, and not due to a significant positive relationship between GMA and best choice scores as expected. These results demonstrate that the higher an individual scores on GMA, the less scale points they are likely to use when rating the effectiveness of SJT item responses. This is may be caused by intelligent individuals being less susceptible to Accentuation biases and thus less likely to make errors in exaggerating their scores. Although the results of this study did not support that intelligence was related to accuracy as was expected, it is possible that individuals who are intelligent may be less likely to commit the types of errors that cause systematic distortion of SJT ratings. This was supported in that there was a negative relationship between intelligence and distance scores.

Additionally, there were unexpected findings with regard to the relationships between SJT scores and criteria. For example, there was a relationship between distance scores for Extraversion and typical performance Extraversion ($r = .36$, $p < .05$) and a significant relationship between best choice scores for Agreeableness and typical Agreeableness ($r = .43$, $p < .01$). This finding may be due to the variables involved. It is possible that there were some differences in scripting the item responses for different traits. Feasibly, Extraversion may lend itself to creating item responses that express the trait at various levels, but it may have been difficult for participants to distinguish which item response was the best choice for this trait. In other words, when scripting Extraversion there may have been some type of ceiling effect, whereby the difference between scripted items as ranked ‘four’s’ and ‘five’s’ out of five trait levels may have been very small and thus undetectable. Conversely, Agreeableness may have

been incorrectly scripted to elicit breadth in the ratings of effectiveness. However, it may be easier to make the best choice response transparent enough to be detected for this particular trait. It is important to remember that the script for item responses was developed specifically to capture these traits, and great care was taken to differential the levels of trait expression. The test was validated using an SME rater. It is possible that existing SJTs that are currently utilized by practitioners may also have issues with differentiation between levels of trait expression.

It is also possible that the Accentuation Effect is stronger for particular constructs. In other words, Accentuation may be exaggerated in very external traits such as Extraversion. However, in other traits such as Agreeableness, Accentuation may be less pronounced. It is possible the strength of these effects varies by dimension. Another possibility is that particular traits are more observable (e.g. Extraversion), and that these traits lend themselves to being distinguished between and perhaps exaggerated at different levels of expression. Other traits may not be as observable (e.g. Agreeableness) and individuals may only be able to make surface judgments regarding the best choice option to deal with the presented scenario. This would be an extension onto Vazire's self-other knowledge asymmetry model (2010), which states that either self-ratings or partner ratings are more predictive of criteria depending upon the observability of the trait. This research would build on this theory and demonstrate that observable traits may affect more than the accuracy of ratings depending on the individual making the ratings. It may be that the observability of a trait affects the cognitive process the individual engages in while making rating judgments. Future research is needed to determine if different scoring techniques are necessary for different traits, and through what theoretical mechanisms these scoring techniques function.

Implicit Trait Policies

Implicit Trait Policies are an individual's implicit believes concerning the effectiveness of different behavioral choices that demonstrate varying levels of targeted traits (Motowidlo, Hooper, and Jackson, 2008). Implicit Trait Policy scoring is based upon three effects: Assimilation Effects, Contrast Effects, and Accentuation. By measuring the distance between the effectiveness ratings for items rated as most effective and items rated as least effective, it is hypothetically possible to capture Accentuation and measure personality implicitly. If the distance scores are indeed capturing Accentuation effects, relationships with self-reported personality should be stronger for distance scores than 'best choice' scores. However, results for distance scores demonstrated non-significant relationships with self-report personality.

Despite this, analyses demonstrated that when predicting typical performance Extraversion, distance scores best choice scores and the interaction term were all significant predictors (see table 12). This demonstrates that both scoring techniques may be useful to collect from participants, as they are both contributing unique variance. When self-report Extraversion is added to the model, the best choice scores, distance scores, and the interaction term remain significant. These results demonstrate that the indices are capturing variance that is unique and distinct from self-report scores. While these SJT scoring indices did not always correlate as was expected with antecedent variables, the significant relationships with criteria and unique variance captured by SJT scores demonstrate the value to utilizing both of these scoring indices.

Validity of Implicit Trait Policies

Results demonstrating correlations between distance scores and self-report personality traits have been difficult to achieve in the past. In his research on ITPs, Dr. Motowidlo has been

able to achieve correlations on the magnitude of roughly $r = .3$ between self-report personality scores and distance scores (Motowidlo, Hooper, & Jackson, 2006). The present study found non-significant between distance scores and criteria. This may be due to criteria used, and due to the fact that the SJTs are capturing specific customer service knowledge that may not generalize to broader criteria or to transparent situations. This may also be because the SJTs capturing unique procedural knowledge not captured by other measures.

One fact that can be gleaned from this data is that Implicit Trait Policies are quite difficult to measure. There are several ways to utilize the theoretical mechanisms, and additionally there are several ways to compile the information gathered from participants into an overall distance score or ITP score for a trait. Different methods of compiling the data that were hypothetically based on the same theoretical tenants have resulted in drastically different results. Again, this emphasizes the importance to considering scoring method, and in determining the theoretical mechanisms that cause the pattern of results demonstrated. Due to incremental validity of some of these scoring techniques in predicting performance criteria, this study may have demonstrated a prediction method which supplements self-reports of personality. However, further analysis is necessary to help explain some of the unexpected results found in this study and to further refine the Implicit Trait Policy capturing method.

Differential Effects of Scoring Technique

As previously mentioned, Ployhart and Ehrhart (2003) have explored some of the differential effects of scoring technique. Specifically, the authors compared three different scoring approaches: one using a forced choice rating of the most effective response only 'best choice,' one using a forced choice rating of the most and least effective response 'best/worst

choice,' and another using a 1-5 rating of effectiveness for each response option (effectiveness ratings; only ratings of the keyed correct and incorrect responses were used). Their results indicated moderate convergence of the SJT scores resulting from these three scoring approaches. Specifically, the best choice scores correlated with the 'best/worst choice' scores at r = .38, whereas the best choice scores correlated with the effectiveness ratings scores moderately at r = .32, and the 'best/worst choice' scores correlated moderately with the effectiveness ratings scores at r = .37. This study builds on the previously study by exploring additional scoring techniques and demonstrating additional validity. In the current study, the correlations between scoring techniques was lower, demonstrating that the differential effects of scoring technique may be more pronounced than previously believed. Distance scores correlated with best choice scores at r = .03. This study explored an innovative scoring technique, and thus it may be expected that inter-correlations would be lower. However, results also demonstrate a drastic moderating effect of scoring technique on relationships between SJT scores, antecedents, and criteria. Significant interactions were demonstrated with regard to scoring technique, demonstrating that scoring technique can affect the prediction of criteria and the adverse impact of SJTs.

Self-reported Versus Peer-reported Trait-related Behavior

Peer reports of the typical expression of trait-related behavior were more strongly associated with SJT scores than were self-reports of the same personality traits. In the following sections, I posit a number of possible explanations for these findings.

Differences between self-perception and reality. Research has previously demonstrated that individuals are able to predict some aspects of their behavior, but not others well. The researchers also found that peer ratings were able to address deficiencies in self ratings (Vazire, 2010). Specifically, the researchers found that individuals were better able to predict their own

behavior regarding traits that were not highly observable (e.g. self-esteem) while friends may be better at predicting traits that are highly observable (e.g. Extraversion).

Other researchers have found similar results. For example, Kolar et al. (1996) found support for the concept that self ratings and peer ratings of personality were both accurate predictors of behavior during an interaction with a stranger. However, they found self-ratings predicted different behaviors (e.g., calmness) than did peer ratings (e.g., humor, likeability, and arrogance). Additionally, Vazire and Mehl (2008) found that self ratings and peer ratings also accurately predicted behavior, but again self-ratings predicted different behaviors (e.g., arguing) than did peer ratings (e.g., socializing).

This idea is also supported by previous research on non-acquaintance ratings of personality, which demonstrates that highly observable traits such as Extraversion are relatively easy to make accurate assessments regarding with very little supplementary information. However, less observable traits such as neuroticism are much more difficult to rate (e.g. Borkenau, Brecke, Möttig, & Paelecke, 2009).

Situation specificity of trait-related behavior. One viewpoint that may be relevant to this particular study is Mishel's (1968) hypotheses regarding the strong influence of the situation on the expression of personality. Mischel proposed that when considering the person by situation interaction that influences behavior, situational factors have a very strong and often underestimated influence on behavior. This would be particularly relevant when analyzing personality through situations. Other researchers have build upon this argument, and modified the theory to be presented as a 'self-presentation account' (Johnson, 1981). Specifically, these researchers believe that an individual may display one behavior in a particular setting that is unrepresentative of their behavior outside of that setting. It is possible that a customer service

based SJT captures the situation-specific expression of personality traits that are relevant to customer service situations. In other words, an individual's expressed level of Extraversion may be different in social settings versus the workplace. This would attenuate relationships between SJT scores and self-reported personality as was found in the present study. Moreover, this explanation is also with the stronger positive relationships between SJT scores and typical performance ratings given that peer raters in the present study rated their partner's tendency to behave and a trait-specific manner in specific situations – the identical situations presented to participants in the SJT.

Additionally, SJTs are notorious for having low reliability and unstable relationships with constructs of interest (Schmidt and Chan, 2006). This study found especially low reliability estimates for best choice SJT scores (Agreeableness $\alpha = .47$, Extraversion $\alpha = .31$). This may explain the difficulty in finding relationships with constructs of interest. It is possible distance scores are tapping into a situation by trait interaction, which would reduce the relationships between distance scores and personality.

Differential Validity for Race and Gender

Construct validity. As previously discussed, GMA tests are often used in the employment selection process. Unfortunately, they also frequently demonstrate some level of adverse impact Hausdorf, LeBlanc & Chawla, 2002; Berry, Clark & McClure, 2011). The adverse impact of GMA tests can even be demonstrated in the results of this study, where Caucasians performed significantly better minorities on the Wonderlic and in the SJT (see tables 3, 4, and 5). This study also demonstrated that distance scores and best choice scores were differentially related to GMA for Caucasian and minority participants (see tables 2 and 3) and these distance and best choice scores also have differential prediction and validity. This is

consistent with previous research which has shown that GMA tests demonstrate some level of differential validity in predicting performance (Berry, Clark & McClure, 2011).

Specifically, for Caucasians, GMA negatively related to distance scores (r = -.36, p < .05). However, this relationship was a very small positive correlation when examining the minority population (r = .04, p > .05). This is interesting, because it demonstrates that GMA may differentially affect responses to SJTs for different races. It is also possible that tests of GMA simply do not capture GMA as well for minorities, as Caucasians scored significant higher on the GMA test.

An additional factor which was considered is fidelity; video-based SJTs tend to result in less adverse impact than written SJTs because video-based SJTs are less cognitively loaded (Chan & Schmitt, 1997). It was hypothesized that the level of fidelity of this SJT would result in less adverse impact relative to other forms of testing. However, because of the very different results found for Caucasians and other races with regard to differential prediction of criteria, it seems that there are effects that may be caused by cultural differences in their interpretation of variables involved. For example, Hispanics often tend to have different cultural expectations with regard to Assertive behavior and Agreeableness (Planellas-Bloom, 1992). These cultural differences may translate into differential prediction for race. Further study should be conducted in order to attain an adequate number of racial minorities to allow for sub-group comparison.

Criterion-related validity. SJTs are often considered to be selection tests which reduce adverse impact (e.g., Clevenger et al., 2001; Harold & Ployhart, 2001; Jenson, 1998; McDaniel & Nguyen, 2001; Motowidlo et al., 1990; Oswald et al., 2004; Pulakos & Schmitt, 1996; Weekley & Jones, 1999). However, this study found significant differential validity and differential prediction. The previous studies examined adverse impact, or the ratio of individuals

who achieved a score higher than a set cut score for each racial subgroup. This study went further to determine if there was differential prediction or differential validity for subgroups. To the author's knowledge, no known study has previously explored differential prediction for SJT scores. As previously discussed, there were differential results in the prediction of typical performance. Specifically, for Caucasians there was a relationship between distance scores Extraversion and typical performance ratings of Extraversion ($r = .36$, $p < .05$) and a significant relationship between best choice scores and typical performance ratings of Agreeableness ($r = .43$, $p < .01$). However, for minorities there were no relationships between SJT scores and typical performance regardless of scoring method. Previous research has demonstrated that individuals tend to have different implicit theories for different racial and ethnic groups regarding the factors that constitute good performance (Wilson, 2010). Thus, it is possible that bias in the rating of typical performance may be causing these differential relationships. It is possible that the ratings for racial minorities are skewed, as the factors that raters utilize when predicting the performance of the minority participants are different than the factors utilized for predicting the performance of Caucasians. Additionally, there may be an interaction between the race of the rater and the race of the participant, such that minority raters are more lenient when rating minority participants and thus less accurate. Previous research has found some support for this possibility. Specifically, Mount, Sytsma, Hazucha, and Holt (1997) found that African American raters tended to rate others of their same race more highly than those of different races. The results for Caucasians were less clear, as racial biases were found for supervisor ratings, but not for subordinate ratings, and not for peer ratings when between subjects analyses were conducted. As peer ratings were used for this study, it is possible that this leniency and inaccuracy by minorities caused the discrepancy in the prediction of typical performance. The results of this study

demonstrated some support for this, as Caucasians were found to have significantly lower typical performance ratings of Agreeableness. However, when analyses were conducted to compare dyads of the same race versus dyads of different races, no significant differences in predictive validities were found. However, it is important to note that this may be cause by the small number of individuals who were comprised in some of the racial matched / unmatched subgroups. Upon examination of the means, it appears that being rated by someone of a different race will result in higher ratings of typical behavior. This may be indicative of some type of patronization effect, whereby individuals inflate ratings for other races. Research has previously found support for this effect occurring in opposite gender dyads (Vescio, Gervais, Snyder, & Hoover, 2005). Future research should further explore the interaction between rater and ratee race in the prediction of performance ratings.

It is important to note that upon examination of validity coefficients, different patters appear between African American and Hispanic participants. The n in these subgroups was relatively small, so the differences in validity coefficients were not significant with the exception of differences in correlations between best choice Agreeableness and GMA (African American r = .55, Hispanic r = -.2, *z-score* = .174, $p < .05$). Often, adverse impact, differential validity, and differential prediction is thought of in terms of Caucasians versus minorities. However, this pattern of results demonstrates differences between Hispanic and African American subgroups that should be explored in future studies. This issue is especially salient as there is a growing number of Hispanic citizens in the United States. These changing population demographics make it important to understand differences between subgroups, and reasons for this differential validity and differential prediction.

It was also found that race and gender interacted such that there were group differences between races in the inflation of typical performance ratings. Specifically, as previously mentioned, it was found that distance scores were higher for Caucasian females (female Agreeableness mean = 6.28, male Agreeableness mean = 5.66, female Extraversion mean = 6.83, male Extraversion mean = 5.42). Comparatively speaking, the distance scores for minority females were much more similar to males (female Agreeableness mean = 6.83, male Agreeableness mean = 6.68, female Extraversion mean = 6.72, male Extraversion mean = 6.42). This finding highlights that when examining bias in SJT scoring, it may be necessary to examine interactions between gender and race. There may be cultural and decision making differences for more specific subgroups than previously explored by research.

Race and Maximum/Typical Performance

Results demonstrated significant differences between racial groups with regard to the predictive validity of different scoring methods. To summarize these results, it was found that typical performance tended to be better predicted by SJT scores for Caucasians. Best choice Agreeableness scores and distance Extraversion scores were stronger predictors of typical performance for the matched trait. Exploration of the data demonstrated no significant differences in the dyads that would explain the differences in typical performance. In other words, it was found that most partners who were Caucasian tended to have Caucasian partners (75%) and many non-Caucasians tended to have non-Caucasian partners (53%). Additionally, although there were significant differences such that partners who were Caucasian were statistically more likely to have Caucasian partners (t = -2.36, p < .01), there were no significant differences in familiarity between these dyads. In other words, same-race dyads were not more familiar than were different-race dyads.

Although typical performance was better predicted by SJT scores for Caucasians, SJT scores tended to better predict maximum performance Extraversion ratings for non-Caucasians. Specifically, both distance and best choice scores were stronger predictors of maximum performance for non-Caucasians. To date, there has been little theorizing which has focused on racial differences in personality characteristics and partner rating (Foldes, Duehr, & Ones, 2008). One meta-analysis demonstrated African American/Caucasian differences such that whites tended to score consistently higher for both self-reported Agreeableness (d = .03) and Extraversion (d = .16). The same pattern was found for Hispanic/Caucasian differences, such that Caucasians tended to score higher for both self-reported Agreeableness (d = .05) and Extraversion (d = .02). Another study, however, has found that the type of criterion will have a strong moderating effect on African American/Caucasian differential validity. Specifically, it was found that using a predictive GMA measure, mean African American validity was lower relative to mean Caucasian validity when examining subjective criterion such as supervisory ratings, but that there were no differences in GMA tests in predicting objective criterion (Berry, Clark, & McClure, 2011). This might explain why typical performance, the subjective criterion in this study, found stronger relationships for Caucasians. However, the objective measure found the opposite pattern. Beyond racial differences, cultural differences have been hypothesized to influence personality; but unfortunately most of this research is conducted on the national level (Hofstede & McCrae, 2004). This makes it very difficult to explain the results found in this study. Further analysis of these results is necessary to aid in understanding how different races and cultures may process information differently, resulting in differences in scoring and prediction for SJT tests.

Practical Implications

Adverse Impact and Scoring Technique

This study highlights how important it is to explore the differential prediction and validity of a selection or assessment test. There were significant mean differences in the SJT scores, and regression analyses demonstrated significant interactions between race and scoring technique in predicting maximum and typical performance. This is very important to note, as a practitioner may not consider the effects on adverse impact when selecting a scoring technique, or when changing the scoring technique of an off-the-shelf assessment. While both scoring techniques demonstrated differential prediction, different scoring techniques were better predictors for Caucasians or minorities, depending on the criteria being measured. These results demonstrate that when administering a newly developed SJT, it may be beneficial to collect as much data as possible (e.g. best choice, worst choice, and effectiveness ratings) in order to determine which score will demonstrate the maximum prediction of criteria, and also to determine which score has the lowest level of adverse impact. Future research should explore the differential prediction of SJT scores measuring a multitude of traits to determine the effects of scoring technique on adverse impact.

Scoring Technique Specific to Dimensions

Results of this study demonstrate that different SJT scoring techniques better captured different dimensions. Specifically, it was found that best choice scores better predicted typical performance Agreeableness, while distance scores better predicted typical performance Extraversion. These results demonstrate that a practitioner may consider utilizing different scoring techniques on the same SJT, if the SJT is measuring multiple dimensions. Additionally, a practitioner should consider the dimension being measured when selecting an SJT scoring

technique. Additionally, as the issue in differential prediction of dimensions may have been an issue in the SJT item response script, practitioners should take care to carefully script different levels of item responses. Future research should address the optimal scoring technique for often-used criteria such as supervisor ratings of performance, customer service performance, job knowledge, and procedural knowledge. Future research should also explore if there are properties (e.g. observability) of Agreeableness or Extraversion that lend themselves to being best measured by specific scoring techniques.

Different Relationships with Criteria

While mean differences were found for typical performance with regards to gender and race, regression analyses determined maximum performance was better predicted by SJT scores for minorities, and typical performance was better predicted by SJT scores for Caucasians. Again, this highlights the necessity to take care when selecting a criterion in order to avoid adverse impact. Additionally, this demonstrates that using an off-the-self assessment tool to predict a related criterion may not necessarily ensure that the assessment tool will not demonstrate adverse impact, even if the assessment tool was initially demonstrated to have minimal adverse impact. In this study, the criteria used were the same (Agreeableness and Extraversion), however the means that this data was collected varied (partner ratings versus simulation). Again, this demonstrates how important it is for practitioners to constantly monitor levels of adverse impact, even if measuring the same criterion, if a new measurement technique is used. Future research should explore why these effects occur, and determine if the differential prediction occurs for other types of assessment tests when utilizing peer ratings or simulation scores as criteria.

Limitations

One limitation of this type of study is the population. This is a familiar problem to researchers who conduct studies in a university setting. In this situation, the population may have made finding results especially difficult. The population had few external motivators to perform well on the SJT or simulation. This will add additional variance due to motivation to the results. The SJT required several judgments regarding the effectiveness of different response items that were very subtly different. This required conscientiousness and attention to detail to complete accurately. Some students may have had difficulty maintaining focus through the entire SJT and simulation, as the research was two and a half hours long in its entirety. Additionally, the simulation was designed to be a measurement of maximum performance; however, it may have been difficult to ensure students were performing at a maximum level as there were no rewards for good performance.

Another limitation of this study was the potential criterion contamination which occurred by having overlap in the scenarios presented in the SJT and the scenarios presented in the simulation. As potential responses to the scenarios were provided in the SJT, it is possible that participants remembered particular item responses from the SJT, and provided similar responses to the scenarios in the simulation. This was unavoidable, as the scenarios in the simulation were expensive to develop and were impossible to interchange. Additionally, the procedure could not be changed to alter the order of presentation, as the SJT was the focus of this study. It could not be risked that exposure to the simulation would alter an individual's responses on the SJT. However, the fact that SJT scores did predict maximum performance Extraversion for minorities suggests that the maximum performance measure was capturing relevant variance.

Another issue may have been the high fidelity of the SJT. This may affect the external validity of these findings. The high fidelity of the test may have presented additional variance. While high fidelity testing is a growing market, it is not entirely clear how these results would generalize to a pen-and-paper test. However, there is no reason to believe the results would not be similar.

Unfortunately, for this study it was not possible to collect actual customer service performance data. This would have been the ideal criteria against which to validate the SJT and explore the effects of scoring technique. The criteria of partner ratings were indicators of typical behavior, but at the same time were difficult to utilize as partners had varying levels of familiarity with the participant. Additionally, some partners may have been unable to predict how extraverted or agreeable the participant may be in a customer service simulation, as the partner may have never witnessed the participant in a workplace scenario. However, it is important to note that supervisor or peer ratings are often used as criteria in the workplace. Additionally, in this study, we were able to measure characteristics of the rater (e.g., personality) and determine that significant relationships were found even after accounting for these variables.

Another limitation was the lack of discriminant validity within assessments of different traits. Certain scoring techniques tended have high correlations between the measurements of different traits. Additionally, typical performance tended to demonstrate strong correlations between ratings of Agreeableness and ratings of Extraversion. This is often a common issue in assessment centers, where the same dimensions often do not converge well across different exercises, but will tend to blur within the same exercise (Lievens, Chasteen, Day, & Christianson, 2006). It is possible there is a general factor captured by the SJT and typical performance (e.g. social intelligence) causing these high correlations.

Conclusion

This study adds to the existing literature in several ways. First, there are several implications for adverse impact, as results demonstrate that differential prediction and differential validity occurred for both scoring techniques. The results demonstrated that scoring techniques may operate differently depending upon the dimension being measured. Finally, this study demonstrates the importance of carefully selecting criteria, as significantly different results were found depending upon the criteria being utilized. It also seems possible from the results that different scoring techniques interact to incrementally add to the prediction of typical performance criteria. This study demonstrates interesting and unexpected findings for future studies to build upon.

An unanticipated result of this study was the extreme differential relationships found for race. SJT scores predicted different criteria for different races. This may be due to differences or biases caused by SJT scoring method, or due to racial differences in relationships to the criteria being used, as mean differences were found between Caucasians and minorities in typical performance ratings. These results should be replicated and explored in future studies in order to determine why these differences exist.

Another unexpected finding was the differences in scoring technique prediction for the two targeted traits. Distance scores predicted typical performance for Extraversion, but not for Agreeableness. Best choice scores predicted typical performance for Agreeableness, but not for Extraversion. These findings may be explained by scripting errors in the item responses, or may be caused by some factor inherent to the traits themselves (e.g. observability). Future studies should explore differential validity for SJT scoring method as a function of the dimension being measured.

Results also demonstrated significant differences for racial subgroups in the prediction of typical or maximum performance. Generally speaking, maximum performance tended to be predicted by SJT scores for minorities, while typical performance was predicted by SJT scores for Caucasians. These results indicate the importance of measuring differential prediction when considering criteria. Future research should explore the differential prediction of different SJT scoring techniques in relation to a multitude of criteria.

Current literature is benefited by further investigation of the differential effects of scoring technique. Often when constructing an SJT, very little thought is given to the scoring technique. SJTs have rarely been used to capture personality, and exploring how to best score SJTs in order to increase the correlations with personality criteria could be of great importance to the development of an implicit measurement of personality. The findings of this study replicate and extend prior research, which demonstrated correlations between SJT scores and self-reported personality using the ITP scoring method (Motowidlo, Hooper, & Jackson, 2006b).

In summary, the current study provides insight into the effects SJT scoring technique has on relationships with criteria of interest. Unexpected findings were demonstrated for differential prediction of race, differential prediction for dimensions, and differential prediction of criteria for SJT scoring techniques. Further, the best method of capturing personality through SJT scoring was explored in order to maximize the prediction of customer service and reduce adverse impact. Finally, this study made progress in answering questions regarding the theoretical framework of SJTs, and also raised new questions regarding the differential effects of scoring technique.

APPENDIX A: INITIAL ITEM POOL

Date: **Participant ID:**

Demographic Info

Demographics Form

Please answer the questions about yourself and your parents/guardians to the best of your knowledge. If you do not know the answer to the question or the question does not apply to you, please write "N/A" to indicate it is not applicable.

1. **How old are you? _____**

2. **What is your sex? (circle one)**
   a. **Male**
   b. **Female**

3. **What is your race or ethnic background? (check "yes" or "no" next to each race or ethnic group; if you choose "Other" as your response, please specify your race or ethnic group)**

Yes	No	
☐	☐	White (Non-Hispanic)
☐	☐	Black or African American (Non-Hispanic)
☐	☐	Asian
☐	☐	American Indian or Alaska Native
☐	☐	Native Hawaiian or Other Pacific Islander
☐	☐	Hispanic or Latino
☐	☐	Other: (Specify) _____

4. **If you chose more than one race or ethnic group in the previous question, which one do you most identify with?**
   a. White (Non-Hispanic)
   b. Black or African American (Non-Hispanic)
   c. Asian
   d. American Indian or Alaska Native
   e. Native Hawaiian or Other Pacific Islander
   f. Hispanic or Latino
   g. Other: (specify)_____

5. **What is your Mother's race or ethnicity? (circle all that apply; if you choose "Other" as your response, please specify your Mother's race or ethnic background)**
   a. White (Non-Hispanic)
   b. Black or African American (Non-Hispanic)
   c. Asian
   d. American Indian or Alaska Native
   e. Native Hawaiian or Other Pacific Islander
   f. Hispanic or Latino

g. Other: (specify)_____

6. **What is your Father's race or ethnicity? (circle all that apply; if you choose "Other" as your response, please specify your Father's race or ethnic background)**
   a. White (Non-Hispanic)
   b. Black or African American (Non-Hispanic)
   c. Asian
   d. American Indian or Alaska Native
   e. Native Hawaiian or Other Pacific Islander
   f. Hispanic or Latino
   g. Other: (specify)_____

7. **Where were you born? (City, State; Country if outside the US)**

   _____

8. **Please indicate if there is a country different from the country in which you were born that you identify with more or it has more cultural influence on you?**

   _____

9. **Where was your Mother born? (City, State; Country if outside the US)**

   _____

10. **Where was your Father born? (City, State; Country if outside the US)**

    _____

11. **Are you fluent in more than one language? If so, please list the languages in the order of which you are most fluent to least fluent.**

    _____

12. **What language does your mother speak? If she speaks more than one language, please list the languages in the order of which she is most fluent to least fluent.**

    _____

13. **What language does your father speak? If he speaks more than one language, please list the languages in the order of which he is most fluent to least fluent.**

    _____

14. **Have you ever lived in a country outside the US? (If your answer is "No", please skip to question 19)**
    a. **Yes**
    b. **No**

15. **If you have lived in a country outside the US,**

    where? _____

how long? _____

16. **Have you ever attended school outside of the US?**
    a. **Yes**
    b. **No**

17. **If you have attended schools outside the US,**

    which country/countries? _____

    during which grades? _____

18. **What is your highest level of education? (grade level or degree) _____**

19. **What is your Mother's highest level of education? (grade level or degree)**

    _____

20. **What is your Father's highest level of education? (grade level or degree)**

    _____

21) Do you have any customer service experience?

◯ **Yes** ◯ **No**

If yes, please give the following for each customer service job you have held:

Job title:	**Employer:**	**Years Experience:**
Job title:	**Employer:**	**Years Experience:**
Job title:	**Employer:**	**Years Experience:**

22) Do you have any work experience in a medical setting?

◯ **Yes** ◯ **No**

23) What is your current G.P.A.? _____

24) If you took the ACT, what was your score? _____

25) If you took the SAT, what was your score? _____

26) What is your middle name? _____

27) What is your major? _____

28)What is your year in school? (freshman, sophomore, junior, senior)?

_____

29) Where were you born (city and state)? _____

30) Do you work? If so, where? _____

31) When is your birthday (MMDDYY)? _____

32) What is your current favorite TV show? _____

33) How many times in your life have you been in an emergency room waiting area?

_____

Read each statement carefully. For each statement, on the scale from 1 – 6, please circle the response that best represents your opinion.
Circle **1** if you ***strongly disagree*** or the statement is definitely false. Circle **6** if you ***strongly agree*** or the statement is definitely true.
Circle only one response for each statement. Respond to all of the statements, making sure that you circle the correct response.

	Strongly Disagree				**Strongly Agree**
1. I am not a worrier.	1	2	3	4	5
2. I like to have a lot of people around me.	1	2	3	4	5
3. I don't like to waste time daydreaming.	1	2	3	4	5
4. I try to be courteous to everyone I meet.	1	2	3	4	5
5. I keep my belongings neat and clean.	1	2	3	4	5
6. I often feel inferior to others.	1	2	3	4	5
7. I laugh easily.	1	2	3	4	5
8. Once I find the right way to do something, I stick to it.	1	2	3	4	5
9. I often get into arguments with family and co-workers.	1	2	3	4	5
10. I am pretty good about pacing myself so as to get things done on time.	1	2	3	4	5
11. When I am under a great deal of stress, I sometimes feel like going to pieces.	1	2	3	4	5
12. I don't consider myself especially "light-hearted."	1	2	3	4	5
13. I am intrigued by the patterns found in art and nature.	1	2	3	4	5
14. It is likely that some people think I am selfish and egotistical.	1	2	3	4	5
15. I am not a very methodical person.	1	2	3	4	5
16. I rarely feel lonely or blue.	1	2	3	4	5
17. I really enjoy talking to people.	1	2	3	4	5

	Strongly Disagree				Strongly Agree
18. I believe that letting students hear controversial speakers can only confuse and mislead them.	1	2	3	4	5
19. I would rather cooperate with others than compete.	1	2	3	4	5
20. I try to conscientiously perform all the tasks assigned to me.	1	2	3	4	5
21. I often feel tense and jittery.	1	2	3	4	5
22. I like to be where the action is.	1	2	3	4	5
23. Poetry has little or no effect on me.	1	2	3	4	5
24. I tend to be cynical and skeptical of others' intentions.	1	2	3	4	5
25. I have a clear set of goals and work toward them in an orderly fashion.	1	2	3	4	5
26. Sometimes, I feel completely worthless.	1	2	3	4	5
27. I usually prefer to do things alone.	1	2	3	4	5
28. I often try new and foreign foods.	1	2	3	4	5
31. I believe that most people will take advantage of you if you let them.	1	2	3	4	5
30. I waste a lot of time before settling down to work.	1	2	3	4	5
31. I rarely feel fearful or anxious.	1	2	3	4	5
32. I often feel as if I am bursting with energy.	1	2	3	4	5
33. I seldom notice the moods or feelings that different environments produce.	1	2	3	4	5
34. It is likely that most of the people who know me like me.	1	2	3	4	5
35. I work hard to accomplish my goals.	1	2	3	4	5
36. I often get angry at the way people treat me.	1	2	3	4	5
37. I am a cheerful, high-spirited person.	1	2	3	4	5

	Strongly Disagree				Strongly Agree
38. I believe we should look to our religious authorities for decisions on moral issues.	1	2	3	4	5
39. It is likely that some people think of me as cold and calculating.	1	2	3	4	5
40. When I make a commitment, I can always be counted on to follow through.	1	2	3	4	5
41. When things go wrong, I often get discouraged and feel like giving up.	1	2	3	4	5
42. I would not be described as a cheerful optimist.	1	2	3	4	5
43. Sometimes when I am reading poetry or looking at a work of art, I feel a chill or wave of excitement.	1	2	3	4	5
44. I am hard-headed and tough-minded in my attitude.	1	2	3	4	5
45. Sometimes I am not as dependable or reliable as he/she should be.	1	2	3	4	5
46. I am seldom sad or depressed.	1	2	3	4	5
47. My life is fast-paced.	1	2	3	4	5
48. I have little interest in speculating on the nature of the universe or the human condition.	1	2	3	4	5
49. I generally try to be thoughtful and considerate.	1	2	3	4	5
50. I am a productive person who always gets the job done.	1	2	3	4	5
51. I often feel helpless and want someone else to solve my problems.	1	2	3	4	5
52. I am a very active person.	1	2	3	4	5
53. I have a lot of intellectual curiosity.	1	2	3	4	5
54. If I don't like a person, I let him/her know it.	1	2	3	4	5
55. I never seem to be able to get organized.	1	2	3	4	5

	Strongly Disagree				Strongly Agree
56. At times, I have been so ashamed I just wanted to hide.	1	2	3	4	5
57. I would rather go my own way than be a leader of others.	1	2	3	4	5
58. I often enjoy playing with theories or abstract ideas.	1	2	3	4	5
59. If necessary, I am willing to manipulate people to get what I want.	1	2	3	4	5
60. I strive for excellence in everything I do.	1	2	3	4	5

	Strongly Disagree				Strongly Agree
1.I can predict other peoples' behavior.	1	2	3	4	5
2. I often feel that it is difficult to understand others' choices	1	2	3	4	5
3. I know how my actions will make others feel.	1	2	3	4	5
4. I often feel uncertain around new people who I don't know.	1	2	3	4	5
5. People often surprise me with the things they do.	1	2	3	4	5
6. I understand other people's feelings.	1	2	3	4	5
7. I fit in easily in social situations.	1	2	3	4	5
8. Other people become angry with me without me being able to explain why.	1	2	3	4	5
9. I understand others' wishes.	1	2	3	4	5
10. I am good at entering new situations and meeting people for the first time.	1	2	3	4	5
11. It seems as though people are often angry or irritated with me when I say what I think.	1	2	3	4	5
12. I have a hard time getting along with other people.	1	2	3	4	5
13. I find people unpredictable.	1	2	3	4	5
14. I can often understand what others are trying to accomplish without the need for them to say anything.	1	2	3	4	5
15. It takes a long time for me to get to know others well.	1	2	3	4	5
16. I have often hurt others without realizing it.	1	2	3	4	5
17. I can predict how others will react to my behavior.	1	2	3	4	5
18. I am good at getting on good terms with new people.	1	2	3	4	5
19. I can often understand what others really mean through their expression, body language, etc.	1	2	3	4	5
20. I frequently have problems finding good conversation topics.	1	2	3	4	5
21. I am often surprised by others' reactions to what I do.	1	2	3	4	5

The following questions concern your familiarity with your experimental partner.

1. How would you describe your relationship with this person?
    a. Relative
    b. Close friend
    c. Acquaintance (e.g. classmate, neighbor)
    d. Roommate
    e. Coworker
    f. Significant other (husband/wife/fiancée; boyfriend/girlfriend)

2. How long would you say you have known this individual? _____

3. On average over the past 6 months, I have interacted with this person:
    a. Almost everyday
    b. More than once a week
    c. About once a week
    d. Less than once a week

4. In the time since we first met, our most frequent level of interaction was:
    a. Almost everyday
    b. More than once a week
    c. Once a week
    d. Less than once a week

5. How often have you observed this person in the following contexts (circle one)?

    a. Interacting with co-workers at work
        1 = never
        2 = only once
        3 = more than once, I would say approximately ____ times
        4 = more times than I can count

    b. Interacting with an authority figure at work (e.g., supervisor, team leader)
        1 = never
        2 = only once
        3 = more than once, I would say approximately ____ times
        4 = more times than I can count

    c. Interacting with professors or instructors at school
        1 = never
        2 = only once
        3 = more than once, I would say approximately ____ times
        4 = more times than I can count

    d. Interacting with other students in class

1 = never
2 = only once
3 = more than once, I would say approximately ____ times
4 = more times than I can count

e. Interacting with you one-on-one
1 = never
2 = only once
3 = more than once, I would say approximately ____ times
4 = more times than I can count

f. Interacting in a group social setting
1 = never
2 = only once
3 = more than once, I would say approximately ____ times
4 = more times than I can count

g. Interacting with his/her family or significant other
1 = never
2 = only once
3 = more than once, I would say approximately ____ times
4 = more times than I can count

h. Interacting with strangers
1 = never
2 = only once
3 = more than once, I would say approximately ____ times
4 = more times than I can count

6. To the best of your knowledge, please answer the following about the peer you selected:

a. What is his/her middle name: _____

b. What is his/her major:_____

c. What is his/her year in school (fr,soph,junior,sr):_____

d. Where was he/she born (state and city)? _____

e. Does he/she currently work? And if so, where?_____

f. What is his/her birthday (MMDDYY?_____

g. What is currently his/her favorite TV show? _____

h. How many times have they been in a E.R. waiting room? _____

Read each statement carefully. For each statement, on the scale from 1 – 5, please circle the response that best represents your opinion.
Circle **1** if you ***strongly disagree*** or the statement is definitely false. Circle **5** if you ***strongly agree*** or the statement is definitely true.
Circle only one response for each statement. Respond to all of the statements, making sure that you circle the correct response.

	Strongly Disagree				**Strongly Agree**
1. My partner is not a worrier.	1	2	3	4	5
2. My partner likes to have a lot of people around them.	1	2	3	4	5
3. My partner doesn't like to waste time daydreaming.	1	2	3	4	5
4. My partner tries to be courteous to everyone they meet.	1	2	3	4	5
5. My partner keeps my belongings neat and clean.	1	2	3	4	5
6. My partner often feels inferior to others.	1	2	3	4	5
7. My partner laughs easily.	1	2	3	4	5
8. Once my partner finds the right way to do something, they stick to it.	1	2	3	4	5
9. My partner often gets into arguments with family and co-workers.	1	2	3	4	5
10. My partner is pretty good about pacing themselves so as to get things done on time.	1	2	3	4	5
11. When my partner is under a great deal of stress, they sometimes feel like going to pieces.	1	2	3	4	5
12. My partner doesn't consider themselves especially "light-hearted."	1	2	3	4	5
13. My partner is intrigued by the patterns found in art and nature.	1	2	3	4	5
14. It is likely that some people think my partner is selfish and egotistical.	1	2	3	4	5
15. My partner is not a very methodical person.	1	2	3	4	5
16. My partner rarely feels lonely or blue.	1	2	3	4	5
17. My partner really enjoys talking to people.	1	2	3	4	5
18. My partner believes that letting students hear controversial speakers can only confuse and mislead them.	1	2	3	4	5

	Strongly Disagree				Strongly Agree
19. My partner would rather cooperate with others than compete.	1	2	3	4	5
20. My partner tries to conscientiously perform all the tasks assigned to them.	1	2	3	4	5
21. My partner often feels tense and jittery.	1	2	3	4	5
22. My partner likes to be where the action is.	1	2	3	4	5
23. Poetry has little or no effect on my partner.	1	2	3	4	5
24. My partner tends to be cynical and skeptical of others' intentions.	1	2	3	4	5
25. My partner has a clear set of goals and works toward them in an orderly fashion.	1	2	3	4	5
26. Sometimes, my partner feels completely worthless.	1	2	3	4	5
27. My partner usually prefers to do things alone.	1	2	3	4	5
28. My partner often tries new and foreign foods.	1	2	3	4	5
29. My partner believes that most people will take advantage of you if you let them.	1	2	3	4	5
30. My partner wastes a lot of time before settling down to Work.	1	2	3	4	5
31. My partner rarely feels fearful or anxious.	1	2	3	4	5
32. My partner often feels as if they are bursting with energy.	1	2	3	4	5
33. My partner seldom notices the moods or feelings that different environments produce.	1	2	3	4	5
34. It is likely that most of the people who know my partner like them.	1	2	3	4	5
35. My partner works hard to accomplish their goals.	1	2	3	4	5
36. My partner often gets angry at the way people treat them.	1	2	3	4	5
37. My partner is a cheerful, high-spirited person.	1	2	3	4	5
38. My partner believes we should look to our religious authorities for decisions on moral issues.	1	2	3	4	5

	Strongly Disagree				Strongly Agree
39. It is likely that some people think of my partner as cold and calculating.	1	2	3	4	5
40. When my partner makes a commitment, they can always be counted on to follow through.	1	2	3	4	5
41. When things go wrong, my partner often gets discouraged and feels like giving up.	1	2	3	4	5
42. My partner would not be described as a cheerful optimist.	1	2	3	4	5
43. Sometimes when my partner is reading poetry or looking at a work of art, they feel a chill or wave of excitement.	1	2	3	4	5
44. My partner is hard-headed and tough-minded in their attitude.	1	2	3	4	5
45. Sometimes my partner not as dependable or reliable as they should be.	1	2	3	4	5
46. My partner is seldom sad or depressed.	1	2	3	4	5
47. My partner's life is fast-paced.	1	2	3	4	5
48. My partner has little interest in speculating on the nature of the universe or the human condition.	1	2	3	4	5
49. My partner generally tries to be thoughtful and considerate.	1	2	3	4	5
50. My partner is a productive person who always gets the job done.	1	2	3	4	5
51. My partner often feels helpless and wants someone else to solve their problems.	1	2	3	4	5
52. My partner is a very active person.	1	2	3	4	5
53. My partner has a lot of intellectual curiosity.	1	2	3	4	5
54. If my partner doesn't like a person, they let him/her know it.	1	2	3	4	5
55. My partner never seems to be able to get organized.	1	2	3	4	5
56. At times, my partner has been so ashamed they just wanted to hide.	1	2	3	4	5

	Strongly Disagree				Strongly Agree
57. My partner would rather go their own way than be a leader of others.	1	2	3	4	5
58. My partner often enjoys playing with theories or abstract ideas.	1	2	3	4	5
59. If necessary, my partner is willing to manipulate people to get what they want.	1	2	3	4	5
60. My partner strives for excellence in everything they do.	1	2	3	4	5

	Strongly Disagree					Strongly Agree
1.My partner can predict other peoples' behavior.	1	2	3	4	5	6
2. My partner often feel that it is difficult to understand others' choices	1	2	3	4	5	6
3. My partner knows how their actions will make others feel.	1	2	3	4	5	6
4. My partner often feel uncertain around new people who he or she doesn't know.	1	2	3	4	5	6
5. My partner is often surprised with the things people do.	1	2	3	4	5	6
6. My partner understands other people's feelings.	1	2	3	4	5	6
7. My partner fits in easily in social situations.	1	2	3	4	5	6
8. Other people become angry with my partner without my partner being able to explain why.	1	2	3	4	5	6
9. My partner understands others' wishes.	1	2	3	4	5	6
10. My partner is good at entering new situations and meeting people for the first time.	1	2	3	4	5	6
11. It seems as though people are often angry or irritated with my partner when my partner says what they think.	1	2	3	4	5	6
12. My partner has a hard time getting along with other people.	1	2	3	4	5	6
13. My partner finds people unpredictable.	1	2	3	4	5	6
14. My partner can often understand what others are trying to accomplish without the need for them to say anything.	1	2	3	4	5	6
15. It takes a long time for my partner to get to know others well.	1	2	3	4	5	6
16. My partner has often hurt others without realizing it.	1	2	3	4	5	6
17. My partner can predict how others will react to their behavior.	1	2	3	4	5	6
18. My partner is good at getting on good terms with new people.	1	2	3	4	5	6
19. My partner can often understand what others really mean through their expression, body language, etc.	1	2	3	4	5	6

20. My partner frequently has problems finding good conversation topics.	1	2	3	4	5	6
21 My partner is often surprised by others' reactions to what they do.	1	2	3	4	5	6

APPENDIX B: EXAMPLE WONDERLIC QUESTION

Note: The Wonderlic Research Donation Program Coordinator advised that the actual Wonderlic Questionnaire could not be included in its entirety, nor could an actual question from the test be reproduced in a dissertation. As such, an example question from the Wonderlic website is used.

Example Question: An instrument store gives a 10% discount to all students off the original cost of an instrument. During a back to school sale an additional 15% is taken off the discounted price. Julie, a student at the local high school, purchases a flute for $306. How much did it originally cost?

A. $325

B. $375

C. $400

D. $408

E. $425

APPENDIX C: SITUATION JUDGMENT TEST TRANSCRIPT

Event 1 (Extraversion) (Simulation Event 1)

Rick: Am I glad to see you. We have been swamped all afternoon and I so need a break. Kelly called at least three times for you, I don't know what it was about but she sounded pretty freaked out about something, but of course she didn't want to leave a message with me so I let it ring through to your voicemail. So hurry up and check your messages so I can get out of here.

Voicemail 1

Kelly: Hey it's Kelly umm I overslept again, I really hate to ask you to do this but do you think you can clock me in? They said if I am late one more time I'll get put on probation, I mean I should get there within five minutes anyway, I'm sure no one will ever notice that I wasn't there. Thank you so much, bye.

Voicemail 2

Kelly: Hey it's me again I hit really bad traffic on the highway, it looks like I'm going to be more like ten minutes late please call me back as soon as you can and let me what time you clocked me in, I just want to make sure we have our stories straight when Lynn asks us.

a) Kelly, I'm sorry. I'm not going to clock you in, because it is against company policy. I'm sure our manager will understand that you were in traffic. *(3)*
b) I don't think it's a good idea for me to clock you in. *(1)*
c) I don't think I should do that for you. Um, clocking you in is going against our company policy, and I might get in trouble. I'm sorry. *(2)*
d) Hey Kelly, I'm really sorry. Getting stuck in traffic is unlucky, but I can't clock you in because it's against company policy. I'm sure our manager will understand that you were in traffic. *(4)*

e) Hey Kelly! Listen, I'm really sorry, but I'm not going to be able to clock you because of company policy. But, getting stuck in traffic is an understandable reason to be late. I'm sure our supervisor will take that into account. Good luck! *(5)*

Event 2 – (Agreeableness) (Simulation Event 2)

Kelly: Oh my gosh it has been the craziest morning, I am so sorry I am late, did Louis or Lynn leave me any special messages for me today?

Rick: Uh yeah, Louis said that umm you need to start wearing tighter shirts to work.

Kelly: No, really.

Rick: Actually he just said that we need to start taking breaks as early in the shift as we can.

Kelly: Okay, this is what I need to do, I want to tell Lynn that I was here at eight but I just forgot to punch in, so if you guys will cover for me, I think I can pull it off, is that a plan?

Rick: Well that depends, what's it worth to you?

Kelly: Can I count on you?

a) I'm sorry no, I won't be able to cover for you. It is not fair to everyone else who makes sure to get here on time. I'm offended that you would expect me to do that for you. *(2)*
b) No, I won't cover for you. Others manage to make it here on time every day, so it is very unfair to those who put forth the effort. You are late, and I do not appreciate that you would ask me to break the rules and do that for you. *(1)*
c) Sorry Kelly, I can't do that. I know it's difficult to be running late, and I feel sorry for you, but I just can't do that for you. Just try to get here as fast as you can. Is there anything else I can do to help? *(4)*

d) I'm so sorry Kelly, I understand how difficult it is to be running late, it happens to everyone and I feel badly that you have to deal with that. I wish I could help, but unfortunately I can't really do that. *(5)*

e) I'm sorry but I can't Kelly, I know it is stressful to be running late, but none-the-less, please never ask me to do anything like this again, it puts me in a difficult situation. *(3)*

Event 3 – (Agreeableness) (Simulation Event 4)

Hispanic Lady: Hi, I'm really sorry he snapped at you like that, It's just he has been here for three hours and he is not getting any help. Um, can you at least send us back into the next room, even if he is not seen any faster, he will think he is getting closer and he will calm down, can you do that for us?

a) I'm sorry ma'am. I know how difficult it is, but we are very busy right now and doing the best we can. We have to follow hospital rules. *(4)*

b) I'm sorry for the wait ma'am. I know how hard it is to be kept waiting. Although I understand your frustration, I'm afraid I cannot make an exception to the rules. *(5)*

c) I'm sorry ma'am, you're just going to have to tell him we'll get to him when we can. The waiting room is swamped right now and my hands are tied because of hospital rules. *(3)*

d) No ma'am, you're just going to have to tell him to deal with the wait. It doesn't make sense that you should get preferential treatment over other patients who are waiting, I can't break the hospital rules. *(1)*

e) Ma'am, he will have to wait just like everyone else. Those are the hospital rules. I can't make an exception for anyone, even if it is just to let him into a back room to calm down. *(2)*

Voicemail from Louis

Louis: This is Louis; I have just received word that we've got a large group of children coming in who were involved in a bus accident so you can expect a crowd of anxious parents to fill the waiting room very shortly. We're not sure how many exactly, some of the children will probably be sent to the Central Park Hospital because of their burn unit. Please tell any inquiring family and friends that we don't have any information at this point on how many patients that will be coming or which patients those will be. I will be emailing you lists of names as I receive them, also please tell the parents to stay in the lobby so they don't miss the updates as they come in, we don't want to have to go searching for them every time there's an update. Thanks, bye.

Event 4 – (Extraversion) (Simulation Event 5)

Kelly: Hey, could you please make an announcement over the PA system regarding the bus accident. These families are dying for any information they can get.

a) Attention everyone; we don't know how many children from the bus accident are going to arrive. Some children might be sent to another hospital. Um, please stay in the lobby area to hear any additional updates as we get additional information. *(2)*
b) Attention friends and family from the school bus accident; we currently don't know how many children from the bus accident will be arriving. Some children might be sent to another hospital. I understand this is a difficult time. Please stay in the lobby area to hear updates. Thank you. *(4)*

c) Attention friends and family from the school bus accident, this is your customer service representative. At the present time, we don't know how many children will be arriving to this hospital from that particular accident, and some children may be sent to another hospital. We will try to provide you with information as soon as it's available, so please remain in the lobby area. Thank you. . *(5)*

d) Attention friends and family; we do not know how many children from the bus accident will be arriving at the hospital. Some children might be sent to another hospital. Please stay in the lobby area to hear updates. Thank you all for your patience. *(3)*

e) We don't have a lot of information about the children from the bus accident. We think some children might be sent to another hospital. We'll probably get more information later so stay in the lobby to hear updates. *(1)*

Interlude

Voicemail from Lynn

Lynn: Louis told us in our morning meeting that from now on we should try to get our breaks over with as early in the shift as possible. So as soon as you get this message I want one of you to go ahead and take your break. Give me a call back and let me know what time the first person leaves so I can give Louis an estimate of when all three breaks will be out of the way.

Event 5 – (Agreeableness) (Simulation Event 7)

Kelly: Hey Rick just asked me whether I minded if he took his break, even though we are really busy right now and can't really afford to have him leave, I felt like I couldn't say no because it's

pretty much my fault that he hasn't had one yet. Maybe when he comes up here and asks you, you could convince him that it'd be better if he waits. Here he comes now.

Rick: Okay, I am ready for my break. Now Lynn told me to go ahead and leave but Tanya was all worried that you and Kelly might need me to stay. Now I already talked to Kelly and she's alright with it, but Tanya wants to hear directly from you. Look, I have been here since early this morning and haven't even eaten lunch yet, so could you hurry up and call her, I'm starving.

a) We need you here, Rick. Everyone else is working hard without a break. You'll have to wait longer, it's just part of the job. I'm afraid that you'll have to make due for the time being. *(1)*
b) Rick, you're absolutely right; it's not fair for you to go this long without a break, and I understand how difficult it can be to go so long without taking one, but unfortunately we have too many patients right now. Would you like me to page you when the lobby begins to clear out? *(5)*
c) We're swamped right now Rick; although I know you could use a break you are needed here to help with these patients to keep the hospital running smoothly. *(3)*
d) Rick, you're right, I understand it has been a while, but I'm afraid we are going to need you here attending to patients, because the waiting room is completely swamped. *(4)*
e) Rick, it's very busy right now. There are many people waiting to be assisted, and we can't afford to be without you, the waiting room is entirely too backed up. *(2)*

Event 6 – (Agreeableness) (Simulation Event 8)

White lady in blue shirt: Excuse me, our son was in a bus accident, his name is Michael Rayfield, we've been waiting here for thirty minutes and were not going to wait any longer, now is somebody going to take us back to see him or do we need to go back there ourselves.

a) Sorry about your wait, often patients do get frustrated but I'm afraid I can't take you back there, and that I don't have any information about your son as of yet. *(4)*
b) I'm sorry that you've been waiting that long, and I understand how difficult it is to be kept waiting. I wish I could take you back there, but I will update you as soon as I know anything. *(5)*
c) Sorry about the wait, but there is nothing I can do, I haven't been updated recently and I have no information regarding your son. I'm afraid you will have to wait a bit longer. *(3)*
d) No, You will have to continue to wait, just like everybody else is waiting. I can't grant you special privileges, as there are several concerned people in the waiting room. I can't just let you back there. *(1)*
e) You will have to wait a little longer, I can't let you back there because I haven't heard anything regarding your son. There are several people who have been waiting longer than you. *(2)*

Event 7 – (Extraversion) (Simulation Event 10)

Intercom (Mr. and Mrs. Rayfield, please meet Dr. Jones in the triage lobby)

Lynn: Didn't you get my message about taking your breaks? Why are all three of you still here?

a) I don't think we should go on break when we're this busy. *(1)*

b) Hey Lynn, I hope I didn't cause you too much concern. We've had so many people in the lobby that we couldn't afford someone taking a break right now. Hopefully I can send someone soon. *(4)*

c) I didn't think we should go on break. Look how full the lobby is, there are a lot of people here. Um, I think we should wait a little until it clears out. *(2)*

d) Lynn – no, I didn't send anyone on their breaks. The lobby was full and I thought it would be in the hospital's best interest if we all worked. Hopefully I can remedy the situation shortly. *(3)*

e) Hey Lynn, I know you're concerned. However, the lobby was full and I needed everyone to work, so I chose not to send anyone at the time. How about I send someone on break when the lobby clears? (5)

Interlude

Lynn: Carl, I heard you were here, what's going on?

White lady in green tank top (Christine): Hey can somebody please see my boyfriend?

Lynn: Christine, you know he is just going through withdrawal like he does once a month when the two of you run out of drugs, don't worry, is he conscious?

Christine: yes

Lynn: well then he is fine as long as he is conscious, let me know if he loses consciousness.

Christine's boyfriend: Christine, come back.

Event 8 – (Extraversion) (Simulation Event 14)

Christine: Hey, can you help me out, that nurse has no heart and she doesn't care what happens to Andrew. Look if you can just have her come back out here and tell her that he has lost

consciousness, then he'll be sure to be seen. She can't prove he didn't lose consciousness when she was putting a Band-Aid on that kids skinned up knee. So if we both say that he has lost consciousness then he'll get his treatment, I mean I know that because that's what she said. So please can you just please help me out?

a) I can't do that, I could get in trouble. *(1)*
b) I'm sorry, um, I don't think that it's a good idea for me to do that for you, I could get in a lot of trouble. *(2)*
c) Miss, I really wish I could help, but I could get in quite a bit of trouble should I be caught doing that for you. I hope you can find another solution! *(4)*
d) Miss, I really wish I could help. While it is very difficult to see your boyfriend like this, I'm afraid that I can't do that or I would get in a lot of trouble. Hopefully everything works out for you! *(5)*
e) Miss, you seem to be very concerned, but I won't be able to do that for you. I could get in a lot of trouble. *(3)*

Event 9 – (Extraversion) (Simulation Event 18)

Black man in maroon shirt (Carl Johnson): Excuse me, my name is Carl Johnson, I'm the husband of Kayla Johnson, she came in earlier with a pregnancy related problem, can you tell me where she is?

Rick: Sir, it's like I told you once before, that name is not in our system.

Carl: Excuse me punk, I wasn't talking to you, look I know my wife came in here because I saw her car in the parking lot, now she is pregnant with my son and I have a right to know how she is

doing. My brother-in-law is a lawyer so if you don't tell me how she is doing your going to be sorry, now are you going to tell me where she is?

a) We don't have that name in our database. I can't provide you any more information. *(1)*
b) Mr. Johnson, I'm afraid her name is not in the database. I know you must be frustrated, and I wish I could help you further. I'd be happy to assist you in any other way. *(5)*
c) Sir, I'm sorry, but at the moment we don't have that name. Although I wish I could help, I'm afraid there is nothing I can do for you. *(3)*
d) I'm not sure we have anyone by that name right now. I wish I could give you more information. *(2)*
e) Mr. Johnson, I'm afraid her name is not in the database. I wish I could help -- let me know if there is any other way in which I can assist you. *(4)*

Interlude

To: C.S Rep

Subject: Rick

Hey, can you believe some of the stuff that comes out of Rick's mouth? I am sick and tired of the way he talks to me! I just mentioned something about it to Lynn---and she just blew me off---like I was overreacting or something. Am I crazy or was he just way out of line earlier?

Lynn's reaction made me so mad that I went to talk to Louis about it too. He tried to make excuses for Rick like he always does---anything to avoid conflict (what a wimp). He tried to tell me that Rick really doesn't mean to offend me and that he probably has no idea that his comments are out of line (yeah right). You saw yourself today how obvious I was today about

letting him know if he couldn't tell by my reaction that I was offended just now, he's just clueless. How can I be any more obvious than that? Don't you agree?

---Kelly

Event 10 – (Agreeableness) (Simulation Event 25)

Lynn: Hey Rick tell me what has been going on between you and Kelly?

Rick: Lynn I don't know what her problem is, I think she secretly has a thing for me, and she is just upset I'm not into her. Look there is nothing that I said to her that would have offended anybody else. Like just today, I said is that a mirror in your pants because I think I can see myself in them.

Lynn: Oh Rick, okay I will try and talk her out of continuing with this.

Rick: yeah please do, I don't want any more problems.

Intercom: (Rick you are needed in the ER)

Rick: Got to go, that's me

Lynn: That Kelly is such a whiner, she is just not cut out for this job, don't you agree?

a) Kelly may exaggerate things at times; you are right about that Lynn. I believe she may be difficult occasionally -- but I'm afraid Rick was inappropriate. Let me know if you need any more information from me. *(4)*
b) No Lynn, I believe you are wrong. Rick was incredibly inappropriate and I support Kelly filing a report. Despite the fact that Kelly may be difficult at times, Rick was out of line. *(1)*
c) Kelly does seem to be bothered by issues that may not bother others; you are absolutely right about that Lynn. I understand she can be difficult to interact with. But Rick did make an offensive comment. Is there any way I can help resolve this? *(5)*

d) Lynn, even if you think Kelly is making a big deal, which may indeed be the case, the fact of the matter is that what Rick said was inappropriate. *(3)*

e) Lynn, I don't agree with what you are saying. Rick was being inappropriate and should be reprimanded, even if Kelly can be difficult at times. *(2)*

Event 11 – (Extraversion) (Simulation Event 27)

Voicemail from Rick

Hey it's Rick, Lynn says that if you tell Louis that you think Kelly is blowing this whole thing out of proportion he will talk to her about dropping the issue. So all I need you to say is that you don't think I said anything wrong. How about it? Call me back right away.

a) Hey Rick, this definitely seems to be a difficult situation. I wish you didn't have to deal with it. Unfortunately, I do believe your comment was somewhat inappropriate so I am not going to be able to say otherwise. *(5)*

b) Rick, in general I'm going to try not to get involved, but if the supervisors ask me, I'll have to tell them the truth. *(2)*

c) I'm afraid I'd rather not get involved with this situation. You might have to handle this on your own. *(1)*

d) Rick, this is a tough situation, but I'm going to have to give my honest opinion of what happened. I'm going to tell the supervisors what I heard if they ask. *(3)*

e) Rick, you do seem to be in a difficult situation, but if I am asked I'll have to give my honest opinion and I do believe your statement was somewhat inappropriate. *(4)*

Interlude

To: C.S. Rep

Subject: Disciplinary action against you

I am writing this email to inform you that I have submitted a disciplinary action against you for leaving the customer service desk for an extended period of time. As you know you missed several critical phone calls regarding a critical patient, as a result I need to know if you are willing to accept responsibility and acknowledge this in writing. As you know, employees shown to give false statements even once are subject to immediate dismissal from their jobs. In the event that you are unwilling to sign the formal disciplinary form the matter will go before a disciplinary board. I need you to reply to this email with your response as soon as possible.

---Lynn

Event 12 – (Agreeableness) (Simulation Event 31)

Voicemail from Louis

Hey this is Louis I need to talk to about that email from Lynn. She just wanted me to fire you on the spot. I convinced her to go with route of just writing you up for disciplinary action. Bottom line is it's your word against hers. You may as well just eat this one. It could turn into a much bigger deal, you know Lynn's temper. What do you say? Just call me back and let me know what you want to do.

a) Louis, what Lynn is saying simply isn't true. I haven't left my desk all day. I don't understand what is happening but I'm not going to sign a document which is inaccurate. *(2)*
b) I do not agree with Lynn. I do not believe I left my desk and I will try to talk to her. I'm afraid that I am not willing to sign something that is not correct. *(3)*

c) I didn't do anything wrong Louis. I will not be signing anything that suggests I did. It seems unethical of you to pressure me to sign something while at the same time admitting that it may not be accurate. *(1)*

d) I think there may be something wrong with the phones Louis. We need to communicate that to Lynn. I don't want to put you in an awkward situation; I know it must be difficult. However, I'm afraid I don't feel right signing a document which may not be accurate. *(4)*

e) You're right Louis -- this is not worth arguing about. I understand you are in a difficult situation, and I feel sorry for you. But I'm still afraid I don't feel right signing a document that states that I left my desk *(5)*

APPENDIX D: SJT ANSWER SHEET

Situational Judgment Test Answer Sheet:

On your answer sheet please rate each response item in terms of the items effectiveness, with **"1" signifying a highly ineffective** response and "**10" signifying a highly effective response.** Note that you can assign the same rating score to multiple responses, if you feel that certain responses are equally effective. After rating each item, please make a check mark on the far left side of your answer sheet in relation which response you feel would be the "best" response and the "worst" response to the situation on the corresponding line.

Event 1:

Best Response: Worst Response:

_____	_____	**A.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
_____	_____	**B.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
_____	_____	**C.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
_____	_____	**D.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
_____	_____	**E.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Event 2:

Best Response: Worst Response:

_____	_____	**A.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
_____	_____	**B.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
_____	_____	**C.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
_____	_____	**D.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
_____	_____	**E.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

APPENDIX E: PARTNER RATING SLIDES

Partner Behavior Ratings
Team Training and Workforce Development Lab
Partner Rating Task

- The Partner Rating Task will present you with scenarios focused on tasks and behaviors regarding situations in the workplace. Each scenario describes a work-related situation and requires you to choose a course of action by responding to multiple-choice questions. For there is a scale on which you will rate how you believe your partner would handle the situation.

Partner Rating Task

- It is important to remember you will be rating tasks as to how your partner would respond to them. Please take time to remember how your partner has responded to situations previously and base your ratings on behaviors you have seen from your partner in the past.

Traits -- Introduction

- During the Partner Rating Task, you will be asked to rate your partner on three specific traits with regard to how they would handle the situation. These three tasks will be explained to you at the beginning of the task, and which will be listed on each slide. These traits are Agreeableness, Extraversion, and Emotional Stability.

Emergency Room Introduction

- For this experiment you are to imagine that your partner is a Customer Service Representative behind the front desk in an emergency room lobby. Customers and coworkers will present scenarios throughout the experiment via video clips, e-mails, and voicemails.

Emergency Room Introduction

- Please click on each icon on each slide to view each video clip or hear each audio clip. Note that not every slide has an audio or video clip, but some slides may contain both an audio clip and a video clip. Also, not every slide requires an answer to be provided. Directions will specify when and when not to provide a response for each event. Note that you can only view each video clip or hear each audio clip <u>ONLY ONCE!!</u>

Emergency Room Introduction

- To the right is an example of the video clips which you will see throughout the experiment
- To play the video clip when it appears, click on the “Play” button which is the middle button on the bottom of the screen
- To close the video clip after the clip has ended, click on the “X” in the upper right hand corner of the clip.

Emergency Room Introduction

- After clicking on each video clip icon a "Yes or No" response will be prompted to which you are to click on "Yes" in order to play the clip.
- Example of video clip icon: Please click on the video icon, and then click "Yes" to play the icon. After playing the clip of the man blinking, please close out the clip.
- Example of audio clip icon: Please double click on the audio icon to hear a party horn.

Emergency Room Introduction

- After you view each video clip or listen to each audio clip for their corresponding event you will be presented with a rating scale that goes from one to five. On one end of the scale is an example response that is very typical of a particular trait. For example, if the trait we are looking at is Extraversion, the scale will give an example of an extraverted response that would be typical of a high, or "five," rating for Extraversion. The scale will also give an example of a response that would be typical of a low, or "one" rating for Extraversion. You are to rate who your partner would respond, on a scale from one to five. It is important to remember, while only two examples will be provided, one on each end of the scale, the rating scale allows you to choose any number from one to five.

Emergency Room Introduction

- For example, after you are presented a scenario where someone is trying to get into the back room of the hospital, and the trait you are rating your partner is agreeableness. You will then be presented with this scale:

Emergency Room Introduction

- As you can see from the scale on the previous page, there are examples provided for each end of the scale. However, these represent the ends of the scale, and you should merely use these as a guide for where the ends of the scale are, and respond with a rating anywhere from one to five, with five being the best example of the trait in question.

Traits --- Agreeableness

- As previously explained, you will be rating how much of a particular trait your partner would portray when responding to a situation. The first of these traits is Agreeableness. Agreeableness is a tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others. Agreeable people tend to be considerate, willing to compromise, conforming, empathetic, and loyal. They tend to be very good at getting along with others, pacifying individuals, and resolving conflict.

Introduction

- You have now completed the introduction to the Partner Rating Task. Please proceed to the next slide to begin the Partner Rating Task.

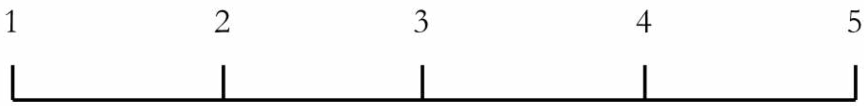
Event 1

- Click on the icon below to view a video clip of the event. After viewing the clip please proceed to the next slide.

Event 1 -- Agreeableness

- After examining the responses below on either end of the scale, rate on a scale of one to five as to how your partner would respond to the previous situation.

Agreeable people tend to be considerate, willing to compromise, conforming, empathetic, and loyal. They tend to be very good at getting along with others, pacifying individuals, and resolving conflict.



No, I won't cover for you. Others manage to make it here on time every day, so it is very unfair to those who put forth the effort. You are late, and I do not appreciate that you would ask me to break the rules and lie for you.

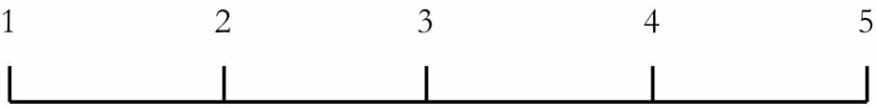
I'm so sorry Kelly, I understand how difficult it is to be running late, it happens to everyone and I feel badly that you have to deal with that. I wish I could help, but unfortunately I can't really do that.

[. . .]

Event 7 -- Extraversion

- After examining the responses below on either end of the scale, rate on a scale of one to five as to how your partner would respond to the previous situation.

Extraverted people tend to be communicative, confident, persuasive, and tend to excel at leadership and taking charge.



I don't think it's a good idea for me to clock you in. I think that is against our company policy.

Hey Kelly! Listen, I'm really sorry, but I'm not going to be able to clock you. But, getting stuck in traffic is an nderstandable reason to be late. I'm sure our supervisor will take that into account. Good luck! .

APPENDIX F: IRB OUTCOME 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 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Daniel Miller** and Co-PI: **Kimberly A. Jentsch**

Date: **February 16, 2011**

Dear Researcher:

On 2/16/2011, the IRB approved the following human participant research until 2/15/2012 inclusive: Type of

Review:	UCF Initial Review Submission Form
Project Title:	Situational Judgment Tests and Implicit Trait Policies
Investigator:	Daniel Miller
IRB Number:	SBE-11-07367
Funding Agency:	
Grant Title:	
Research ID:	N/A

The Continuing Review Application must be submitted 30days prior to the expiration date for studies that were previously expedited, and 60 days prior to the expiration date for research that was previously reviewed at a convened meeting. Do not make changes to the study (i.e., protocol, methodology, consent form, personnel, site, etc.) before obtaining IRB approval. A Modification Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at https://iris.research.ucf.edu .

If continuing review approval is not granted before the expiration date of 2/15/2012, approval of this research expires on that date. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

Use of the approved, stamped consent document(s) is required. The new form supersedes all previous versions, which are now invalid for further use. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Participants or their representatives must receive a copy of the consent form(s).

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual. On behalf

of Joseph Bielitzki, DVM, UCF IRB Chair, this letter is signed by:

Signature applied by Joanne Muratori on 02/16/2011 01:57:52 PM EST

IRB Coordinator

REFERENCES

Bauer, T., Maertz, C., Dolen, M., & Campion, M. (1998). Longitudinal assessment of applicant reactions to employment testing and test outcome feedback. *Journal of Applied Psychology*, *83*(6), 892-903. doi:10.1037/0021-9010.83.6.892.

Bauer, T., Truxillo, D., & Paronto, M. (2004). The measurement of applicant reactions to selection. *Comprehensive handbook of psychological assessment, Vol. 4: Industrial and organizational assessment* (pp. 482-506).

Becker, T. (2005). Information Exchange Article Development and Validation of a Situational Judgment Test of Employee Integrity. *International Journal of Selection and Assessment*, *13*(3), 225-232. doi:10.1111/j.1468-2389.2005.00319.x.

Bergman, M. E., Drasgow, F, Donovan, M. A. (2006). Scoring Situational Judgment Tests: Once You Get the Data, Your Troubles Begin. *International Journal of Selection and Assessment, 14* (3) 223 -- 235. doi:10.1111/j.1468-2389.2006.00345.x

Berry, C. M., Clark, M. A., & McClure, T. K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology, 96*(5), 881-906. doi:10.1037/a0023222

Borkenau, P., Brecke, S., Möttig, C., & Paelecke, M. (2009). Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, *43*, 703–706.

Borman, Walter C. (1991). *Job Behavior, Performance, and Effectiveness.* In: Handbook of industrial and organizational psychology, Vol. 2 (2nd ed.). Dunnette, Marvin D.; Hough, Leaetta M.; Palo Alto, CA, US: Consulting Psychologists Press, 1991. pp. 271-326. [Chapter]

Bruce, M., & Learner, D. (1958). A supervisory practices test. *Personnel Psychology*, 207-216. doi:10.1111/j.1744-6570.1958.tb00015.x.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin, 114*(3), 542-551. doi:10.1037/0033-2909.114.3.542

Cardall, A. (1942). *Test of Practical Judgment,* and the *Preliminary Manual,* Chicago. Science Research Associates.

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. doi:10.1037/0021-9010.82.1.143.gf

Chan, D., Schmitt, N., DeShon, R., Clause, C., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, *82*(2), 300-310. doi:10.1037/0021-9010.82.2.300.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance*, *15*(3), 233-254. doi:10.1207/S15327043HUP1503_01.

Clause, C., Mullins, M., Nee, M., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternative predictors and an example. *Personnel Psychology*, *51*(1), 193-208. doi:10.1111/j.1744-6570.1998.tb00722.x.

Clevenger, J., Pereira, G. M., Wiechman, D., Schmitt, N., & Harvey, V. S., (2001). Incremental Validity of Situational Judgment Tests. *Journal of Applied Psychology, 86* 410-417. doi:10.1111/j.1744-6570.1999.tb00176.x.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, M., & Morris, S. (2008). Testing for adverse impact when sample size is small. *Journal of Applied Psychology*, *93*(2), 463-471. doi:10.1037/0021-9010.93.2.463.

Cope, J., & Watts, G. (2000). Learning by Doing – An Exploration of Experience, Critical Incidents and Reflection in Entrepreneurial Learning. *International Journal of Entrepreneurial Behavior and Research, 6*(3), 104 – 124. doi: 10.1108/13552550010346208.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P., Terracciano, A., & McCrae, R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322-331. doi:10.1037/0022-3514.81.2.322.

Cullen, M., Sackett, P., & Lievens, F. (2006). Threats to the Operational Use of Situational Judgment Tests in the College Admission Process. *International Journal of Selection and Assessment*, *14*(2), 142-155. doi:10.1111/j.1468-2389.2006.00340.x.

Dalessio, A. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology*, *9*(1), 23-32. doi:10.1007/BF02230984.

Davis, M. (2000, May). Influence of assessor individual differences on rating errors and rating accuracy in assessment centers. *Dissertation Abstracts International*, 60

Dobbins, G. H., Farh, J. L., and Werbel, J. D. (1993). The influence of self-monitoring and inflation of grade-point averages for research and selection purposes. *Journal of Applied Social Psychology, 23,* 321-334.

Donner, A., and Eliasziw, M. (1994). Statistical Implications of the Choice between a Dichotomous or Continuous Trait in Studies of Interobserver Agreement. *Biometrics, 50*(2). 550-555. doi:10.1111/j.1550-1555.1994.01740.x.

Drexler JA Jr, Beehr TA, Stetz TA. (2001). Peer appraisals: Differentiation of individual performance on group task. *Human Resource Management, 40,* 333-345. doi:10.1002/hrm.1023

DuBois, C. L., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and White-Black differences. *Journal of Applied Psychology*, 78(2), 205-211. doi:10.1037/0021-9010.78.2.205

Eiser, J. R., & van der Pligt, J. (1984). Accentuation theory, polarization, and the judgment of attitude statements. In J. R. Eiser (Ed.), *Attitudinal judgment* (pp. 43–63). New York: Springer-Verlag. doi:10.1037/0022-3514.42.2.224

Ferguson, C. J. (2010). A meta-analysis of normal and disordered personality across the life span. *Journal of Personality and Social Psychology, 98*(4), 659-667. doi:10.1037/a0018770

Flake, W. L., and Goldman, B. A. (1991). Comparison of grade point averages and SAT scores between reporting and nonreporting men and women and freshmen and sophomores. *Perceptual and Motor Skills, 72,* 177-178. doi:10.2466/PMS.72.1.177-178

Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five U.S. racial groups. *Personnel Psychology, 61*(3), 579-616. doi:10.1111/j.1744-6570.2008.00123.x

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. doi: 10.1207/s15327043hup1101_1

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

Frucot, V. G., and Cook, G. L. (1994). Further research on the accuracy of students' self-reported grade point averages, SAT scores, and course grades. *Perceptual and Motor Skills, 79,* 743-746.

Gilliland, S. (1993). The perceived fairness of selection systems: An organizational justice perspective. *The Academy of Management Review*, *18*(4), 694-734. doi:10.2307/258595.

Gini, G. (2006). Brief report: Adaptation of the Italian version of the Tromsø Social Intelligence Scale to the adolescent population. *Journal of Adolescence*, *29*(2), 307-312. doi:10.1016/j.adolescence.2005.05.003.

Goffin, R. D., & Christiansen, N. D. (2003). Correcting *personality* tests for faking: A review of popular *personality* tests and an initial survey of researchers. *International Journal of Selection and Assessment*, *11*, 340–344. doi:10.1111/j.0965-075X.2003.00256.x

Goldman, B. A., Flake, W. L., and Matheson, M. B. (1990). Accuracy of college students' perceptions of their SAT scores and high school and college grade point averages relative to their ability. *Perceptual and Motor Skills, 70,* 514.

Graziano, W., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. *Handbook of personality psychology* (pp. 795-824). San Diego, CA US: Academic Press. doi:10.1016/B978-012134645-4/50031-7.

Harris, J., Vernon, P., & Jang, K. (1999). Intelligence and personality characteristics associated with accuracy in rating a co-twin's personality. *Personality and Individual Differences*, *26*(1), 85-97. doi:10.1016/S0191-8869(98)00133-0.

Haynes, S., Richard, D., & Kubany, E. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238-247. doi:10.1037/1040-3590.7.3.238.

Harold, C., & Ployhart, R. (2008). What do applicants want? Examining changes in attribute judgments over time. *Journal of Occupational and Organizational Psychology*, *81*(2), 191-218. doi:10.1348/096317907X235774.

Hauenstein, N., & Alexander, R. (1991). Rating ability in performance judgments: The joint influence of implicit theories and intelligence. *Organizational Behavior and Human Decision Processes*, *50*(2), 300-323. doi:10.1016/0749-5978(91)90024-N.

Hausdorf, P. A., LeBlanc, M., & Chawla, A. (2002). Cognitive ability testing and employment selection: Does test content relate to adverse impact?. *Applied H.R.M. Research*, *7*(1-2), 41-48.

Hedlund, J., Forsythe, G., Horvath, J., Williams, W., Snook, S., & Sternberg, R. (2003). Identifying and assessing tacit knowledge: Understanding the practical intelligence of

military leaders. *The Leadership Quarterly*, *14*(2), 117-140. doi:10.1016/S1048-9843(03)00006-7.

Hofstede, G., & McCrae, R. R. (2004). Personality and Culture Revisited: linking traits and dimensions of culture. *Cross-Cultural Research: The Journal of Comparative Social Science, 38*(1), 52-88.

Hogan, R., & Briggs, S., (Eds.), *Handbook of personality psychology* (pp. 795–824). Academic Press: San Diego.

Hogan, J., Hogan, R., & Busch, C. (1984). How to measure service orientation. *Journal of Applied Psychology*, *69*(1), 167-173. doi:10.1037/0021-9010.69.1.167.

Holden, R., & Jackson, D. (1979). Item subtlety and face validity in personality assessment. *Journal of Consulting and Clinical Psychology*, *47*(3), 459-468. doi:10.1037/0022-006X.47.3.459.

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

Hough, L., Eaton, N., Dunnette, M., Kamp, J., & McCloy, R. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, *75*(5), 581-595. doi:10.1037/0021-9010.75.5.581.

Hovland, C., & Sherif, M. (1952). Judgmental phenomena and scales of attitude measurement: item displacement in Thurstone scales. *The Journal of Abnormal and Social Psychology*, *47*(4), 822-832. doi:10.1037/h0056372.

Huang, J. L., & Ryan, A. (2011). Beyond personality traits: A study of personality states and situational contingencies in customer service jobs. *Personnel Psychology, 64*(2), 451-488. doi:10.1111/j.1744-6570.2011.01216.x

Hunter, D. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *International Journal of Aviation Psychology*, *13*(4), 373-386. doi:10.1207/S15327108IJAP1304_03.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72–98. doi:10.1037/0033-2909.96.1.72

James, L. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods*, *1*(2), 131-163. doi:10.1177/109442819812001.

Jensen, A. (1998). *The g factor: The science of mental ability*. Westport, CT US: Praeger Publishers/Greenwood Publishing Group

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A.Pervin & O. P.John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York: Guilford Press.

Johnson, J. A. (1981). The "self-disclosure" and "self-presentation" views of item response dynamics and personality scale validity. *Journal of Personality and Social Psychology*, 40(4), 761-769. doi:10.1037/0022-3514.40.4.761.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35*(1), 1-19. doi:10.1207/S15327906MBR3501_1.

Jones, C., and DeCotiis, T. A. (1996). Video-Assisted Selection of Hospitality Employees. *The Cornell H.R.A. Quarterly,* 68-73. doi:10.1177/001088048602700222

Judge, T. A., Jackson, C. L., Shaw, J. C., Scott, B. A., & Rich, B. L. (2007). Self-efficacy and work-related performance: The integral role of individual differences. *Journal of Applied Psychology*, *92*(1), 107-127. doi:10.1037/0021-9010.92.1.107

Kane JS, Lawler EE. (1978). Methods of peer assessment. *Psychological Bulletin, 88,* 80-81. doi:10.1037/0033-2909.85.3.555

Kanning, U., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view: Reactions to different types of situational judgment items. *European Journal of Psychological Assessment*, *22*(3), 168-176. doi:10.1027/1015-5759.22.3.168.

Kaukiainen, A., Björkqvist, K., Lagerspetz, K., Österman, K., Salmivalli, C., Rothberg, S., et al. (1999). The relationships between social intelligence, empathy, and three types of aggression. *Aggressive Behavior*, *25*(2), 81-89. doi:10.1002/(SICI)1098-2337(1999)25:2<81::AID-AB1>3.0.CO;2-M.

Klehe, U., & Anderson, N. (2007). Working hard and working smart: Motivation and ability during typical and maximum performance. *Journal of Applied Psychology*, 92(4), 978-992. doi:10.1037/0021-9010.92.4.978

Klehe, U., König, C., Richter, G., Kleinmann, M., & Melchers, K. (2008). Transparency in structured interviews: Consequences for construct and criterion-related validity. *Human Performance*, *21*(2), 107-137. doi:10.1080/08959280801917636

Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology*, *78*(6), 988-993. doi:10.1037/0021-9010.78.6.988.

Kluger, A., & Rothstein, H. (1993). The influence of selection test type on applicant reactions to employment testing. *Journal of Business and Psychology*, *8*(1), 3-25. doi:10.1007/BF02230391.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality*, *64*, 311–337.

König, C., Melchers, K., Kleinmann, M., Richter, G., & Klehe, U. (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. doi:10.1111/j.1468-2389.2007.00388.x.

Konradt, U., Hertel, G., & Joder, K. (2003). Web-based assessment of call center agents: Development and validation of a computerized instrument. *International Journal of Selection and Assessment*, *11*(2-3), 184-193. doi:10.1111/1468-2389.00242.

Labrador, J. (2007). Assessing applicant personality using low-fidelity simulations. *Dissertation Abstracts International*, 67.

Legree, P., Psotka, J., Tremble, T., & Bourne, D. (2005). Using Consensus Based Measurement to Assess Emotional Intelligence. *Emotional intelligence: An international handbook* (pp. 155-179). Ashland, OH US: Hogrefe & Huber Publishers

Levashina, J., Morgeson, F., & Campion, M. (2009). They don't do it often, but they do it well: Exploring the relationship between applicant mental abilities and faking. *International Journal of Selection and Assessment*, *17*(3), 271-281. doi:10.1111/j.1468-2389.2009.00469.x.

Lievens, F., Buyse, T., & Sackett, P. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, *58*(4), 981-1007. doi:10.1111/j.1744-6570.2005.00713.x.

Lievens, F., Chasteen, C. S., Day, E., & 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. doi:10.1037/0021-9010.91.2.247

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment*, *10*(4), 245-257. doi:10.1111/1468-2389.00215.

Lievens, F., Coetsier, P., De Fruyt, F., & De Maeseneer, J. (2002). Medical students' personality characteristics and academic performance: A five-factor model perspective. *Medical Education*, *36*(11), 1050-1056. doi:10.1046/j.1365-2923.2002.01328.x.

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology, 96*(5), 927-940. doi:10.1037/a0023496.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426-441. doi:10.1108/00483480810877598

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, *91*, 1181–1188. doi:10.1037/0021-9010.91.5.1181.

Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology*, 63(4), 507-512. doi:10.1037/0021-9010.63.4.507

Lippa, R., & Dietz, J. (2000). The relation of gender, personality, and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior*, *24*(1), 25-43. doi:10.1023/A:1006610805385.

Marcus, B., Goffin, R., Johnston, N., & Rothstein, M. (2007). Personality and cognitive ability as predictors of typical and maximum managerial performance. *Human Performance*, *20*(3), 275-285.

Markel, N., Phillis, J., Vargas, R., & Howard, K. (1972). Personality traits associated with voice types. *Journal of Psycholinguistic Research*, *1*(3), 249-255. doi:10.1007/BF01074441.

Mascaro, G. (1969). Attitude extremity and latitudes of acceptance, rejection and indifference. *Perceptual and Motor Skills*, *28*(3), 859-863.

McClough, A., & Rogelberg, S. (2003). Selection in teams: An exploration of the Teamwork Knowledge, Skills, and Ability test. *International Journal of Selection and Assessment*, *11*(1), 56-66. doi:10.1111/1468-2389.00226.

McCroskey, J., Heisel, A., & Richmond, V. (2001). Eysenck's BIG THREE and communication traits: Three correlational studies. *Communication Monographs*, *68*(4), 360-366. doi:10.1080/03637750128068.

McDaniel, M., Hartman, N., Whetzel, D., & Grubb, W. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, *60*(1), 63-91. doi:10.1111/j.1744-6570.2007.00065.x.

McDaniel, M., Morgeson, F., Finnegan, E., Campion, M., & Braverman, E. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*(4), 730-740. doi:10.1037/0021-9010.86.4.730.

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

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). Job experience correlates of job performance. *Journal of Applied Psychology*, *73*, 327–330. doi:10.1037/0021-9010.73.2.327

McDaniel, M., & Whetzel, D. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence*, *33*(5), 515-525. doi:10.1016/j.intell.2005.02.001.

McDaniel, M., & Whetzel, D. (2005). *Situational Judgment Tests.* Presented at IPMAAC Workshop June 20, 2005. doi:10.1111/j.1744-6570.2007.00065.x

McDaniel, M., Whetzel, D., Hartman, N., Nguyen, N., & Grubb, W. (2006). Situational Judgment Tests: Validity and an Integrative Model. *Situational judgment tests: Theory, measurement, and application* (pp. 183-203). doi:10.1037/0021-9010.86.4.730

McHenry, J., & Schmitt, N. (1994). Multimedia testing. *Personnel selection and classification* (pp. 193-232). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Melchers, K., Klehe, U., Richter, G., Kleinmann, M., Konig, C., & Lievens, F. (2009). 'I know what you want to know": The impact of interviewees' ability to identify criteria on interview performance and construct-related validity. *Human Performance*, *22*(4), 355-374. doi:10.1080/08959280903120295.

Meyer, J. (2010). *Reliability*. New York, NY US: Oxford University Press.

Miller, D., Smith-Jentsch, K. A., Hall, C. M., Schwartz, J., & Rivera-Cruz., C. (2010). Using Situational Judgment Tests to Measure Teamwork and Communication. In Kurtessis, J. and Krokos, K. (Co-Chairs), *Using Situational Judgment Tests to Measure Teamwork and Communication.* Symposium conducted at the annual meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of *personality* tests in personnel selection contexts. *Personnel Psychology*, *60*, 683–729. doi:10.1111/j.1744-6570.2007.00089.x

Morgeson, F., Delaney-Klinger, K., & Hemingway, M. (2005). The Importance of Job Autonomy, Cognitive Ability, and Job-Related Skill for Predicting Role Breadth and Job Performance. *Journal of Applied Psychology*, *90*(2), 399-406. doi:10.1037/0021-9010.90.2.399.

Motowidlo, S., & Beier, M. (2010). Differentiating specific job knowledge from Implicit Trait Policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology*, *95*(2), 321-333. doi:10.1037/a0017975.

Motowidlo, S., Crook, A., Kell, H., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology*, *24*(3), 281-288. doi:10.1007/s10869-009-9106-4.

Motowidlo, S., Dunnette, M., & Carter, G. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*(6), 640-647. doi:10.1037/0021-9010.75.6.640.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006a). A theoretical basis for situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests*. Mahwah, NJ: Erlbaum.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). Implicit Policies About Relations Between Personality Traits and Behavioral Effectiveness in Situational Judgment Items. *Journal of Applied Psychology*, 91, 4, 749 – 761. doi:10.1037/0021-9010.91.4.749

Motowidlo, S., & 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.

Mount, M. K., Sytsma, M. R., Hazucha, J., & Holt, K. E. (1997). Rater–ratee race effects in developmental performance ratings of managers. *Personnel Psychology*, 50(1), 51-69. doi:10.1111/j.1744-6570.1997.tb00900.x

Mullins, M.E., & Schmitt, N. (1998). Situational judgment testing: Will the real constructs please present themselves? Paper presented at the 13th Annual Conference of the Society of Industrial Organizational Psychology, Dallas, TX.

Nelson, C. (2009). Job type as a moderator of the relationship between situational judgment and personality. *Dissertation Abstracts International*, 69.

Nguyen, N., Biderman, M., & McDaniel, M. (2005). Effects of Response Instructions on Faking a Situational Judgment Test. *International Journal of Selection and Assessment*, *13*(4), 250-260. doi:10.1111/j.1468-2389.2005.00322.x.

O'Connell, M., Doverspike, D., Norris-Watts, C., & Hattrup, K. (2001). Predictors of organizational citizenship behavior among Mexican retail salespeople. *International Journal of Organizational Analysis*, *9*(3), 272-280. doi:10.1108/eb028936.

O'Connell, M., Hartman, N., McDaniel, M., Grubb, W., & 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. doi:10.1111/j.1468-2389.2007.00364.x.

Olney, R. J. (1982). How employers view resumes: 1974… 1981. *Journal of College Placement*, *42*, 64–67.

Olson-Buchanan, J., & Drasgow, F. (2006). Multimedia Situational Judgment Tests: The Medium Creates the Message. *Situational judgment tests: Theory, measurement, and application* (pp. 253-278

Olson-Buchanan, J., Drasgow, F., Moberg, P., Mead, A., Keenan, P., & Donovan, M. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology*, *51*(1), 1-24. doi:10.1111/j.1744-6570.1998.tb00714.x.

Ones, D., Dilchert, S., Viswesvaran, C., & Judge, T. (2007). In support of personality assessment in organizational settings. *Personnel Psychology*, *60*(4), 995-1027. doi:10.1111/j.1744-6570.2007.00099.x.

Ones, D., & Viswesvaran, C. (2007). Labor market influences on personality scale scores among job applicants: Four field studies in personnel selection settings. *Zeitschrift für Personalpsychologie*, *6*(2), 71-84. doi:10.1026/1617-6391.6.2.71.

Ones, D., Viswesvaran, C., & Dilchert, S. (2005). Cognitive Ability in Selection Decisions. *Handbook of understanding and measuring intelligence* (pp. 431-468). Thousand Oaks, CA US: Sage Publications, Inc.

Ones, D., Viswesvaran, C., & Schmidt, F. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*(4), 679-703. doi:10.1037/0021-9010.78.4.679.

Oswald, F., Friede, A., Schmitt, N., Kim, B., & Ramsay, L. (2005). Extending a Practical Method for Developing Alternate Test Forms Using Independent Sets of Items. *Organizational Research Methods*, *8*(2), 149-164. doi:10.1177/1094428105275365.

Oswald, F., Schmitt, N., Kim, B., Ramsay, L., & Gillespie, M. (2004). Developing a Biodata Measure and Situational Judgment Inventory as Predictors of College Student Performance. *Journal of Applied Psychology*, *89*(2), 187-207. doi:10.1037/0021-9010.89.2.187.

Planells-Bloom, D. (1992). Latino cultures: Framework for understanding the Latina adolescent and assertive behavior. In I. G. Fodor, I. G. Fodor (Eds.), *Adolescent assertiveness and social skills training: A clinical handbook* (pp. 113-128). New York, NY US: Springer Publishing Co.

Pauls, C., & Crost, N. (2005). Cognitive Ability and Self-Reported Efficacy of Self-Presentation Predict Faking on Personality Measures. *Journal of Individual Differences*, *26*(4), 194-206. doi:10.1027/1614-0001.26.4.194.

Payton, M. E., Greenstone, M. H., & Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science*, 3, 34–39.

Peeters, H., & Lievens, F. (2005). Situational Judgment Tests and Their Predictiveness of College Students' Success: The Influence of Faking. *Educational and Psychological Measurement*, *65*(1), 70-89. doi:10.1177/0013164404268672.

Penley, J., & Tomaka, J. (2002). Associations among the Big Five, emotional responses and coping with acute stress. *Personality and Individual Differences*, *32*(7), 1215-1128. doi:10.1016/S0191-8869(01)00087-3.

Pennington, D. (2003). *Essential personality*. London England: Arnold.

Ployhart, R., & Ehrhart, M. (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. doi:10.1111/1468-2389.00222.

Ployhart, R., Lim, B., & Chan, K. (2001). Exploring relations between typical and maximum performance ratings and the five factor model of personality. *Personnel Psychology*, *54*(4), 809-843. doi:10.1111/j.1744-6570.2001.tb00233.x.

Ployhart, R., & Ryan, A. (1998). Toward and explanation of applicant reactions: An examination of original justice and attribution frameworks. *Organizational Behavior and Human Decision Processes*, *74*(1), doi:10.1006/obhd.1998.2777.

Potosky, D., & Bobko, P. (1998). The Computer Understanding and Experience Scale: A self-report measure of computer experience. *Computers in Human Behavior*, *14*(2), 337-348. doi:10.1016/S0747-5632(98)00011-9.

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

Pulakos, E., Schmitt, N., Dorsey, D., Arad, S., Hedge, J., & Borman, W. (2002). Predicting adaptive performance: Further tests of a model of adaptability. *Human Performance*, *15*(4), 299-324. doi:10.1207/S15327043HUP1504_01.

Putka, D. J., & McCloy, R. A. (2004). Preliminary AIM validation based on GED Plus program data. In D. J.Knapp, E. D.Heggestad, & M. C.Young (Eds.), *Understanding and improving the assessment of individual motivation (AIM) in the Army's GED Plus program* (pp. 3–19–3–30). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Quiñones, M., Ford, J., & Teachout, M. (1995). The relationship between work experience and job performance: A conceptual and meta-analytic review. *Personnel Psychology*, *48*(4), 887-910. doi:10.1111/j.1744-6570.1995.tb01785.x.

Richman-Hirsch, W., Olson-Buchanan, J., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology*, *85*(6), 880-887. doi:10.1037/0021-9010.85.6.880.

Rupp, D. E., & Spencer, S. (2006). When customers lash out: The effects of customer interactional injustice on emotional labor and the mediating role of discrete emotions. *Journal of Applied Psychology*, *91*, 971–978. doi:10.1037/0021-9010.91.4.971

Sackett, P. R. (2007). Revisiting the origins of the typical-maximum performance distinction. *Human Performance, 20*(3), 179-185.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73,* 482–486. doi:10.1037/0021-9010.73.3.482

Salgado, J., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. (2003). A Meta-Analytic Study of General Mental Ability Validity for Different Occupations in the European Community. *Journal of Applied Psychology*, *88*(6), 1068-1081. doi:10.1037/0021-9010.88.6.1068.

Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach*. Thousand Oaks, CA: Sage. doi:10.2174/138161206777698909

Schmitt, N., & Chan, D. (2006). Situational Judgment Tests: Method or Construct?. *Situational judgment tests: Theory, measurement, and application* (pp. 135-155). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers. doi:10.1002/bmc.568

Schubert, S., Ortwein, H., Dumitsch, A., Schwantes, U., Wilhelm, O., & Kiessling, C. (2008). A situational judgement test of professional behaviour: Development and validation. *Medical Teacher*, *30*(5), 528-533. doi:10.1080/01421590801952994.

Schyns, B., Felfe, J. (2006). *The Personality of Followers and its Effect on the Perception of Leadership: An Overview, a Study, and a Research Agenda.* Small Group Research, 37, 5, 522-539. doi:10.1177/1046496406293013

Seberhagen, L. (1997). *Court Rules Against Another Job Element Test. From PTC/MW Newsletter, 8/97)*

Shepperd, J. A. (1993). Student derogation of the Scholastic Aptitude Test: Biases in perceptions and presentations of College Board scores. *Basic and Applied Social Psychology, 14,* 455-473. doi:10.1207/s15324834basp1404_5

Smiderle, D., Perry, B., & Cronshaw, S. (1994). Evaluation of video-based assessment in transit operator selection. *Journal of Business and Psychology*, *9*(1), 3-22. doi:10.1007/BF02230983.

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

Smith-Jentsch, K. A. (2007). The impact of making targeted dimensions transparent on relations with typical performance predictors. *Human Performance*, *20*, 187–203. doi:10.1037/h0e083472

Smither, J., & Reilly, R. (1987). True intercorrelation among job components, time delay in rating, and rater intelligence as determinants of accuracy in performance ratings. *Organizational Behavior and Human Decision Processes*, *40*(3), 369-391. doi:10.1016/0749-5978(87)90022-7.

Smither, J., Reilly, R., Millsap, R., & Pearlman, K. (1993). Applicant reactions to selection procedures. *Personnel Psychology*, *46*(1), 49-76

Stemler, S., & Sternberg, R. (2006). Using Situational Judgment Tests to Measure Practical Intelligence. *Situational judgment tests: Theory, measurement, and application* (pp. 107-131). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Sternberg, R.J., & the Rainbow Project Team. (2002, February 16). *The rainbow project: Augmenting the validity of the SAT*. Paper presented to American Academy of Arts and Sciences, Boston, MA.

Sternberg, R., Forsythe, G., Hedlund, J., Horvath, J., Wagner, R., Williams, W., Snook, S., Grigorenko, E. (2000). *Practical intelligence in everyday life*. New York, NY US: Cambridge University Press doi:10.1016/S0160-2896(01)00081-2

Stevens, M., & Campion, M. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management*, *25*(2), 207-228. doi:10.1016/S0149-2063(99)80010-5.

Stöber, J., Dette, D., & Musch, J. (2002). Comparing continuous and dichotomous scoring of the Balanced Inventory of Desirable Responding. *Journal of Personality Assessment*, *78*(2), 370-389. doi:10.1207/S15327752JPA7802.

Tajfel, H. (1957). Value and the perceptual judgment of magnitude. *Psychological Review*, *64*(3), 192-204. doi:10.1037/h0047878.

Tesluk, P., & Jacobs, R. (1998). Toward an integrated model of work experience. *Personnel Psychology*, *51*(2), 321-355. doi:10.1111/j.1744-6570.1998.tb00728.x.

Tett, R., & Guterman, H. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, *34*(4), 397-423. doi:10.1006/jrpe.2000.2292.

Thorndike, E. L. (1920). "Intelligence and its use". *Harper's Magazine* 140: 227–235.

Thorndike, E. L. (1949). *Personnel selection*. New York: Wiley.

Thornton, G. & Byham, W. (1982). *Assessment Centers and Managerial Performance.* New York: Academic Press.

Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M. & Barry, W. J. (2006). Modelling personality features by changing prosody in synthetic speech. *Proc. Speech Prosody 2006,* Dresden, Germany.

Van der Pligt, J., & Eiser, J. (1984). Attribution of traits to self and others: Situationality vs. uncertainty. *Current Psychological Research & Reviews*, *3*(1), 45-51. doi:10.1007/BF02686531.

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98*(2), 281-300. doi:10.1037/a0017908

Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology*, *95*, 1202–1216.

Vescio, T. K., Gervais, S. J., Snyder, M., & Hoover, A. (2005). Power and the Creation of Patronizing Environments: The Stereotype-Based Behaviors of the Powerful and Their Effects on Female Performance in Masculine Domains. *Journal of Personality and Social Psychology*, 88(4), 658-672. doi:10.1037/0022-3514.88.4.658

Veselka, L., Schermer, J., Petrides, K., & Vernon, P. (2009). Evidence for a heritable general factor of personality in two studies. *Twin Research and Human Genetics*, *12*(3), 254-260. doi:10.1375/twin.12.3.254.

Weekley, J., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology*, *50*(1), 25-49. doi:10.1111/j.1744-6570.1997.tb00899.x.

Weekley, J., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology*, *52*(3), 679-700. doi:10.1111/j.1744-6570.1999.tb00176.x.

Weekley, J., & Ployhart, R. (2005). Situational Judgment: Antecedents and Relationships with Performance. *Human Performance*, *18*(1), 81-104. doi:10.1207/s15327043hup1801_4.

Weekley, J., Ployhart, R., & Harold, C. (2004). Personality and Situational Judgment Tests Across Applicant and Incumbent Settings: An Examination of Validity, Measurement, and Subgroup Differences. *Human Performance*, *17*(4), 433-461. doi:10.1207/s15327043hup1704_5.

Weekley, J., Ployhart, R., & Holtz, B. (2006). On the Development of Situational Judgment Tests: Issues in Item Development, Scaling, and Scoring. *Situational judgment tests: Theory, measurement, and application* (pp. 157-182). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Wernimont, P., & Campbell, J. (1968). Signs, Samples, and Criteria. *Journal of Applied Psychology*, *52*(5), 372-376. doi:10.1037/h0026244.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, 21(3), 291-309. doi:10.1080/08959280802137820

Wilson, K. (2010). An analysis of bias in supervisor narrative comments in performance appraisal. *Human Relations*, 63(12), 1903-1933. doi:10.1177/0018726710369396

Wingrove, J., Glendinning, R., & Herriot, P. (1984). Graduate pre-selection: A research note. *Journal of Occupational Psychology*, *57*, 169–171.

Witt, L. A., & Spitzmüller, C. (2007). Person-situation predictors of maximum and typical performance. *Human Performance*, 20(3), 305-315.

Workforce Central Florida. (2006). *State of the Workforce Report.* Retrieved July 15, 2007, from http://www.workforcecentralflordia.com/assets/State_of_the_Workforce_Report06.pdf

Zagorseck, H., Stough, S. J., Jaklic, M. (2006). *Analysis of the Reliability of the Leadership Practices Inventory in the Item Response Theory Framework.* International Journal of Selection and Assessment, 14, 2, 180-191. doi:10.1111/j.1468-2389.2006.00343.x