2006

The Construct Validity Of A Situational Judgment Test In A Maximum Performance Context

Kevin Stagl
University of Central Florida

THE CONSTRUCT VALIDITY OF A SITUATIONAL JUDGMENT TEST IN A MAXIMUM PERFORMANCE CONTEXT

by

KEVIN C. STAGL
B.S. University of North Florida, 1997
M.S. University of Central Florida, 2004

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

Summer Term
2006

Major Professor: Eduardo Salas

ABSTRACT

A Predictor Response Process model (see Ployhart, 2006) and research findings were leveraged to formulate research questions about, and generate construct validity evidence for, a new situational judgment test (SJT) designed to measure declarative and strategic knowledge. The first question asked if SJT response instructions (i.e., 'Should Do', 'Would Do') moderated the validity of an SJT in a maximum performance context. The second question asked what the upper-bound criterion-related validity coefficient is for SJTs in talent selection contexts in which typical performance is the criterion of interest. The third question asked whether the SJT used in the present study was fair for gender and ethnic-based subgroups according to Cleary's (1968) definition of test fairness. Participants were randomly assigned to complete an SJT with either 'Should Do' or 'Would Do' response instructions and their maximum decision making performance outcomes were captured during a moderate fidelity poker simulation. The findings of this study suggested knowledge, as measured by the SJT, interacted with response instructions when predicting aggregate and average performance outcomes such that the 'Should Do' SJT had stronger criterion-related validity coefficients than the 'Would Do' version. The findings also suggested the uncorrected upper-bound criterion-related validity coefficient for SJTs in selection contexts is at least moderate to strong (β = .478). Moreover, the SJT was fair according Cleary's definition of test fairness. The implications of these findings are discussed.

In dedication to my family and friends whose guidance, patience, and support made this possible.

I would especially like to thank co-chairs Eduardo Salas and Barbara A. Fritzsche and members William Wooten and C. Shawn Burke for serving on my dissertation committee.

ACKNOWLEDGMENTS

I would like to thank Eduardo Salas, Barbara A. Fritzsche, William Wooten, C. Shawn Burke, Patrick J. Rosopa, and Kelly A. Rutkowski for their insightful comments on this work. I would also like to thank Gabriella Severe, Brandy Burke, Kaoruko Nakano, and Linda Burks for their help with data collection. Finally, I would like to thank Texas Hold'em poker subject matter experts Joseph W. Guthrie, Cameron Klein, and James S. Painter III for their suggestions.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION

Situational judgment tests (SJTs) have been used since 1873 (e.g., Ansbacher, 1941; Binet, 1905; File, 1945; Kite, 1916; Moss, 1926; Simoneit, 1938) but modern variations have recently remerged as popular assessment tools (e.g., Weekley & Ployhart, 2006 a). SJTs are comprised of item stems, response instructions, standardization rules, and when a forced choice format is used, a set of response alternatives. SJTs are often used as assessment tools in human capital management initiatives such as those involving talent selection (see Alignmark, 2001; Chan & Schmitt, 2002; Clevenger, Jockin, Morris, & Anselmi, 1999; Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; Harvey, Morath, Christopher, & Anderson, 2003; Joiner, 2002; McDaniel, Yost, Ludwick, Hense, & Hartman, 2004; Motowidlo, Dunnette, & Carter, 1990; Motowidlo, Hanson, & Crafts, 1997; Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Peters & Lievens, 2005; Ployhart, Weekley, Holtz, & Kemp, 2003; Pulakos & Schmitt, 1996; Schmitt & Mills, 2001; Weekley & Jones, 1997; Weekley & Jones, 1999), talent development (see Hanson, Horgen, & Borman, 1998; Hedge, Borman, & Hanson, 1996; Hunter, 2003; Hunter, Martinussen, & Wiggins, 2003; Mullins, 2000; Weekly & Ployhart, 2002; www.faa.gov; www.aimmconsult.com), and talent retention (see www.talentkeepers.com).

The burgeoning popularity of SJTs is not surprising given their numerous practical benefits. A major advantage of SJTs is that they can explain incremental variance in a wide range of occupational and educational criteria over cognitive ability tests and personality measures (Clevenger et al., 2001; Hedlund, Plamondon, Wilt, Nebel, Ashford, & Sternberg, 2001; Lievens, Buyse, & Sackett, 2005; Schmitt & Mills, 2001; Weekley & Ployhart, 2005). SJTs also offer a versatile means of conducting training needs assessment, delivering scenario-based training content, and/or assessing learning during training evaluation (Fritzsche, Stagl,

Salas, & Burke, 2006). SJTs are also more efficient relative to some traditional diagnostic tools, as respondents can complete up to 20 low-fidelity SJT scenarios in the same amount of time it takes to finish one high-fidelity exercise (Joiner, 2002). Moreover, respondents report favorable reactions to the use of SJTs (Lievens et al., 2005; Shotland, Alliger, & Sales, 1998), likely in part because they are perceived to be face valid (Anderson, 2003; Chan & Schmitt, 1997; Hanson & Borman, 1987; Hausknecht, Day, & Thomas, 2004; Rosen, 1961; Truxillo & Hunthausen, 1999). Given these advantages, the use of SJTs will only continue to increase in the foreseeable future.

Although SJTs are useful for a variety of purposes and for a variety of reasons, they are often used as measurement methods with less regard for the constructs they measure (Schmitt & Chan, 2006). This practice may persist because most SJT users are predominantly concerned with the prediction of performance and have a more limited interest in the latent constructs measured by SJTs. The secondary status afforded to the constructs measured by SJTs can be seen in a recent meta-analytic initiative which was undertaken to examine SJTs as a measurement method (e.g., McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). This is problematic because the validity evidence for an assessment tool is less meaningful without a concomitant understanding of the constructs it measures (Guion, 1998; Nunnally & Bernstein, 1994; Schmitt & Chan, 1998; Schmitt & Chan, 2006). In fact, several sources of evidence should be collected to support the inferences that are drawn about the constructs measured by a measurement method because when these constructs are ambiguous it is difficult to improve its psychometric properties, meaningfully interpret observed predictor-criterion relationships, generate supporting validity evidence for its applications, and professionally and legally defend its use as an assessment tool (Ployhart & Ryan, 2000).

Designing SJTs to measure specific dimensions identified via either job analysis or criterion-theory is the first step in creating assessment tools that have meaningful construct validity evidence, are representative of the performance domain, and are subsequently job-related (Weekley, Ployhart, & Holtz, 2006). Yet, most SJTs are developed via an iterative domain sampling approach grounded in critical incidents (Ployhart & Ehrhart, 2003). The use of a domain sampling approach to develop an SJT can ultimately cloud explanations about the constructs it measures because the performance domains critical incidents are sampled from are often multidimensional (Schmitt & Chan, 2006). This practice can result in an inadequate preoperational explication of constructs, a threat to the inferences that can be drawn about the construct validity of measures and manipulations (Cook & Campbell, 1979).

One line of research that can yield construct validity evidence for SJTs examines the item characteristics of these assessment tools. Researchers have examined various aspects of SJT item characteristics such as the origin of stem development (Weekley et al., 2006), stem fidelity (Motowidlo et al., 1997), stem comprehensibility (Sacco, Scheu, Ryan, Schmitt, Schmidt, & Rogg, 2000), and stem complexity (Reynolds, Sydell, Scott, & Winter, 2000). The findings from these studies can be leveraged to better develop, structure, and score SJTs and thereby increase the construct validity evidence for, and utility of, SJTs (Weekley et al., 2006). The importance of this kind of research is implicit in the tenets of the *Standards for Educational and Psychological Testing* (i.e., *Standards*), which states an unambiguous rationale must exist for the item characteristics of assessment tools (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Research conducted to examine the response instructions of SJTs can also provide construct validity evidence for their applications. Substantial variation in the response

instructions of SJTs used in lab and field studies (see Clevenger, 1999) has prompted repeated calls for research on the issue (Fritzsche et al., 2006; Horgen, 2004; McDaniel, Hartman, & Grubb, 2003; McDaniel, Hartman, Nguyen, & Grubb, 2006; McDaniel & Nguyen, 2001; Nguyen, McDaniel, & Biderman, 2002; Ployhart, 2006; Ployhart & Ehrhart, 2003; Weekley et al., 2006). The importance of this kind of research is underscored by findings that suggested response instructions affected the construct and criterion-related validity evidence for biodata-based assessment tools and interviews (Campion, Palmer, & Campion, 1997).

In response to the above calls for research, two studies have recently been conducted to investigate two general types of SJT response instructions (i.e., 'Should Do', 'Would Do') (McDaniel et al., 2003; Ployhart & Ehrhart, 2003). The findings of both McDaniel and colleagues' meta-analysis and Ployhart and Ehrhart's study suggested response instructions may moderate the criterion-related validity of SJTs. The results of these studies were mixed, however, with more favorable criterion-related validity evidence for SJTs with 'Should Do' instructions in the former meta-analytic initiative and 'Would Do' SJTs in the latter study. These results are intriguing, but their implications are difficult to discern because the SJTs included in these studies were predominantly developed via the use of critical incidents rather than a construct oriented approach and thus, the constructs measured in these studies are somewhat ambiguous.

To help guide research on SJTs and their response instructions, Ployhart (2006) recently adapted a general Predictor Response Process (PRPR) model (see Appendix A) to illustrate some of the factors impinging upon the response processes of SJT respondents (see Appendix B). According to the PRPR model, response instructions contribute a contaminating source of variance to SJT scores. The PRPR Model suggests that when the response instructions of an SJT change, the primary determinants of the response processes respondents engage in when

formulating a response to an item also change (see Appendix C). Because the determinants of the cognitive processes test takers engage in when formulating responses to SJT items change when response instructions change, instructions can affect the psychometric properties of, and construct validity evidence for, SJTs (Ployhart, 2006). Thus, the PRPR model provides an explanation for the findings of the above studies which suggested response instructions affected the construct validity, and may have moderated the criterion-related validity, of the SJTs examined. The present study leveraged the PRPR model to guide the generation of confirmatory and exploratory hypotheses about the validity and fairness of an SJT.

Construct validity evidence for SJTs can also be generated by examining their relationships with different types of criteria. For example, the studies described by McDaniel et al. (2003) and Ployhart and Ehrhart (2003) included self-, supervisor-, and/or peer-reports of typical performance as the primary operationalizations of the criterion constructs examined. The findings from this line of research are informative about the SJTs examined in these studies; however, questions remain about the prediction of peak performance criteria via an SJT that measures specific constructs. Commenting on the issue, Ployhart and Ehrhart (2003) stated "'Should Do' measures might best predict maximum performance or performance during high transitions, whereas 'Would Do' measures might be most predictive of stable performance" (p. 13). A similar assertion was made about the prediction of peak performance criteria in a training context (Fritzsche et al., 2006). The present study examined the nomological network of an SJT when its response instructions were manipulated and participant aggregate and average maximum decision making performance outcomes were objectively measured.

A related issue worthy of consideration is the maximum economic value or utility of SJTs. The utility of an assessment tool is directly proportional to its predictive validity

coefficient (Schmidt, Hunter, McKenzie, & Muldrow, 1979). In regards to this issue, prior meta-analytic research findings suggested the estimated mean population correlation coefficient for SJTs is .34 (McDaniel et al., 2001). More recent meta-analytic findings suggested SJTs with instructions to pick the best option had an estimated mean population correlation coefficient of .40 (McDaniel et al., 2003). Collectively, these findings suggest SJTs can have substantial utility as assessment tools; which is important because their use should provide a return on investment.

Based on the above findings it seems SJTs can provide some amount of dollar benefit to organizations. The amount of this benefit may be underestimated, however, if a utility estimate is based in part on a predictive validity coefficient from a study that included subjective ratings of typical performance as a criterion. The use of subjective ratings of typical performance to examine the predictive validity, and thereby the utility of SJTs, is problematic, as typical performance is much more variable and multiply determined than maximum performance. This is because typical performance is driven by a host of factors such as knowledge, skill, personality, and motivation. By contrast, maximum performance is characterized by uniformly high levels of motivation and the identification and execution of optimal responses (Campbell, 1990). Thus, when subjective ratings of typical performance are used as criteria, these extraneous factors can serve to attenuate criterion-related validity, and thereby estimates of utility, of an SJT. In fact, recent meta-analytic findings suggested the use of performance ratings as a criterion resulted in lower validity coefficients for a wide range of commonly used assessment tools than did the use of objective indices (Schmitt, Gooding, Noe, & Kirsch, 1984).

If prior research has underestimated the strength of the relationships between the constructs measured by SJTs and performance criteria, then utility estimates based on these statistics are also downwardly biased. In order to estimate the maximum possible utility of SJTs,

it is first necessary to estimate the upper-bound validity coefficient associated with the use of SJTs. Specifically, an estimate is needed of the population parameter of SJTs when predicting peak performance. A closer approximation of the population parameter is the regression coefficient between an SJT with 'Should Do' instructions and a maximum performance outcome. Moreover, this estimate should be based upon predictor and criterion constructs that have been purposively sampled from theories of job performance. The present study met this need.

In sum, research was needed to generate construct validity evidence for an SJT in a maximum performance context in order to extend prior mixed research results and provide guidance for the development and use of SJTs. The present study met this need by applying the PRPR model to generate hypotheses about an SJT using a between-subjects experimental design. The PRPR model was leveraged to examine three questions about an SJT when its response instructions were manipulated and participant maximum decision making performance outcomes were measured. The first question asked if SJT response instructions (i.e., 'Should Do', 'Would Do') moderated the construct and criterion-related validity of a newly developed SJT that was designed to measure declarative and strategic knowledge. The second question asked what the upper-bound criterion-related validity coefficient is for SJTs in talent selection contexts in which typical performance is the criterion of interest. The third question asked whether the SJT used in this research was fair for gender and ethnic-based subgroups according to the standards set forth in Cleary's (1968) definition of test fairness. The implications of the findings from this study for the design and use of SJTs are discussed.

CHAPTER TWO: CONCEPTUAL FOUNDATION

Electronic searches of computerized databases were conducted using the key words *situational judgment tests* and *situational judgment testing*. The electronic databases EBSCOhost, Academic Search Premier, Business Source Premier, Military and Government Collection, PsycARTICLES, and PsycINFO were searched for pertinent research published between 1900 and 2005. Collectively, these databases contain over 15,000 sources of scientific literature. In addition to the electronic searches, relevant literature was also identified by searching the author's archives on SJTs. Pertinent findings are synthesized in this section.

Situational Judgment Tests

As noted in the introduction, SJTs are assessment tools comprised of item stems, response instructions, standardization rules, and when a forced choice format is used, a set of response alternatives. These assessment tools can be administered via paper-and-pencil, video, or orally. As noted by Fritzsche et al. (2006, p. 1), the use of SJTs as assessment tools is predicated upon the assumption that those respondents:

> …who can identify more effective and less effective responses to job-related situations have greater job-relevant knowledge or better job-related judgment and reasoning skills and are thus, expected to have higher levels of job performance than those who are less able to identify appropriate responses to job-related situations.

The item stem of an SJT is a realistic job-related scenario presented in narrative form. When a constructed response format is used, SJT respondents provide narrative responses to item stems. In contrast, when a forced-choice format is used, several response alternatives are included to present plausible courses of action which can be chosen by respondents in response to a scenario posed in an item stem. The response instructions of an SJT often ask respondents to

choose a response alternative that best matches what they should do or would do in response to a scenario. SJT response instructions are discussed in greater detail later in this text.

Although considerable debate continues about the mechanisms via which respondent SJT scores relate to subsequent performance, at least three theoretical rationales (i.e., behavioral consistency principle, theory of planned behavior, performance-related constructs) have been advanced in the literature which account for this relationship (Ployhart & Ehrhart, 2003). These three mechanisms offer theoretically anchored explanations for why SJTs are related to typical and maximum performance criteria.

Behavioral Consistency

One reason why SJT scores are predictive of subsequent performance is offered by the principle of behavioral consistency (Motowidlo et al., 1990). The behavioral consistency principle states the best predictor of future behavior is past behavior (Wernimont & Campbell, 1968). According to the behavioral consistency principle, assessment tools which present job-related scenarios to respondents produce responses which are best characterized as samples rather than signs of the respondent's future performance (Motowidlo et al., 1990). The courses of action chosen by SJT respondents are a sample of the actual behavior they will take in the performance context because SJT scenarios are sampled directly from the performance domain. In light of the behavioral consistency principle, SJTs are work simulations which should be designed to match the nature of behavior that is expected in the future performance context.

Intentions

A second reason advanced in the literature for why SJTs predict performance is because they measure intentions (Motowidlo et al., 1990). This rationale is especially plausible if

respondents are unfamiliar with the specific situations comprising an SJT, as responses to novel scenarios may reflect intentions to perform the chosen response option (Ployhart & Ehrhart, 2003). When novel situations are encountered, respondents draw from their prior experiences to determine how they intend to respond to the scenarios comprising an SJT.

If intentions are indeed the reason why SJTs predict performance, then the theory of planned behavior suggests SJTs should be constructed to match pertinent criteria in terms of target, action, context, and time (see Ajzen, 1991; Ajzen & Fishbein, 1977; Fishbein & Ajzen, 1975). More specifically, the theory of planned behavior suggests respondent SJT scores will be more strongly related to future behavior if the predictor is constructed to reflect the situational contingencies which characterize the performance context.

Performance-related Constructs

The third reason advanced in the literature for why SJTs are predictive of performance is because they assess performance-related constructs. For example, SJTs can be designed to assess constructs such as declarative, procedural, and strategic knowledge (Motowidlo et al., 1997; Schmidt & Hunter, 1993), tacit knowledge (Sternberg, Forsythe, Hedlund, Horvath, Wagner, Williams, Snook, & Grigorenko, 2000), and cognitive ability (McDaniel et al., 2001; Weekley & Jones, 1997). Each of these constructs has been advanced in the literature as either a direct or indirect determinant of performance (see Campbell, 1990).

The three previously noted rationales offer theoretical explanations for why SJTs are predictive of performance (Motowidlo, Borman, & Schmitt, 1997). Unfortunately, however, the preponderance of research conducted to date has sought to determine if SJTs predict various criteria rather than why. Therefore, the relative validity of these three explanations remains uncertain (Ployhart & Ehrhart, 2003). Research is needed to compare the validity of these three

explanations by manipulating item content, instruction set, and targeted construct, using a fully-crossed experimental design (Ployhart & Ehrhart, 2003).

Although comparative validity evidence is lacking in the current body of SJT literature, it should be noted evidence for each of these three explanations exists in the wider domains of testing and individual assessment. For example, the behavioral consistency principle underlies assessment centers conducted for selection, development, and certification (Wernimont & Campbell, 1968). Moreover, intentions are argued to be the mechanism via which situational interviews predict performance (Motowidlo, 1999). A substantial base of evidence also suggests constructs such as declarative knowledge, procedural knowledge, and motivation are critical for effective performance (Campbell, 1990). Thus, each of these three explanations for why SJTs predict performance has indirect supporting evidence.

Research Investigating SJT Response Instructions

Research examining SJT response instructions has recently been conducted. The findings from this line of research suggest different item instructions lead to different test taker responses and thereby affect the psychometric properties of, and validity evidence for, SJTs (Ployhart & Ehrhart, 2003). For example, the findings of Ployhart and Ehrhart's (2003) study suggested response instructions affected the psychometric properties of the SJTs examined. Specifically, SJTs with 'Would Do' response instructions had more favorable psychometric properties than SJTs with 'Should Do' response instructions. The use of 'Should Do' instructions resulted in higher means and lower standard deviations than 'Would Do' response instructions. Furthermore, all three versions of the 'Should Do' SJT investigated by Ployhart and Ehrhart had significant skewness and kurtosis, whereas only one of the three 'Would Do' SJT versions had a score distribution that was skewed.

Ployhart and Ehrhart's (2003) findings also suggested SJTs with 'Would Do' instructions had higher criterion-related validities than SJTs with 'Should Do' instructions. Specifically, student SJT scores derived from the use of 'Would Do' instructions were more strongly correlated with self- and peer-reports of student performance than SJT scores derived from the use of 'Should Do' response instructions. In fact, scores from the 'Would Most/Least Likely Do' SJT version were significantly correlated with self-reports of student performance, whereas scores from the 'Should Do' SJT were not. Similarly, student test scores derived from the use of 'Would Most/Least Likely Do' SJT instructions were significantly correlated with peer-reports of student performance, whereas test scores from the 'Should Do' SJT version were not.

Meta-analytic research also investigated SJT response instructions (McDaniel et al., 2003). After coding primary studies as to which type of response instructions were used, McDaniel et al.'s meta-analytic findings suggested SJT response instructions may moderate the relationship between the constructs measured by SJTs and job performance. Specifically, both the average effect sizes (i.e., the estimated population correlation coefficients) for the 'Would Do' and 'Should Do' SJT versions were statistically significant and noticeably different.

If response instructions moderate the relationship between the constructs measured by SJTs and performance criteria, it is important to understand why this occurs. The next subsection offers a theory-based explanation by examining both the latent cognitive processes respondents engage in when formulating a response to an SJT item and the individual, methodological, and contextual factors which impinge upon these response processes.

Response Instruction Effects

One issue implicitly raised by Ployhart and Ehrhart's (2003) and McDaniel and colleagues' (2003) research is why response instructions affected, and possibly moderated, the

relationship between SJTs and typical performance. In order to understand the effects of response instructions on the psychometric properties of, and validity evidence for, an assessment tool it is important to first consider the latent cognitive response processes, and other individual, methodological, and contextual factors which impact a response to an SJT item. This subsection addresses the latent cognitive processes respondents engage in, and some of the myriad of factors which impinge on these processes, when a test item is answered.

The PRPR model (Krosnick, 1999; Tourangeau, Rips, & Rasinski, 2000; see Appendix A) provides a theoretical basis for examining why item characteristics such as response instructions affect test takers' responses and thereby the validity evidence in support of assessment tools. Thus, the PRPR model provides a framework for: (1) illuminating the multitude of factors contributing to a participant's response to an SJT item, (2) explaining why SJT response instructions can moderate the relationships between the constructs measured by SJTs and performance criteria, and (3) identifying nuisance factors for control in research. In fact, according to both the *Standards* (2003) and the *Principles for the Validation and Use of Personnel Selection Procedures* (2004), examining response processes provides substantive validity evidence for assessment tools (see American Educational Research Association et al., 1999; Messick, 1995; Society for Industrial and Organizational Psychology, 2004).

The PRPR model suggests test takers engage in four sequential psychological processes (i.e., comprehension, retrieval, judgment, response selection) when answering an item of a test or measure (Tourangeau et al., 2000; see Appendix A). These four cognitive processes are latent, whereas the actual response of a test taker to an assessment item is manifest. As can be seen in the PRPR model presented in Appendix A, latent response processes or constructs are illustrated with circles, the manifest variable or response is depicted with a box, one-headed arrows

represent theoretical causal relationships, and two-headed arrows depict covariance.

Ployhart (2006) recently adapted this general model to depict some of the factors impacting the cognitive processes respondents engage in when answering an SJT item (see Appendix B). When the PRPR model is applied to SJTs, it suggests respondents engage in the same four psychological processes in which all test and measure respondents engage in. In addition to these core processes, the figure in Appendix B also illustrates several factors which influence a manifest response. For example, sources of true score variance are contributed by the latent individual differences that an SJT is designed to measure such as declarative knowledge and strategic knowledge. Sources of unwanted latent variance also contribute to the total variance of SJT scores. These sources of variance are construct irrelevant (Messick, 1995), and some are method bias (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003).

The model illustrated in Appendix B suggests multiple factors impinge upon a respondent's cognitively-grounded response processes as s/he formulates a response to a given SJT item. Some of these factors contribute contaminating sources of systematic variance to the total variation of SJT scores. The figure presented in Appendix B provides a means of framing some of the issues to consider when conducting research to examine SJTs.

The modified PRPR model presented in Appendix C illustrates the effects of response instructions on a respondent's response processes and thereby on the responses that are summed to scale scores (Ployhart, 2006). The PRPR model suggests the primary determinants of the response processes enacted by respondents are different when 'Should Do' instructions are used than when 'Would Do' instructions are used. Although there are multiple influences on SJT scores, the primary determinant of the processes underlying a response to an SJT item with 'Should Do' instructions is test taker knowledge. In contrast, the primary determinant of the

response processes triggered by an item with 'Would Do' instructions is respondent personality. This offers one explanation for why prior research has consistently documented only a moderate correlation between 'Should Do' and 'Would Do' SJT scores (McDaniel & Nguyen, 2001; Ployhart & Ehrhart, 2003). Moreover, this model offers a theoretically grounded explanation for why the findings of prior research investigating SJT item characteristics suggested SJT response instructions (i.e., 'Should Do', 'Would Do') affected, and may have moderated, the validity of the SJTs examined (e.g., McDaniel et al., 2003; Ployhart & Ehrhart, 2003).

The modified PRPR model suggests SJTs with 'Should Do' instructions will trigger knowledge driven cognitive processes, whereas SJTs with 'Would Do' instructions will elicit personality driven response processes. Given the validity of these assertions, SJT scores derived from knowledge driven response processes (i.e., 'Should Do' instructions) should be more predictive of 'can do' performance than SJT scores derived from personality driven response processes. This is because in maximum performance contexts the volitional behavior of individuals is suppressed such that their motivation to exert effort is uniformly high. Thus, peak performance is primarily driven by ability, knowledge, and expertise. This issue is expounded upon in the next subsection which addresses the nature of maximum performance contexts.

Maximum Performance Contexts

The development or identification of criterion-related assessment tools begins with a systematic explication of both the criterion domain and the context of performance measurement (Campbell, 1990; Kozlowski & Klein, 2000). A number of typologies have been advanced which can assist those charged with conceptualizing and operationalizing criteria. For example, criteria can be conceptualized along three dimensions: (1) short term versus long term, (2) general versus specific, and (3) proximal versus distal in respect to organizational goals (Smith, 1976).

A fourth dimension along which criteria can be scaled is typical versus maximum performance (Kane, 1982; Sackett, Zedeck, & Fogli, 1988). A similar distinction is frequently made in regards to predictor tests and measures (see Cronbach, 1960). Typical and maximum performance measures differ in the degree to which ability versus nonability driven constructs are measured (DuBois, Sackett, Zedeck, & Fogli, 1993). In fact, prior research results suggest there is only a weak correlation between measures of peak and typical performance, even when the measures included in the studies were highly reliable (Sackett et al., 1988).

Typical versus maximum performance contexts differ in the degree to which contextual features constrain the volitional choices of individuals (Lim & Ployhart, 2004). In maximum performance contexts the volitional choices of participant motivation such as: (1) the choice to engage in effort, (2) the choice of what level of effort to expend, and (3) the choice to persist with effort, are constrained such that motivation to exert effort is maximized (DuBois et al., 1993). Each motivational choice is constrained by a corresponding demand characteristic present in maximum performance contexts including: (1) an awareness of being evaluated, (2) the receipt and acceptance of instructions to maximize effort, and (3) a limited measurement time frame in which the evaluated individual can maintain a high level of effort (Sackett et al., 1988).

The distinction between typical and maximum performance is best characterized as a continuum of motivational constraint rather than as a dichotomy (Sackett et al., 1988). Maximum performance contexts primarily trigger what respondents can do rather than what they would typically do (Smith-Jentsch, Salas, & Brannick, 2001). Individuals in a maximum performance context are energized to identify and enact the most effective course of action available. By contrast, typical performance is much more variable and multiply determined because individuals may or may not choose to pursue an optimal course of action. Thus, respondent SJT

scores derived from an SJT with 'Should Do' response instructions will likely be more highly predictive of maximum performance than scores from and SJT with 'Would Do' instructions. This is because 'Should Do' instructions ask respondents to identify the most effective response available and thus observed responses are a result of knowledge driven response processes. By contrast, 'Would Do' instructions result in response processes and thereby SJT scores which are more closely aligned to the behavioral tendencies which characterize typical or long-run performance. This issue is addressed again in the criterion-related validity subsection.

The PRPR model underscores the importance of precisely specifying the nature of the criterion domain and performance context when designing SJTs. SJT item stems, response instructions, and response alternatives should all be carefully constructed to match the nature of performance criteria and the context in which performance occurs. In fact, the *Standards* (1999) state that maximizing the fidelity between test instructions and performance criteria affords another form of validity evidence (American Educational Research Association et al., 1999).

Prior Research Investigating SJT Test Fairness

The preponderance of evidence from 85 years of research suggests tests of general mental ability are often the most valid predictors of both future job performance and learning (Hunter & Hunter, 1984; Schmidt & Hunter, 1998). Yet, the use of cognitive ability tests can result in mean differences of approximately one standard deviation between majority and minority ethnic-based subgroups (Gottfredson, 1988). Thus, when a cognitive ability test is used in a top down selection system, and the selection ratio is low, very few members of the focal subgroup will be selected and disparate impact may ensue. Subgroup differences resulting in adverse impact can undermine the achievement of staffing goals and trigger increased scrutiny of talent management

practices. This presents a quandary for stakeholders who want to simultaneously maximize the validity and utility of assessment tools, create a more diverse workforce, and avoid litigation.

In response to this dilemma, researchers have sought alternative assessment tools such as SJTs to achieve comparable levels of validity to cognitive ability tests while minimizing subgroup differences. The findings from a number of studies suggest the use of SJTs can result in smaller mean subgroup differences than traditional cognitive ability and skill tests (see Chan & Schmitt, 1997; Clevenger et al., 2001; Hanson & Borman, 1995; Motowidlo et al., 1990; Motowidlo & Tippins, 1993; Oswald et al., 2004; Pulakos & Schmitt, 1996; Schmidt, Greenthal, Hunter, Berner, & Seaton, 1977; Schmitt & Mills, 2001; Weekley & Jones, 1999). For example, the results of one study suggested race-based effect sizes for SJTs are approximately one third that of cognitive ability tests (Motowidlo & Tippins, 1993).

Although the above research suggests the use of SJTs is less likely to result in disparate impact than the use of cognitive ability tests, other pertinent findings suggest mean subgroup differences resulting from the use of SJTs are not negligible. For example, research findings have repeatedly suggested gender-based differences on SJTs (Motowidlo et al., 1990; Motowidlo & Tippins, 1993; Weekly & Jones, 1999). The results of these studies suggested females scored up to approximately .20 of a standard deviation higher than males on SJTs. Moreover, Weekley and Jones reported that both African Americans and Hispanics scored .52 and .36 standard deviation units lower on an SJT respectively than Caucasians.

The above findings are informative about mean SJT score differences, but it is also important to consider whether a single regression line is fair for both the focal and reference groups of interest. This question addresses the fairness of an SJT according to Cleary's (1968) definition of test fairness, a standard supported by the U.S. Equal Employment Opportunity

Commission. The present study addressed this question by using moderated multiple regression to examine whether an SJT that was designed to measure declarative and strategic knowledge was a fair assessment tool for gender and ethnic-based majority and minority subgroup members.

CHAPTER THREE: THE PRESENT STUDY

The present study extended both Ployhart and Ehrhart's (2003) and McDaniel and colleagues' (2003) research findings by applying the PRPR model to examine the validity evidence for, and fairness of, an SJT in a maximum performance context. A between-subjects experimental design was used to investigate three questions about the construct validity, criterion-related validity, and fairness of an SJT when its response instructions were manipulated and maximum decision making performance outcomes were objectively measured. The first research question asked if response instructions (i.e., 'Should Do', 'Would Do') moderated the validity of an SJT that was designed to measure declarative and strategic knowledge. The second question asked about the upper-bound criterion-related validity coefficient for SJTs in contexts such as talent selection in which typical performance is the criterion of interest. The third question asked whether the SJT used in this present study was fair for gender and ethnic-based subgroups according to the standards set forth in Cleary's (1968) definition of test fairness.

To answer the above questions, participant's demographic characteristics and propensity to take risks were measured. Next, participants completed a brief computer-based training module. This self-paced presentation provided basic facts about Texas Hold'em poker and the poker simulation completed later in the study. After training, each participant completed a traditional multiple-choice test of declarative knowledge of Texas Hold'em poker. Participants were then randomly assigned to complete one of the two SJT versions. Both SJT versions were designed with identical item content and response alternatives but had different response instructions (i.e., 'Should Do', 'Would Do'). The instructions of the 'Should Do' SJT asked participants to identify the option they *should do* in terms of an optimal response to the scenario.

The instructions of the 'Would Do' SJT asked participants to identify the option they *would do* in terms of the response they would actually choose in the situation.

There were two criteria included in the present research, aggregate and average performance outcomes. The aggregate performance outcome indexed the total dollar amount each participant had at the end of the poker simulation. The average performance outcome indexed the mean dollar amount of each participant at the end of the simulation. Both of these dependent variables were outcomes of maximum decision making performance. Decision making has been defined as "the ability to gather and integrate information, use sound judgment, identify alternatives, select the best solution, and evaluate the consequences" (Cannon-Bowers, Tannenbaum, Salas, & Volpe, 1995, p. 346). Decision making in a Texas Hold'em poker game is an unfolding process whereby a decision maker draws upon his or her declarative and strategic knowledge to identify and diagnose an unfolding situation, recognize familiar patterns, allocate financial resources to capitalize on presented opportunities, and evaluate the outcome(s) of resource allocation decisions. The purpose of this activity is to maximize one's monetary winnings and thereby secure additional financial resources for future investments.

The remainder of this section consists of several subsections which detail the evidence that was collected in support of the SJT used in the present study. A description of the construct validity evidence, criterion-related validity evidence, and fairness evidence collected is provided and hypotheses are advanced.

Construct Validity Evidence

A construct is "some postulated attribute of people assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 281). Validity is "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and

appropriateness of interpretations and actions on the basis of test scores or other modes of assessment" (Messick, 1995, p. 741). Construct validity is defined in terms of whether a measure or test allows for "accurate inferences about an individual's standing on a psychological construct of particular interest" (Binning & Barrett, 1989, p. 479). As noted by Nunnally and Bernstein (1994, pp. 86-87), the process of construct validation involves three steps including:

> …(1) specifying the domain of observables related to the construct; (2) determining the extent to which observables tend to measure the same thing, several different things, or many different things from empirical research and statistical analyses; and (3) performing individual differences studies and/or experiments to determine the extent to which measures of the construct are consistent with best guesses about the construct.

The SJT used in the present study purported to measure knowledge of Texas Hold'em poker. This knowledge was conceptualized as being multidimensional, comprised of both declarative and strategic aspects. Declarative knowledge is knowledge about performance relevant tasks and behaviors (Campbell et al., 1993). Exemplars of declarative knowledge include knowledge of principles, facts, goals, and self (Campbell et al., 1993). In contrast, strategic knowledge is information about "…why, when, and where to apply one's knowledge and skills" (Bell & Kozlowski, 2002, p. 274).

In the present study, inferences were drawn about a participant's standing on the constructs of declarative and strategic knowledge of Texas Hold'em poker on the basis of their SJT score. Inferences were also drawn about participants' declarative knowledge on the basis of their multiple choice test of declarative knowledge score. Similarly, participants' propensity to take risks was inferred from their scale score on the self-report measure of risk taking.

SJT & Declarative Knowledge Test

As part of a broader effort to generate construct validity evidence for the SJT, the present study included a traditional multiple-choice test of declarative knowledge. This test was designed to measure participant's knowledge of basic facts about Texas Hold'em poker such as how many cards are dealt at specific points in the game and which hands are better than others. Given that both the SJT and the traditional multiple-choice test purported to measure declarative knowledge, participants' scores on these two assessment tools were expected to be positively correlated. The correlation between participants' scores on the two tests was calculated to generate convergent validity evidence for the SJT used in the present study.

Construct validity evidence for the SJT used in the present study was also generated by examining its response instructions. Of particular interest was whether SJT response instructions moderated the relationship between declarative knowledge as measured by the traditional multiple-choice test and knowledge as measured by the SJT. As suggested by the PRPR model, declarative knowledge was expected to be more highly correlated with SJT scores when they are derived from 'Should Do' response instructions.

According to the PRPR model, there are multiple sources of systematic variance that influence the response processes in which respondents engage when formulating a response to an SJT item (Ployhart, 2006). For example, response instructions may contribute a contaminating source of systematic variance to SJT scores because they influence the proximal determinants of the response processes respondents engage in when formulating responses to SJT items. As noted in the conceptual foundation section, 'Should Do' response instructions trigger response processes that are primarily driven by respondent knowledge of what ought to be done in response to an SJT scenario. In contrast, 'Would Do' instructions trigger personality driven

response processes (see Appendix C). Thus, when SJT scores are derived from the use of 'Would Do' response instructions, systematic variance in those scores may be due to both the construct(s) the SJT was designed to measure (e.g., declarative and strategic knowledge) and other constructs it was not designed to measure such as respondent personality characteristics. The systematic error variance attributable to personality in 'Would Do' SJT scores may serve to attenuate the relationships between this assessment tool and other assessment tools that do not also measure similar constructs. Given this conjecture, the following hypothesis was advanced:

> Hypothesis 1: Declarative knowledge was expected to correlate more highly with 'Should Do' SJT scores than 'Would Do' SJT scores.

SJT & Risk Taking Questionnaire

A self-report measure of risk taking (see Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, & Gough, 2006) was also included in the present study in order to help generate construct validity evidence for the newly developed SJT. Individuals with a propensity for taking risks are more likely to choose suboptimal courses of action. There are multiple reasons why risk takers may not do what should be done in a situation. For example, risk takers enjoy adventure and are inclined to seek the danger and excitement of speculative, high risk investments. In poker, these tendencies can result in the use of tactics that may be effective in any given situation, but when used routinely ultimately undermine long-run performance. For example, a poker player prone to taking risks is much more likely to raise or call with a hand of cards that has a low probability of winning in an attempt to bluff. Risk takers also tend to break more rules than their counterparts so they may be more likely to ignore poker rules of thumb such as never chase a straight or flush. Moreover, risk takers are typically more reckless and this can translate into unnecessarily aggressive wagers that expose betters to large losses relative to their chip

total. If risk takers choose options that depart from what ought to be done in a series of situations, they will have lower SJT scores, which may, in turn, result in a negative correlation between risk taking and SJT scores. Thus, it was expected that there would be a negative correlation between risk taking and SJT scores.

Also of interest was whether SJT response instructions moderated the relationship between risk taking and knowledge as measured by the SJT. In light of the PRPR model, risk taking may interact with SJT response instructions when predicting knowledge as measured by the SJT such that risk taking is more predictive of SJT scores when those scores are derived from 'Would Do' response instructions than when they are derived from 'Should Do' response instructions. As noted, 'Would Do' instructions primarily trigger personality driven response processes, whereas 'Should Do' instructions primarily elicit knowledge driven response processes (see Appendix C). Thus, when SJT scores are derived from the use of 'Would Do' response instructions, systematic variance in those scores may be due to both the construct(s) the SJT was designed to measure (e.g., declarative and strategic knowledge) and other personality-based characteristics such as risk taking that it was not designed to measure. The systematic error variance attributable to personality in 'Would Do' SJT scores may serve to inflate the observed relationship between this assessment tool and other assessment tools that also assess personality constructs such as the risk taking measure that was included in the present study.

Of note, however, the above theoretical rationale is contrary to meta-analytic findings suggesting SJT scores derived from knowledge instructions were more strongly correlated to openness to experience than scores derived from behavioral tendency instructions (McDaniel et al., 2003). These findings were, however, based on just five effect sizes, an insufficient sample to

reach firm conclusions about the construct validity of the SJTs investigated. Given the totality of theoretical and empirical evidence presented above, the following hypothesis was advanced:

> Hypothesis 2: Risk taking was expected to correlate more highly with 'Would Do' SJT scores than 'Should Do' SJT scores.

Criterion-related Validity Evidence

This subsection discusses the criterion-related validity evidence generated in support of the SJT investigated in the present research study. The criterion-related validity evidence generated by testing the hypotheses advanced below also served to contribute to an overall understanding about the construct validity of the SJT examined (see Nunnally & Bernstein, 1994). There are two criterion-related validity issues addressed in this subsection. The first issue addresses whether response instructions moderate the relationship between participant SJT scores and maximum decision making performance outcomes. The second issue addresses the value added by the SJT over the traditional multiple-choice declarative knowledge test when predicting peak performance. More specifically, the second issue addresses the incremental variance in maximum decision making performance outcomes that was explained by the SJT.

SJT & Maximum Decision Making Performance

As noted in the conceptual foundation section, prior research findings have suggested that response instructions may moderate the criterion-related validity of SJTs (McDaniel et al., 2003; Ployhart & Ehrhart, 2003). Unfortunately, the results of these studies were mixed and because the SJTs used in these studies measured somewhat indeterminable constructs, the meaning of these findings is difficult to discern. The present research study sought to leverage the PRPR model to advance the current understanding about the criterion-related validity of an SJT that

measured declarative and strategic knowledge and 'can do' performance criteria. In the present study, SJT response instructions were manipulated and participant maximum decision making performance outcomes were objectively measured to determine if SJT scores interacted with response instructions when predicting peak performance criteria.

The suppositions advanced in the conceptual foundation section of this text about the latent cognitive processes which give rise to a response to an SJT item collectively suggest SJT scores may interact with SJT response instructions when predicting peak performance outcomes (Ployhart, 2006). This is because response instructions influence a respondent's comprehension, retrieval, judgment, and response selection processes during the response to an SJT item (see Appendix C). The PRPR model suggests that even when an SJT with 'Would Do' instructions has been purposively designed to measure a performance-related construct such as strategic knowledge, its response instructions will still trigger personality driven response processes. Thus, when SJT scores are derived from the use of 'Would Do' response instructions, systematic variance in those scores may be due to both the construct(s) the SJT was designed to measure and other personality-based characteristics that were not targeted for measurement.

The systematic error variance attributable to personality in 'Would Do' SJT scores may serve to attenuate relationships between this assessment tool and peak performance criteria such that SJT scores derived from 'Would Do' instructions are less strongly correlated with maximum performance outcomes than SJT scores derived from 'Should Do' instructions. This is because a larger proportion of the systematic variance in 'Should Do' SJT scores is accounted for by constructs such as declarative and strategic knowledge which empirical evidence suggests are predictive of 'can do' performance. In regards to this issue, Ployhart and Ehrhart (2003) stated that "'Should Do' measures might best predict maximum performance or performance during

high transitions, whereas 'Would Do' measures might be most predictive of stable performance" (p. 13). Given these suppositions, a noncrossing ordinal interaction was expected between SJT scores and SJT response instructions when predicting both aggregate and average peak performance outcomes. Based on this conjecture, the following hypothesis was advanced:

> Hypothesis 3: 'Should Do' SJT scores were expected to correlate more highly with peak performance criteria than 'Would Do' SJT scores.

SJT vs. Traditional Test of Declarative Knowledge

SJTs can be used as assessment tools to measure a wide variety of performance-related constructs such as personality characteristics, cognitive abilities, and job knowledge. For example, the present study used an SJT that was designed to measure declarative and strategic knowledge. Given their complexity, however, SJTs take longer and cost more to develop than traditional multiple-choice tests of knowledge. Thus, an important question is whether SJTs that measure knowledge can explain incremental variance in organizationally valued criteria beyond that already explained by less expensive and readily available methods of measuring knowledge. Answering this question sheds light on the 'value added' of using SJTs as assessment tools.

There has been no prior research of which the author is aware that has investigated the incremental variance in maximum decision making performance outcomes that is accounted for by an SJT that measures declarative and strategic knowledge. However, there was reason to believe that the SJT would account for additional variance in maximum decision making performance outcomes that was not already explained by the traditional test of declarative knowledge. This is because the SJT was designed to be a multidimensional assessment tool, measuring both declarative and strategic aspects of knowledge. Strategic knowledge goes beyond the mere memorization of facts or declarative knowledge as it involves a deeper level processing

of information which ultimately results in an understanding of why, when, and where to apply one's resources (Bell & Kozlowski, 2002; Fritzsche, Stagl, Burke, & Salas, under review).

Participant 'Should Do' SJT scores should account for incremental variance in maximum decision making performance outcomes beyond that already explained by the traditional test of declarative knowledge because the SJT used in the present study better represented the criterion dimensionality of maximum decision making performance (see Schmitt & Chan, 2006). The SJT used in this research was comprised of scenarios that required participants to have an understanding of why a particular response option was effective (i.e., strategic knowledge) in order to know which of the presented response options was the most effective (i.e., declarative knowledge). This deeper appreciation of the task domain is critical to making a sequence of effective decisions in a maximum performance context. A similar argument can not be advanced for the 'Would Do' SJT version used in the present study because the response instructions of this assessment tool trigger personality-based response processes. Given these assertions, the following hypothesis was advanced:

> Hypothesis 4: It was expected that knowledge as measured by the 'Should Do' SJT would account for incremental variance in maximum decision making performance outcomes beyond that explained by declarative knowledge as measured by the traditional multiple-choice test.

Test Fairness

As noted in the conceptual foundation section, repeated calls have been made for studies to investigate possible sex and ethnic-based mean subgroup differences in SJT scores (Hanson & Ramos, 1996; McDaniel & Nguyen, 2001; Weekly & Jones, 1999; Weekley et al., in press). This line of research is undoubtedly important to conduct, as one of the oft noted advantages of SJTs

is that their use produces comparable levels of validity but smaller subgroup differences than other widely vaunted assessment tools such as cognitive ability tests (Chan & Schmitt, 1997; Clevenger et al., 2001; Hanson & Borman, 1995; Motowidlo et al., 1990; Motowidlo & Tippins, 1993; Oswald et al., 2004; Pulakos & Schmitt, 1996; Schmidt, Greenthal, Hunter, Berner, & Seaton, 1977; Schmitt & Mills, 2001; Weekley & Jones, 1999).

While it seems the use of SJTs can result in smaller mean subgroup differences than cognitive ability tests, reported mean SJT differences between focal and reference groups are not negligible. For example, findings suggested females scored up to approximately .20 of a standard deviation higher than males on SJTs (Motowidlo et al., 1990; Motowidlo & Tippins, 1993; Weekly & Jones, 1999). Research findings also suggested that African Americans and Hispanics scored one half and one third standard deviation units lower, respectively, than Caucasians on SJTs (Weekley & Jones, 1999). The results of these studies raise the question of whether or not these observed predictor score differences are proportional to criterion score differences and independent of the level of the predictors examined. The answer to this question speaks to the fairness of an SJT according to the standards of Cleary's (1968) definition of test fairness.

There was reason to believe that the cognitively loaded nature of an SJT measuring both declarative and strategic knowledge would produce an assessment tool that was ultimately biased for certain subgroups. It is a well documented finding that the use of cognitive ability tests results in large subgroup differences (Gottfredson, 1988) and that cognitive ability is a direct determinant of knowledge (Campbell, 1990). Thus, it seemed plausible that an SJT which measured declarative and strategic knowledge would result in an assessment tool that was ultimately biased for gender and ethnic-based subgroups according to Cleary's (1968) definition of test fairness. To investigate the fairness of the SJT used in this present study moderated

multiple regression analyses were conducted to examine whether the regression equations for the focal and reference groups under consideration were statistically equivalent. These analyses were conducted to examine whether knowledge, as measured by the SJT, interacted with gender and/or ethnicity when predicting maximum decision making performance outcomes.

Prior to using moderated multiple regression, power analyses were conducted which are detailed in a later subsection. Unfortunately, the limited base of prior research examining the fairness of SJTs made it difficult to meaningfully estimate the expected effect size of the interaction between SJT scores and either gender or ethnicity when predicting maximum performance outcomes. Because the expected interaction effect size was unknown, it was difficult to accurately estimate the sample size needed in the present study to achieve a desirable level of power to detect an effect if one existed. Given the lack of actionable insight that can be distilled from the current body of research about the expected interaction effect size, and thereby how many participants were needed to detect an interaction if one did exist, a decision was made to investigate the fairness of the SJT used in the present study for gender and ethnic-based subgroups on an exploratory basis. Given the above rationales, the following exploratory hypothesis was advanced about the fairness of the SJT used in the present study:

> Hypothesis 5: It was expected that SJT scores would interact with gender and ethnicity when predicting maximum decision making performance outcomes such that the slopes would be different for the focal and reference groups investigated.

CHAPTER FOUR: METHOD

This section details the method used in the present between-subjects experiment. The power analyses, participants, procedure, performance context, predictor measures, and criterion domain are discussed.

Power Analyses

Several power analyses were conducted to estimate the number of participants required to detect the various main effects and interactions that were discussed in the previous subsections. The first set of power analyses were conducted to determine the number of participants that were required to detect a medium to large main effect at $p < .05$, one-tailed. These analyses were geared toward detecting a medium to large effect because prior research suggested this is approximately the magnitude of the effect size for SJTs when predicting typical performance criteria (see McDaniel et al., 2003; Ployhart & Ehrhart, 2003). Moreover, these analyses were geared toward a one-tailed test because directional hypotheses were advanced.

The power analyses suggested a sample of 68 participants was required to have an 80% chance of detecting a medium effect at $p < .05$, one-tailed. In contrast, only 22 participants were needed to have an 80% chance of detecting a large effect at $p < .05$, one-tailed (Cohen, 1977). Splitting the difference between the results of these two power analyses suggested approximately 45 participants were required to have an 80% chance of detecting a hypothesized moderate to strong main effect. This information is pertinent to testing some of the hypotheses in this study.

Of note, however, many of the hypotheses advanced in the previous section were about expected interactions between variables (e.g., response instructions moderating the SJT peak performance relationship). In regards to this issue, the findings from a simulation-based study examining the power of moderated multiple regression to detect interactions suggested that

strong ordinal interactions can be detected with as few as 48 participants, while weak ordinal interactions require up to 240 participants to detect (Stone, Austin, & Shetzer, 1986). Given the general lack of SJT research addressing the specific contingent relationships investigated in the present study, it was difficult to discern the magnitude of the expected interaction effect size, and thereby how many participants were needed to detect a hypothesized interaction if one did truly exist. Ultimately, the results of the above noted power analyses and prior research findings were considered thoroughly and 110 participants were recruited for participation in the present study.

Participants

Participants were recruited from the student body of a large public university located in the Southeast. The demographic composition of the obtained sample was fairly evenly split between genders (63% female); young ($M = 20.03$, $SD = 2.86$); and ethically diverse (60% Caucasian, 18% Hispanic, 13% African American, 7% Asian). Moreover, 34 different nationalities and combinations of nationalities were represented in the achieved sample. Although participants were predominantly in their first year of college (49%), sophomores (24%), juniors (11%), seniors (12%), and graduate students (4%) were also represented. Finally, almost 40 college majors were represented in the achieved sample included in the present study.

Procedure

Participants completed the present study in a private laboratory on the main campus of a large public university located in the Southeast. Upon arrival, participants were introduced to their experimenter, surroundings, and planned activities via scripted dialogue. Participants were then asked to complete an informed consent form. After completing the informed consent form, participants were asked to complete a demographics questionnaire and risk taking measure.

Each participant then completed a self-paced computer-based training session that lasted approximately ten minutes. The purpose of this training module was to impart knowledge about Texas Hold'em poker and to familiarize participants with the Texas Hold'em poker simulation that was used later in the study. Following the training session, participants were asked to complete a traditional multiple choice test of declarative knowledge. Next, participants were randomly assigned to complete one of the two versions of the SJT (i.e., 'Should Do', 'Would Do') used in the present study. Afterwards, each participant was asked to participate in a moderate fidelity computer-based simulation of Texas Hold'em poker. At the conclusion of the simulation, participants were fully debriefed and any questions or issues which arose during the experiment were addressed via a two-way dialogue.

Performance Context

As noted in the conceptual foundation section, the present study was characterized by several demand characteristics which served to constrain the motivational choices of participants and thereby helped foster a maximum performance context. In maximum performance contexts, the volitional choices of participant motivation including: (1) the choice to engage in effort, (2) the choice of what level of effort to expend, and (3) the choice to persist with effort, are constrained (Sackett et al., 1988). Sackett and colleagues' assert that each of these three motivational choices is constricted by a corresponding demand characteristic present in maximum performance contexts including: (1) an awareness of being evaluated, (2) the receipt and acceptance of instructions to maximize effort, and (3) a limited time frame in which an individual can maintain a high level of effort (Sackett et al.).

Each of the three above noted demand characteristics were present in the proposed experiment. For example, participants were consciously aware they were being evaluated as part

of their participation in the laboratory-based experiment. Moreover, participants were instructed to maximize their effort. In an effort to help ensure instructions to maximize effort were accepted, participants were informed that the top performing individual would be rewarded with an Apple iPod©. Finally, the poker simulation lasted no longer than 10 minutes, a limited time frame in which participants could energize and maintain a high level of effort.

Evidence for the establishment of a maximum performance context was collected from a variety of sources. For instance, participants had to make an effort to arrive at the campus locale the study was conducted at on time. Also, experimenter reports suggest that each participant expressed excitement about the opportunity to win an Apple iPod©. Moreover, all participants who started the approximately one hour long study finished it, evidencing their persistence. Given this evidence, it seems reasonable to assume that the motivational choices of participants were constrained, and a maximum performance context was created, in the present study.

Predictor Measures

Three predictor measures were used in the present research study; a traditional multiple-choice test of declarative knowledge, a self-report measure of risk taking, and an SJT that was designed to measure declarative and strategic knowledge of Texas Hold'em poker. Participant's aggregate and average maximum decision making performance outcomes were objectively indexed as peak performance criteria.

The traditional multiple-choice declarative knowledge test was comprised of seven items. Each of these items included a single sentence item stem and four response options. This assessment tool was administered to participants after they completed the computer-based training module but prior to the Texas Hold'em poker simulation. Participant's responses to the

items of this test were scored as either correct or incorrect. One point was awarded for a correct response. Participant's correct responses were summed to calculate their scale score.

A self-report paper-and-pencil measure of risk taking was also included in the present study (see Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, & Gough, 2006). The risk taking measure was obtained from the archives of the International Personality Item Pool, a scientific collaboratory for the development of personality and individual difference measures (see http://ipip.ori.org). The risk taking measure included 10 items, each of which was presented on a 7 point Likert scale. This measure was administered to participants after they completed the demographics questionnaire but prior to the start of the computer-based training module.

The forced-choice SJT used in the present study consisted of 6 items with 3 response options per item. These items were presented in a paper-and-pencil format. Item stem consisted of a paragraph long hypothetical Texas Hold'em poker situation. The SJT was designed via a construct oriented approach as described in the construct validity section above. A multiple correct answer scoring scheme was used with the SJT. Participants received 3 points for choosing the response designated as most effective, 2 points for choosing the next most effective response option, and 1 point for choosing the least most effective response option. Three subject matter experts were consulted to determine the relative effectiveness of the response options.

Two versions of the SJT were created, one with 'Should Do' response instructions and the other with 'Would Do' response instructions. The response instructions of the 'Should Do' SJT asked participants to identify the response option that they *should do* in terms of an optimal response to the posed scenario. A should do approach has been widely used in research investigating SJTs (e.g., Phillips, 1993; Reynolds, Winter, & Scott, 1999; Strong & Najor, 1999; Weekley & Jones, 1997). In contrast, the response instructions of the 'Would Do' SJT asked

participants to identify the option they *would do* in terms of the response they would actually choose in response to the posed situation. A 'Would Do' approach was used by Bruce and Learner (1958) to develop a supervisory practice test.

The SJT used in the present study was designed to measure knowledge of Texas Hold'em poker. The construct domain measured by the SJT was conceptualized as multidimensional, comprised of both declarative and strategic knowledge. Declarative knowledge is knowledge about performance relevant tasks and behaviors (Campbell et al., 1993). The SJT was used to scale participant declarative knowledge of basic facts about Texas Hold'em poker like which hands are superior to other hands. Moreover, the scenarios comprising the SJT were designed so that respondents needed an understanding of why particular actions were appropriate in order to receive the maximum amount of points for a given item. Thus, the SJT used in the present study was also designed to scale participant's strategic knowledge (see Bell & Kozlowski, 2002).

The SJT used in the present study was designed to measure knowledge because prior taxonomic efforts and empirical research findings suggest knowledge is one of the three proximal antecedents of task performance (Campbell, 1990). In fact, some have argued that it is likely most SJTs used in organizational settings as part of broader talent selection, development, or retention initiatives contain a knowledge component (Schmidt, 1994; Schmidt & Hunter, 1993). Thus, designing the SJT used in this research to measure knowledge is consistent with prior assertions that most SJTs assess knowledge.

In order to generate some initial empirical support the newly developed SJT, both versions of it were piloted in a small sample (N = 10) of incumbents at an applied research institute in the Southeast using a within subjects design. The results of this preliminary analysis indicated that both the 'Should Do' (r = .705) and 'Would Do' (r = .730) SJT versions had

acceptable levels of internal consistency reliability for research purposes. While preliminary, this evidence provided initial support for the assertion that the SJT measured knowledge of poker. These initial estimates were used to revise and tailor the SJT for use in the present study.

Criterion Domain & Indices

In the present study, decision making performance outcomes were captured during a moderate fidelity Texas Hold'em poker simulation. Researchers have defined decision making as "the ability to gather and integrate information, use sound judgment, identify alternatives, select the best solution, and evaluate the consequences" (Cannon-Bowers, Tannenbaum, Salas, & Volpe, 1995, p. 346). In a Texas Hold'em poker game, decision making is an unfolding process whereby a decision maker draws upon his or her knowledge to diagnose and capitalize on presented opportunities by allocating financial resources. The goal of the decision maker in a poker game is to maximize one's monetary winnings while minimizing exposure to losses.

Effective decision making in Texas Hold'em poker demonstrates job-specific task proficiency (see Campbell, McCloy, Oppler, & Sager, 1993). Job-specific task proficiency is one of eight factors which have been argued to comprise individual performance. Campbell and colleagues' suggest individual performance can be represented via a hierarchical taxonomy. At the highest level of this taxonomy, performance is comprised of eight factors including: job-specific task proficiency, non-job-specific task proficiency, written and oral communication proficiency, demonstration of effort, maintenance of personal discipline, facilitation of peer and team performance, supervision/leadership, and management/administration.

The first factor in this taxonomy of performance, job-specific task proficiency, represents the level of proficiency with which an individual can perform the substantive technical tasks that distinguish the core content of one job from another. As noted by Campbell "The question of

how well individuals can do such tasks is meant to be independent of their level of motivation…" (1990, p. 709). Thus, the only opportunity to meaningfully assess job-specific task proficiency is when an individual's volitional choices are constrained (i.e., in a maximum performance context), such as in the present study.

The two criteria included in the present research objectively indexed participant's aggregate and average maximum decision making performance outcomes. The aggregate performance outcome indexed the total amount of money each participant had at the end of the poker simulation. The average performance outcome indexed the total amount of money each participant had at the end of the simulation divided by the nine hands that were completed by all participants. Both of these dependent variables were the outcomes of several sequential maximum decision making performance episodes.

CHAPTER FIVE: FINDINGS

The results are organized around 6 subsections: (1) data screening, (2) scale descriptive statistics, (3) internal consistency reliability, (4) construct validity, (5) criterion-related validity, and (6) test fairness. The statistical analyses were generated using SPSS© version 12.0 for Windows©. A majority of the hypotheses advanced in the present experiment dealt with the interactions amongst independent variables when predicting dependent variables. These hypotheses were tested via the use of moderated multiple regression (MMR) analysis. MMR analysis was used to determine if SJT response instructions moderated the relationships between: (1) declarative knowledge and SJT scores, (2) risk taking and SJT scores, (3) SJT scores and an aggregate performance outcome, and (4) SJT scores and an average performance outcome. Moreover, MMR was also used to determine if gender and/or ethnicity moderated the relationships between SJT scores and both aggregate and average performance outcomes.

Data Screening

Prior to hypothesis testing, the data were screened for errors, missing data, and outliers. Each successive entry into the computerized data file was compared against the original data provided by participants. In this manner, 100% of the data was visually checked. During this process, two data entry errors were identified and corrected. The range of data values were also examined to ensure that all entries were within acceptable limits. All values were within range.

Next, the data set was screened for missing data. Three variables were identified as having missing data in the data set. Specifically, one case was missing ‘Should Do’ SJT data, one case was missing ‘Would Do’ SJT data, and two cases were missing average performance data. All of these missing values were randomly dispersed across different variables and cases

and thus all cases with missing data were dropped from subsequent analyses per the recommendation of Tabachnick and Fidell (2001).

The final step of data screening was conducted to identify and treat cases with a univariate outlier (i.e., an extreme value on one variable). As recommended by Tabachnick and Fidell (2001), outliers were identified by inspecting histograms and by identifying standardized scores in excess of 3.29. Three cases were found to have a univariate outlier value on the traditional test of declarative knowledge (-3.51, -3.51, -3.51). One case was found to have a outlier value on the risk taking measure (3.43). Six cases were found to have an outlier value on the aggregate performance measure (-3.87, 4.02, -3.55, -3.72, 4.02, 4.02). Two cases had an outlier on the average performance measure (-5.08, -3.29). After cases with missing data and univariate outliers were dropped, complete data were available for: 105 participants for the declarative knowledge test, 107 participants for the risk taking measure, 51 participants for the 'Should Do' SJT, 55 participants for the 'Would Do' SJT, 102 participants for the aggregate performance variable, and 104 participants for average performance variable.

The participants in the present study were instructed to do their best in the Texas Hold'em poker simulation in terms of maximizing their monetary winnings without taking unnecessary risks. It seems 6 of the 108 participants made unnecessarily large wagers (i.e., went all in) during a single hand of the nine card hands played during the simulation. These large wagers meant that they either lost most, if not all, of their allotted $5,000 in a single hand, or that their opponent lost most, if not all, of their money in a single hand. Because this type of betting was not consistent with the instructions provided to participants, the 6 outliers were deleted from subsequent analyses involving the aggregate performance measure and the 4 outliers were deleted from subsequent analyses involving the average performance measure. Because these

values were so far outside the range of the remainder of the performance values, their inclusion would serve to distort statistical inferences about the variables measured in the present study (Cohen, Cohen, West, & Aiken, 2003; Stevens, 1984).

Scale Descriptive Statistics

Descriptive statistics and variable intercorrelations were calculated for the tests and measures used in this study (see Table 1). Specifically, means, standard deviations, internal consistency reliability estimates, and variable intercorrelations are presented in Table 1. Internal consistency reliability estimates are discussed in the next subsection. Indices of skew and kurtosis are not presented in Table 1. Of note, however, neither the 'Should Do' (-.106) or 'Would Do' (.001) SJT distributions were significantly skewed. The kurtosis was -1.119 for the 'Should Do' SJT and .045 for the 'Would Do' SJT. The observed means were similar for the 'Should Do' (M = 12.90) and 'Would Do' (M = 12.15) SJTs. Moreover, the standard deviations of the 'Should Do' (SD = 2.435) and 'Would Do' (SD = 2.094) SJTs were also similar.

Neither the 'Should Do' SJT version (r = .180, $p \geq .05$) or the 'Would Do' SJT version (r = .161, $p \geq .05$) was significantly correlated with declarative knowledge as measured by the traditional multiple choice test. In contrast, both the 'Should Do' SJT version (r = -.254, $p \leq .05$) and the 'Would Do' SJT version (r = -.369, $p \leq .05$) were correlated with risk taking. One apparent difference in the nomological network of the two SJT versions was their respective relationships to both aggregate and average maximum decision making performance outcomes. For example, participant scores on the 'Should Do' SJT version were more predictive of the aggregate maximum performance outcome (β = .478, $p \leq .01$) than 'Would Do' SJT scores (β = -.158, $p \geq .05$). Similarly, the 'Should Do' SJT was more predictive of the average maximum performance outcome (β = .346, $p \leq .01$) than the 'Would Do' SJT (β = -.048, $p \geq .05$).

Internal Consistency Reliability

In the present study, Cronbach's coefficient alpha was used to estimate the internal consistency reliability for the risk taking questionnaire (.776), declarative knowledge test (.494), 'Should Do' SJT (.546), and 'Would Do' SJT (.348). The low internal consistency reliability estimate for the declarative knowledge test is not surprising, given that this tool was only comprised of seven items and there was a low standard deviation (*SD* = .964) on this scale in this sample. The internal consistency reliability estimate for the 'Should Do' SJT (.546) is similar in magnitude to the estimate for the 'Should Do' SJT (.520) provided by Ployhart and Ehrhart (2003). Conversely, the internal consistency reliability estimate for the 'Would Do' SJT (.348) used in the present study is somewhat smaller in magnitude than the estimate for the 'Would Do' SJT (.570) investigated by Ployhart and Ehrhart (2003). The internal consistency estimates for both the 'Should Do' and 'Would Do' SJTs are likely low because these tests only used 6 items. This issue, and the others noted above, are expounded upon in the discussion section.

Construct Validity Evidence

The first hypothesis examined in this study stated declarative knowledge would interact with SJT response instructions when predicting participant SJT scores such that declarative knowledge was more predictive of 'Should Do' SJT scores than 'Would Do' SJT scores. Specifically, it was asserted that SJT response instruction type (i.e., 'Should Do', 'Would Do') would moderate the relationship between declarative knowledge as measured by the traditional multiple-choice test and SJT scores such that declarative knowledge would explain more variance in 'Should Do' SJT scores than in 'Would Do' SJTs. The results of MMR analysis which was used to determine if SJT response instructions moderated the relationship between declarative knowledge and SJT scores did not support hypothesis 1. Specifically, when SJT

scores were regressed on declarative knowledge (β = .142, p ≥ .05), response instructions (β = .137, p ≥ .05), and the interaction term (β = .040, p ≥ .05), a multiple correlation of .237 was obtained. The overall F test ($F_{(3,99)}$ = 1.967, p ≥ .05) and specific t-tests for the beta weights were not significant. Thus, hypothesis 1 was not supported in the present study.

The second hypothesis tested in this study stated risk taking would interact with SJT response instructions when predicting participant SJT scores such that risk taking was more predictive of 'Would Do' SJT scores than 'Should Do' SJT scores. In other words, it was expected that SJT response instructions (i.e., 'Should Do', 'Would Do') would moderate the relationship between risk taking scores and SJT scores. The results of MMR analysis conducted to determine if SJT response instructions moderated the relationship between risk taking and SJT scores failed to support hypothesis 2. When SJT scores were regressed on risk taking (β = -.375, p ≤ .05), response instructions (β = .166, p ≥ .05), and the interaction term (β = .101, p ≥ .05), a multiple correlation of .352 was obtained. The overall F test ($F_{(3,101)}$ = 4.752, p ≤ .05) and t-test for the risk taking beta weight were significant but the beta weights for the moderator and interaction terms were not. Thus, hypothesis 2 was not supported in this study.

Criterion-related Validity Evidence

The third hypothesis tested in this study stated that knowledge measured by the SJT would interact with SJT response instructions when predicting aggregate and average maximum decision making performance outcomes such that 'Should Do' SJT scores are more predictive of peak performance criteria than 'Would Do' SJT scores. Specifically, it was expected that SJT response instructions (i.e., 'Should Do', 'Would Do') would moderate the relationship between SJT scores and both aggregate and average maximum performance criteria. The results of MMR analyses conducted to determine if SJT response instructions moderated the relationship between

knowledge as measured by the SJT and 'can do' performance supported hypothesis 3. When the aggregate maximum performance outcome variable was regressed on SJT scores (β = -.210, p ≥ .05), response instructions (β = -.116, p ≥ .05), and the interaction term (β = .399, p ≤ .05), a multiple correlation of .295 was obtained. The overall F test ($F_{(3,96)}$ = 3.041, p ≤ .05) and t-test for the interaction term beta weight were both significant. This suggests SJT scores interacted with response instructions when predicting the aggregate maximum performance outcome such that 'Should Do' SJT scores (β = .478, p ≤ .05) were more predictive of peak performance than 'Would Do' SJT scores (β = -.158, p ≥ .05), even though both SJTs had identical item content and the same response alternatives. This evidence supported hypothesis 3.

The second part of hypotheses 3 was concerned with the prediction of the average maximum performance outcome via participant SJT scores. When the average maximum performance outcome variable was regressed on SJT scores (β = -.049, p ≥ .05), response instructions (β = -.227, p ≤ .05), and the interaction term (β = .288, p = .05), a multiple correlation of .323 was obtained. The overall F test ($F_{(3,98)}$ = 3.810, p ≤ .05) and the specific t-tests for the moderator and interaction terms were significant. This suggests that SJT scores interacted with response instructions when predicting the average maximum performance outcome such that the 'Should Do' SJT (β = .346, p ≤ .05) was more predictive of peak performance than the 'Would Do' SJT (β = -.048, p ≥ .05), even though both SJTs had identical item content and the same response alternatives. This evidence also supported hypothesis 3.

The fourth hypothesis tested in this study stated that knowledge as measured by the 'Should Do' SJT version would account for incremental variance in maximum decision making performance outcomes beyond that explained by declarative knowledge as measured by the traditional multiple-choice test. It was expected that 'Should Do' SJT scores would explain

incremental variance in both the aggregate and average maximum decision making performance outcomes. This hypothesis speaks to the 'value added' of using SJTs as assessment tools. In order to test this hypothesis, two multiple regression analyses were conducted, one for each of the dependent variables of concern. The results of the first of these two analyses suggested knowledge as measured by the 'Should Do' SJT ($\beta = .510$, $p \leq .05$) accounted for 25% incremental variance in the aggregate maximum performance outcome variable beyond that already explained by declarative knowledge ($F\Delta_{(2,43)} = 7.207$, $p \leq .05$). Similarly, knowledge as measured by the 'Should Do' SJT ($\beta = .342$, $p \leq .05$) accounted for 11% incremental variance in the average maximum performance outcome variable beyond that already explained by declarative knowledge ($F\Delta_{(2,46)} = 2.931$, $p \leq .05$). This evidence supports hypothesis 4.

Test Fairness

The sixth hypothesis tested in this study was investigated on an exploratory basis to help determine whether the SJT used in the present study was fair for various gender (i.e., female vs. male) and ethnic-based (i.e., majority vs. minority) subgroups according to Cleary's (1968) definition of test fairness. The sixth hypothesis stated that SJT scores would interact with gender and ethnicity when predicting maximum decision making performance outcomes such that the slopes would be different for the focal and reference groups studied.

To test the above assertion, the aggregate maximum performance outcome variable was regressed on SJT scores ($\beta = .115$, $p \geq .05$), gender ($\beta = .075$, $p \geq .05$), and the interaction term ($\beta = -.095$, $p \geq .05$). The overall F test ($F_{(3,96)} = .479$, $p \geq .05$) and the specific t-tests for the moderator and interaction terms were not significant. When the average maximum performance outcome variable was regressed on SJT scores ($\beta = .054$, $p \geq .05$), gender ($\beta = .163$, $p \geq .05$), and the interaction term ($\beta = .069$, $p \geq .05$), the overall F test ($F_{(3,98)} = 1.629$, $p \geq .05$) and specific t-

tests for the beta weights were not significant. Thus, the results of these two MMR analyses suggested that gender did not moderate the relationship between SJT scores and either aggregate or average maximum performance outcomes.

The second part of hypotheses 5 was concerned with whether ethnicity moderated the relationships between SJT scores and peak performance criteria. To test this assertion, the aggregate maximum performance outcome variable was regressed on SJT scores (β = -.073, p ≥ .05), ethnicity (β = .119, p ≥ .05), and the interaction term (β = .148, p ≥ .05). The overall F test ($F_{(3,96)}$ = .825, p ≥ .05) and the specific t-tests for beta weights were not significant. When the average maximum performance outcome variable was regressed on SJT scores (β = .062, p ≥ .05), ethnicity (β = .086, p ≥ .05), and the interaction term (β = .062, p ≥ .05), the overall F test ($F_{(3,98)}$ = .834, p ≥ .05) and specific t-tests for the beta weights were still not significant. Thus, the results of these two MMR analyses suggested that ethnicity did not moderate the relationship between SJT scores and either aggregate or average maximum performance outcomes. This evidence, and that presented above, does not support hypothesis 5.

CHAPTER SIX: DISCUSSION

The purpose of this study was to generate construct validity evidence for an SJT in a maximum performance context in order to extend prior mixed research results and provide actionable guidance for the development and use of SJTs. A PRPR model (Ployhart, 2006) and criterion theory (Campbell, 1990) were leveraged to craft research questions and specify five hypotheses about the validity of an SJT. The PRPR model suggests there are multiple individual, methodological, and contextual factors which impinge upon the latent cognitive response processes that respondents engage in each time they respond to an SJT item. It is important to understand, model, and control these factors when examining the validity of multidimensional assessment tools like SJTs because they can lead to systematic measurement error that provides an alternative explanation for the relationships observed in validity studies.

The PRPR model suggests a test score is the outcome of a psychological process, a point ignored by most validation research that focuses strictly on scale scores and the relationships between them. The present study did not directly measure or manipulate the latent response processes that give rise to a response to an SJT item, but it did consider one of the myriad of factors that impinge upon these processes, response instructions. According to the PRPR model, response instructions contribute a contaminating source of variance to SJT scores because when response instructions change, so to do the primary determinants of the response processes respondents engage in when answering an SJT item (see Appendix C). Because the determinants of the cognitive processes test takers engage in when formulating responses to SJT items change when response instructions change, instructions can ultimately affect the psychometric properties of, and construct validity evidence for, SJTs (Ployhart & Ehrhart, 2003). Thus, the PRPR model provided a framework for collecting and interpreting construct validity evidence for the new SJT.

Given the tenets of the PRPR model, it was expected that part of the total variance in SJT scores would be 'true score' variance, attributable to individual differences in the constructs that the SJT was designed to measure (i.e., declarative and strategic knowledge). In addition, response instructions were expected to contribute a contaminating source of systematic error variance to participant SJT scores via their influence on the proximal determinants of the response processes that respondents engage in when answering SJT items. According to the PRPR model, 'Should Do' response instructions trigger response processes that are primarily driven by respondent knowledge of what ought to be done in response to a posed scenario, whereas 'Would Do' instructions primarily trigger personality driven response processes (see Appendix C). This line of thinking suggests systematic variance in 'Would Do' SJT scores may reflect both the constructs the SJT was designed to measure (e.g., declarative and strategic knowledge) and other constructs it was not meant to measure (e.g., personality characteristics).

Based on the tenets of the PRPR model, differences were expected in the psychometric properties of, and validity evidence for, the 'Should Do' and 'Would Do' SJTs. The findings of the present study largely confirmed these expectations. For instance, noticeable differences existed between the estimates of internal consistency reliability for the 'Should Do' (.546) and 'Would Do' (.348) SJTs. These estimates are low as compared to conventional standards; likely because both tests were comprised of only six items, but their difference in magnitude is striking given that both SJT versions were comprised of identical item content and response alternatives. An explanation for this difference in reliability is provided by the PRPR model. The PRPR model suggests that 'Would Do' instructions invoke personality driven response processes which culminate in 'Would Do' SJT scores that reflect the declarative and strategic knowledge the SJT was designed to measure, as well as construct irrelevant personality characteristics. If systematic

variance due to personality was reflected in 'Would Do' SJT scores, then this variance would contribute to the heterogeneity of the construct domain measured by this test and thus it would have a lower estimate of internal consistency reliability than the 'Should Do' SJT which measured a relatively more homogenous construct domain.

In addition to differences in the psychometric properties of the SJTs, the PRPR model suggested that differences in the validity evidence for two SJT versions would exist because of their response instructions. In regards to this issue, the findings of the present study suggested there were differences in the criterion-related validity of the 'Should Do' and 'Would Do' SJTs. The results of MMR analyses suggested that response instructions moderated the relationships between SJT scores and both aggregate and average maximum decision making performance outcomes. Specifically, SJT scores interacted with SJT response instructions such that the 'Should Do' SJT version was more predictive of both peak performance criteria. It seems the systematic error variance attributable to personality in 'Would Do' SJT scores may have ultimately served to attenuate the criterion-related validity of this test with the peak performance criteria included in the study. This is likely because effectiveness in maximum performance contexts is ability rather than volitionally driven.

Of note, however, the evidence generated by the present study in support of the core tenets of the PRPR model is not entirely unequivocal. Specifically, the results of MMR analyses suggested the relationships between declarative knowledge and SJT scores, and risk taking and SJT scores, were not contingent upon SJT response instructions. Moreover, contrary to expectations fueled by the PRPR model, both the 'Should Do' and 'Would Do' SJT versions used in the present study were fair for gender (i.e., female vs. male) and ethnic-based (i.e., majority vs. minority) subgroups according to Cleary's (1968) definition of test fairness.

In sum, the findings of the present study suggested response instructions moderated the criterion-related validity evidence for an SJT that was designed to measure declarative and strategic knowledge. These findings bolster prior research findings that suggested response instructions affected the nomological network of SJTs. The findings generated by the present study also complement prior findings from research examining SJT response instructions in typical performance contexts. Collectively, the findings from this line of research underscore the fact that response instructions are not arbitrary; they must be valid for the manners in which they are applied (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Response instructions must be aligned with the purposes of measurement, the constructs measured, and the context of measurement. Ignoring this caveat can produce a response instruction method effect or systematic measurement error that provides an alternative explanation for the relationships observed in validity studies.

In addition to extending prior empirical findings on response instructions, and generating evidence in support the core tenets of the PRPR model, the findings of the present study also hold implications for the use of SJTs in talent selection contexts. Specifically, the present study investigated both the upper-bound criterion-related validity coefficient associated with the use of SJTs in talent selection contexts, and the incremental variance in performance outcomes that can be explained by SJTs beyond that already accounted for by declarative knowledge as measured by a traditional multiple choice test. This kind of information is important to stakeholders who use SJTs in human capital initiatives in the workplace. For example, the upper-bound validity coefficient for the use of SJTs is important because it is a primary factor in determining the practical economic value or utility of SJTs as measurement methods. Moreover, the incremental variance in criteria accounted for SJTs speaks to their ‘value added’ as assessment tools.

In regards to economic considerations, meta-analytic research findings suggested SJTs can have utility as assessment tools (see McDaniel et al., 2003). Of note, however, the primary studies included in prior meta-analyses were largely subject to the problems inherent to using subjective ratings of typical performance as the primary operationalizations of the dependent variables investigated. As noted in the conceptual foundation section, the use of subjective ratings of typical performance criteria can result in lower validity coefficients for assessment tools. If the findings of prior studies have underestimated the validity of SJTs, then utility estimates based on these findings will also be downwardly biased; as the utility of an assessment tool is directly proportional to its predictive validity coefficient (Schmidt et al., 1979).

One byproduct of investigating the criterion-related validity of an SJT when its response instructions are manipulated, and maximum decision-making performance outcomes are objectively measured, is that a closer approximation of the upper-bound validity coefficient associated with the use of SJTs in talent selection contexts was established. The validity coefficient provided by the present study is a more accurate estimate of the upper-bound of the validity coefficients associated with the use of SJTs in selection contexts because "…'should do' responses would better predict tightly controlled simulations that represent 'can do' behavior of the same skill than they would predict other aspects of job performance or less tightly controlled on-the-job measures…" (Fritzsche et al., 2006, p. 22). Moreover, the estimate provided by the present study was based on predictor and criterion constructs that were purposively sampled from theories of performance which helped ensure the fidelity between these constructs. The findings of the present study suggested the uncorrected upper-bound criterion-related validity coefficient associated with SJTs in talent selection contexts is at least moderate to strong (β = .478). Thus, it seems, SJTs can have substantial utility as assessment tools.

The present study also examined the incremental variance in peak performance outcomes that was explained by an SJT designed to measure declarative and strategic knowledge beyond that already accounted for by declarative knowledge as measured by a traditional multiple choice test. This is an important issue to stakeholders who use SJTs in selection systems comprised of multiple assessment tools, because one of the oft noted advantages of SJTs is that they can explain incremental variance in occupational criteria over cognitive ability tests and personality measures. The present study also extended prior research by examining the incremental variance accounted for by an SJT beyond that explained by a traditional multiple choice test of declarative knowledge. This kind of information is important to SJT users because given their complexity, SJTs take longer and cost more to develop than traditional multiple-choice tests of knowledge.

The findings of the present study suggested declarative and strategic knowledge, as measured by the 'Should Do' SJT version, accounted for 25% incremental variance in the aggregate maximum decision making performance outcome variable beyond that already explained by declarative knowledge as measured by the traditional multiple choice test. Moreover, declarative and strategic knowledge, as measured by the 'Should Do' SJT version, accounted for 11% incremental variance in the average maximum decision making performance outcome variable beyond that already explained by declarative knowledge as measured by the traditional multiple choice test. Collectively, these findings suggest SJTs that measure declarative and strategic knowledge can explain a sizable amount of incremental variance in organizationally valued criteria beyond that already explained by less expensive and readily available methods of measuring knowledge such as traditional multiple choice tests.

Limitations

The present study leveraged the PRPR model to examine the construct validity of an SJT in a maximum performance context. The PRPR model was used to craft hypotheses about the aggregate relationships between SJT scores and several independent and dependent variables. The PRPR model suggests that manipulations of SJT response instructions affect the proximal determinants of the response processes respondents engage in when responding to the items of assessment tools. Of note, however, the present study did not include any direct manipulations or measures of the proximal determinants of the response processes respondents engage in when answering items, or the response processes themselves. Thus, the findings from the present study only indirectly support the assertions of Ployhart's (2006) PRPR model. Additional research is needed to directly test the assertions of the PRPR model and, thereby generate substantive validity evidence for an SJT, by directly examining the response processes of SJT respondents.

Another limitation of the present study is that it only included one kind of criterion, job-specific task proficiency. As noted by Campbell (1990), job-specific task proficiency is just one of many dimensions underlying job performance and it may not be the most important factor to effective performance. Thus, research is needed to extend the findings of the present study by examining the relationships between SJTs that are also designed to measure declarative and strategic and other specific facets and/or types of job performance. For example, research could address whether response instructions moderate the relationships between SJTs designed to measure specific constructs and contextual, team, and/or adaptive performance criteria. The results of this line of research would complement the current initiative which examined maximum performance criteria and prior research which examined typical performance criteria.

CHAPTER SEVEN: CONCLUSION

An increasingly informed and litigious society has helped create an impetus for organizations to use professionally developed assessment tools in lieu of either outmoded or outlawed approaches when gathering information for their human capital initiatives. Although the options available to gather data are more limited and scrutinized, identifying exceptional performers is no less important, as these workers can accomplish up to twice as much as poor performers (Schmidt & Hunter, 1981). SJTs offer one means of balancing these concerns.

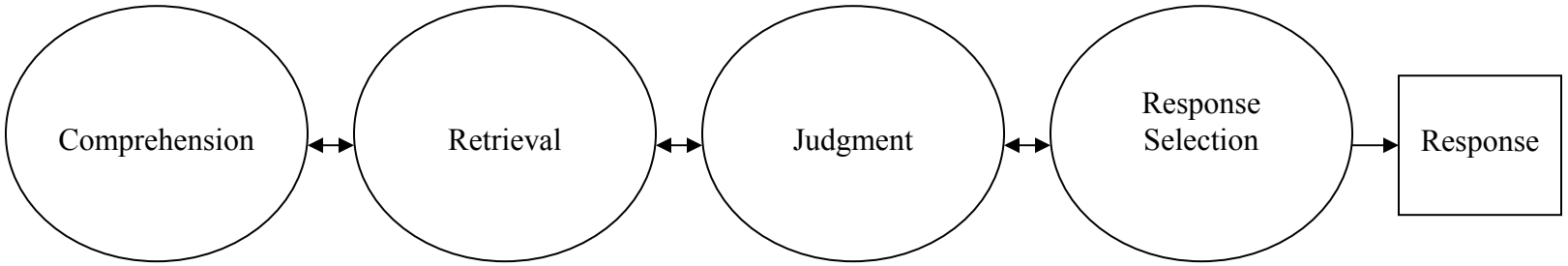
SJTs have proven useful for a variety of purposes and for a variety of reasons over the course of the 130 years they have been included in human capital initiatives; yet, some fundamental questions about their nature and appropriate uses are just beginning to garner serious attention. Questions about the constructs SJTs measure, why they measure them, and how these constructs fit within in a nomological network of lawful relations, are being addressed by a new wave of individual differences studies, experiments, and meta-analytic initiatives. The findings from the studies have provided much needed insight about SJTs. The lessons learned from this line of research can be leveraged by users to better develop, structure, and score SJTs and thereby increase the construct validity evidence for, and utility of, these assessment tools.

The present study continued in this tradition by generating validity evidence for a newly developed SJT in a maximum performance context via the use of a between-subjects design. Validity evidence was generated for the SJT by leveraging the PRPR model to examine the constructs it purported to measure, the relationships between those constructs and other independent and dependent variables, and the role of its response instructions in moderating those relationships. In regards to this latter issue, a series of MMR analyses were conducted to determine if SJT response instructions moderated the relationships between: (1) declarative

knowledge and SJT scores, (2) risk taking and SJT scores, (3) SJT scores and an aggregate performance outcome, and (4) SJT scores and an average performance outcome. This approach allowed for the collection of validity evidence to support the new SJT, while concurrently providing a means to extend prior mixed research results, gather evidence to support the PRPR model, and generate actionable guidance for the development and use of SJTs.

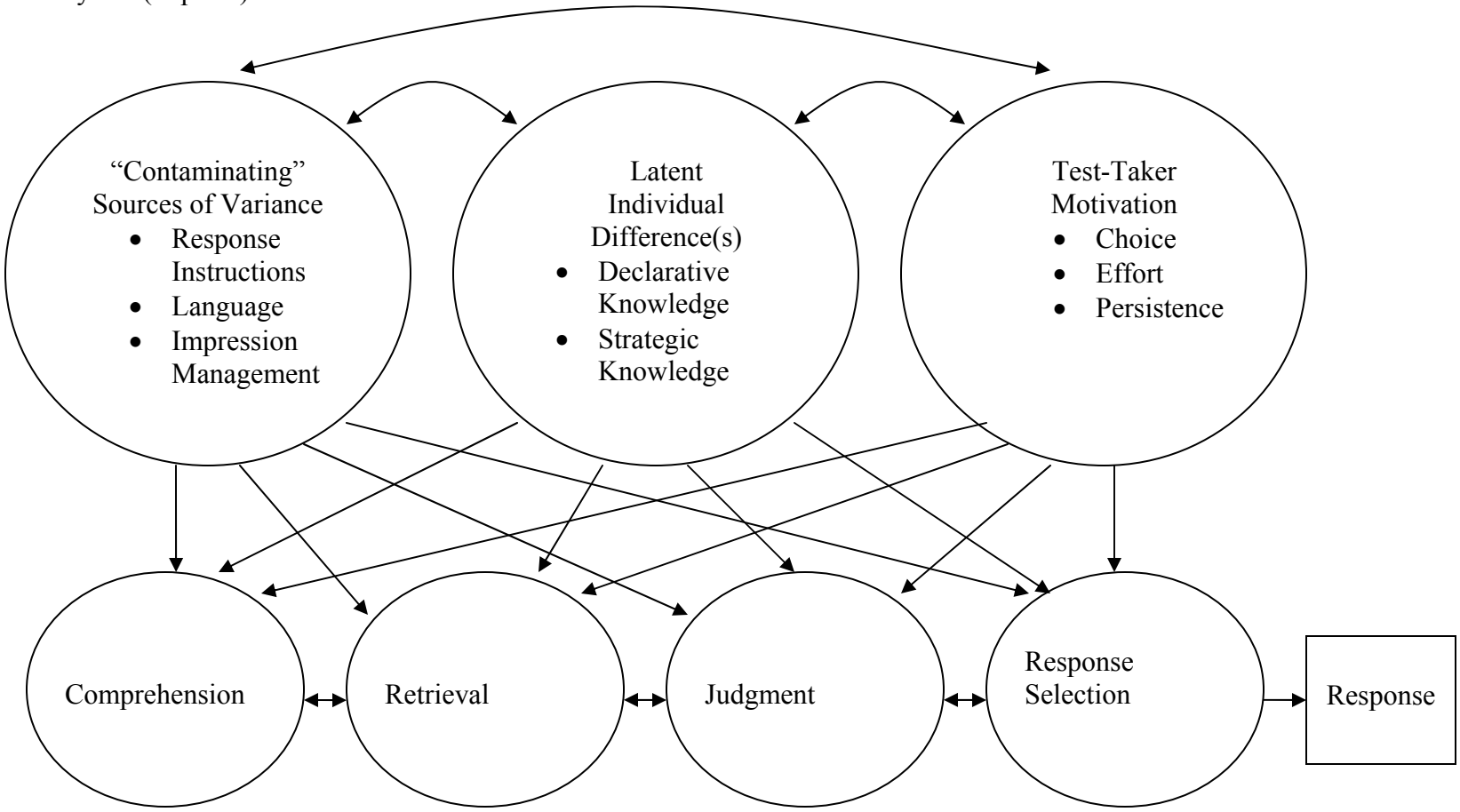
The findings of the present study suggested response instructions played an important role in shaping the nomological network of the SJT examined. Specifically, SJT response instructions moderated the relationships between SJT scores and both aggregate and average maximum decision making performance outcomes. The findings of the present study also suggested that SJTs can account for large amounts of incremental variance in peak performance criteria and can provide a substantial dollar benefit or utility in talent selection contexts.

APPENDIX A: PREDICTOR RESPONSE PROCESS MODEL


Comprehension
Retrieval
Judgment
Response Selection
Response

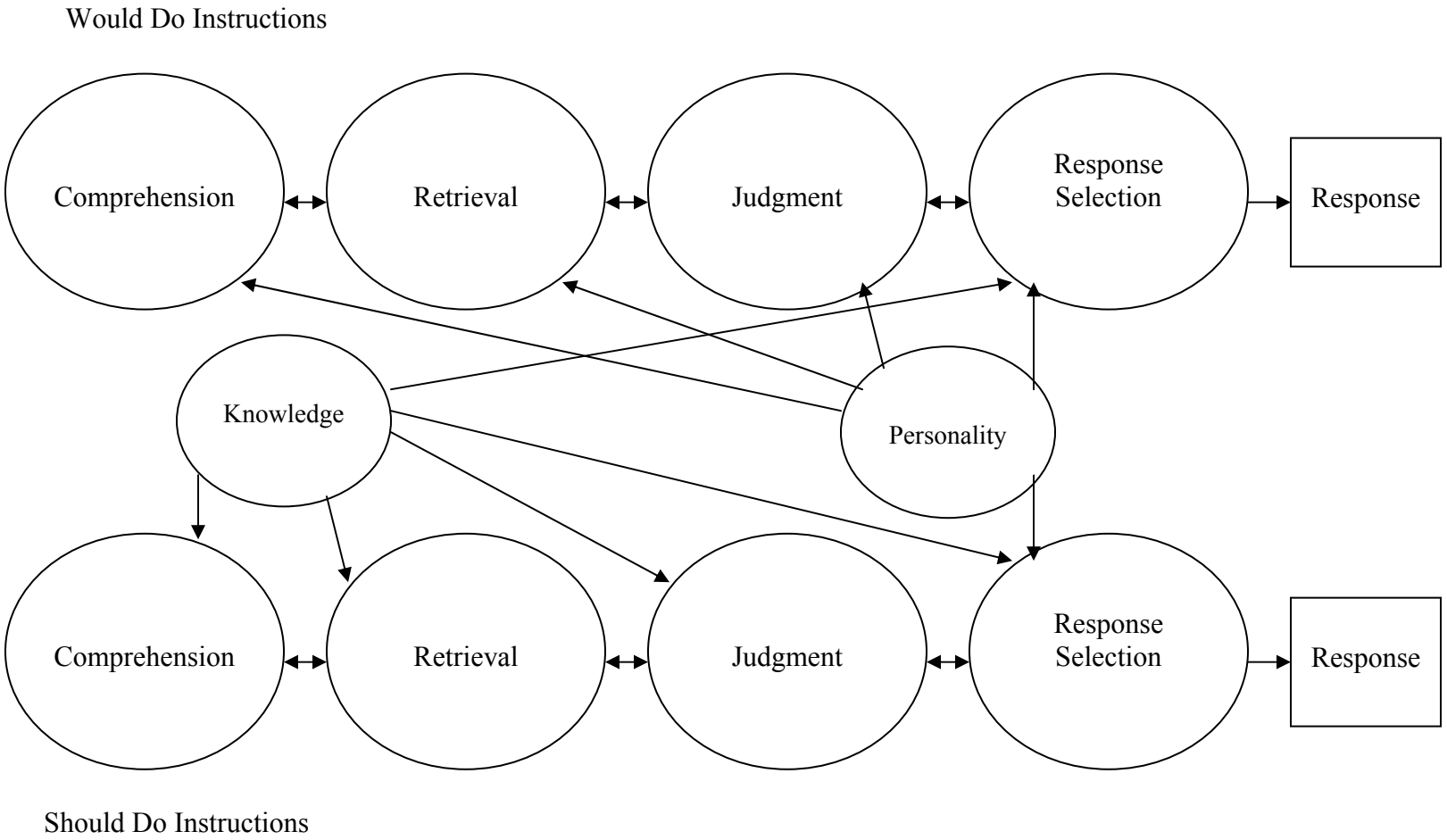
APPENDIX B: DETERMINANTS OF PREDICTOR RESPONSE PROCESSES

Figure 2
Determinants of Predictor Response Processes
Adapted from Ployhart (in press)



APPENDIX C: RESPONSE PROCESSES INVOLVED WITH WOULD DO AND SHOULD DO INSTRUCTIONS

Figure 3
Response Processes Involved with Would Do and Should Do Instructions
Adapted from Ployhart (in press)



APPENDIX D: TABLE 1 SCALE DESCRIPTIVES & INTERCORRELATIONS

Table 1

Scale Descriptives & Intercorrelations

Test/Measure	M	SD	α	1	2	3	4	5	6
1. Traditional Declarative Knowledge Test	6.48	.786	.494	-	-.060	.181	.160	.132	.085
2. Risk Taking Measure	41.04	8.49	.776		-	**-.254**	**-.369**	.026	-.027
3. 'Should Do' SJT	12.90	2.44	.546			-		**.478**	**.346**
4. 'Would Do' SJT	12.15	2.09	.348				-	-.158	-.048
5. Aggregate Maximum Performance Outcome	4898	521						-	**1**
6. Average Maximum Performance Outcome	4968	302							-

Note. Correlations that are statistically significant at $p < .05$ one tailed are bolded.

LIST OF REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211.

Ajzen, I. & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84*, 888-918.

Alignmark (2001). AccuVision customer service system validation report. Maitland, FL: Alignmark.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anderson, N. (2001). Applicant and recruiter reactions to new technology in selection: A critical review and agenda for future research. *International Journal of Selection and Assessment, 11*, 121-136.

Ansbacher, H. L. (1941). German military psychology. *Psychological Bulletin, 38*, 370-392.

Bell, B. S. & Kozlowski, S. W. J. (2002). Adaptive guidance: Enhancing self-regulation, knowledge, and performance in technology-based training. *Personnel Psychology, 55*, 267-307.

Binet, A. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'Annee Psychologique, 12*, 191-244.

Binning, J. F. & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494.

Bruce, B. M. & Learner, D. B. (1958). A supervisory practices test. *Personnel Psychology, 11*, 207-216.

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

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35-70). San Francisco: Jossey-Bass.

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50*, 655-702.

Cannon-Bowers, J. A., Tannenbaum, S.I., Salas, E., & Volpe, C.E. (1995). Defining competencies and establishing team training requirements. In R. Guzzo, & E. Salas (Eds.), *Team effectiveness and decision-making in organizations*. San Francisco, CA: Jossey-Bass.

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

Chan, D. & Schmitt, N. (2002). Video-based versus paper-and-pencil method of assessment in situational judgment tests. *Journal of Applied Psychology, 82*, 143-159.

Clevenger, J. (1999, April). *The construct validity of the situational judgment inventory*. Symposium conducted at the 14th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Clevenger, J., Jockin, T., Morris, S., & Anselmi, T. (1999, April). *A situational judgment test for engineers: Construct and criterion-related validity of a less adverse alternative*. Paper presented at the 14th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

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

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cook, T. D. & Campbell, D. T. (1979). (Eds.), *Quasi-Experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin Company.

Costa, P. T. & McCrae, R. R. (1985). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd Ed.). New York: Harper & Row.

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

DuBois, C. L. Z., 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*, 205-211.

File, Q. W. (1945). The measurement of supervisory practices test. *Personnel Psychology, 29*, 323-337.

Fishbein, M. & Ajzen, I. (1977). *Belief, attitude, intentions, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Fritzsche, B. A., Stagl, K. C., Salas, E. & Burke, C. S. (2006). Enhancing the design, delivery and evaluation of scenario-based training: Can situational judgment tests contribute? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests*. Mahwah, NJ: Lawrence Erlbaum Associates.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Gottfredson, L. S. (1988). Reconsidering fairness: A matter of social and ethical priorities. *Journal of Vocational Behavior, 33*, 293-319.

Guion, R. M. (1991). Personnel assessment, selection, and placement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp.687-732). Palo Alto, CA: Consulting Psychologists Press.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum & Associates.

Hanson, M. A. & Borman, W. C. (1995, April). *Construct validation of a measure of supervisory job knowledge*. Poster presented at the 10th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Hanson, M. A. & Borman, W. C. (1989). *Development of a situational judgment test to be used as a job performance measure for first line supervisors in the U.S. Army*. Paper presented at the 4th annual conference of the Society for Industrial and Organizational Psychology, Boston, MA.

Hanson, M. A., Horgen, K. E., & Borman, W. C. (1998, March). *Situational judgment: An alternative approach to selection test development*. Paper presented at the 38th annual conference of the International Military Testing Association.

Hanson, M. A. & Ramos, R. A. (1996). Situational judgment tests. In R. S. Barrett (Ed.), *Fair employment strategies in human resource management* (pp. 119-124). Westport, CT: Greenwood Publishing Group.

Harvey, J. L., Morath, R., Christopher, S. C., & Anderson, L. (2001). *Validity of the selection instrument developed for the senior accountant job class*. Fairfax, VA: Caliber Associates.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639-683.

Hedge, J. W., Borman, W. C. & Hanson, M. A. (1996, March). *Videotaped crew resource management scenarios for selection and training applications*. Paper presented at the 38th annual conference of the International Military Testing Association.

Hedlund, K., Plamondon, K., Wilt, J., Nebel, K., Ashford, S., & Sternberg, R. J. (2001, April). Practical intelligence for business: Going beyond the GMAT. In J. Cortina (Chair), *Out with the old, in with the new: Looking above and beyond what we know about cognitive ability predictors*. Symposium conducted at the 16th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Horgen, K. E. (2004). *Construct and criterion-related validity of two situational judgment tests*. Unpublished Ph.D. dissertation. University of South Florida, Tampa, Florida.

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

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

Hunter, D. R., Martinussen, M., & Wiggins, M. (2003). Understanding how pilots make weather-related decisions. *The International Journal of Aviation Psychology, 13,* 73-87.

Joiner, D. A. (2002). Assessment centers: What's new? *Public Personnel Management, 31,* 179-185.

Kane, J. S. (1982, November). *Rethinking the problem of measuring performance: Some new conclusions and a new appraisal method to fit them.* Paper presented at the 4th Johns Hopkins University National Symposium on Educational Research, Washington, DC.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of ASTD, 13*, 3-9.

Kite, E. S. (1916). *The development of intelligence in children*. Vineland, NJ: Publications of the Training School at Vineland.

Kozlowski, S. W. J. & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. San Francisco, CA: Jossey-Bass.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 3*, 442-452.

Lim, B. & Ployhart, R. E. (2004). Transformational leadership: Relations to the five-factor model and team performance in typical and maximum performance contexts. *Journal of Applied Psychology, 4*, 610-621.

McDaniel, M. A., Hartman, N. S., & Grubb, W. L. (2003, April). *Situational judgment tests, knowledge, behavioral tendency, and validity: A meta-analysis*. Paper presented at the 18th annual conference of the Society for Industrial and Organizational Psychology. Orlando, FL.

McDaniel, M. A., Hartman, N. S., Nguyen, N., & Grubb, W. L. (2006). Situational judgment tests: Validity and an integrative model. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests*. Mahwah, NJ: Lawrence Erlbaum & Associates.

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

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

McDaniel, M. A., Yost, A. P., Ludwick, M. H., Hense, R. L., & Hartman, N. S. (2004, April). *Incremental validity of a situational judgment test*. Paper presented at the 19th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

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

Moss, F. A. (1926). Do you know how to get along with people? Why some people get ahead in the world while others do not. *Scientific American, 135*, 26-27.

Motowidlo, S. J. (1999). Asking about past behavior versus hypothetical behavior. In R. W. Eder and M. M. Harris (Eds.), *Employment interview handbook*, pp. 179-190. Thousand Oaks, CA: Sage.

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

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

Motowidlo, S. J., Hanson, M. A. & Crafts, J. L. (1997). Low-fidelity simulations. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology*. Palo Alto, CA: Consulting Psychologists Press.

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

Mullins, M. E. (2000). *The effects of practice variability and velocity feedback on the development of basic and strategic training skills*. Unpublished Ph.D. dissertation. Michigan State University, Ann Arbor, Michigan.

Nguyen, N. T., McDaniel, M. A., & Biderman, M. D. (2002, April). *Response instructions in situational judgment tests: Effects on faking and construct validity*. Symposium conducted at the 17th annual conference of the Society for Industrial and Organizational Psychology. Atlanta, GA.

Nunnally, J. & Bernstein, I. (1994). *Psychometric theory*. New York: McGraw Hill.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment test as predictors of college student performance. *Journal of Applied Psychology, 89*, 187-207.

Peters, H. & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational & Psychological Measurement, 65*, 70-89.

Phillips, J. E. (1993). Predicting negotiation skills. *Journal of Business and Psychology, 7*, 403-411.

Ployhart, R. E. (2006). The predictor response process model. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests*. Mahwah, NJ: Lawrence Erlbaum & Associates.

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

Ployhart, R. E., Porr, W., & Ryan, A. (2004, April). A construct oriented approach for developing situational judgment tests in a service context. In P. R. Sackett (Chair), *New developments in SJTs: Scoring, coaching, and incremental validity*. Presented at the 19th annual conference of the Society for Industrial and Organizational Psychology, Chicago.

Ployhart, R. E. & Ryan, A. M. (2000, April). *Integrating personality tests with situational judgment tests for the prediction of customer service performance*. Paper presented at the 15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology, 56*, 733-752.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879-903.

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. D., Schmitt, N., & Chan, D. (1996). Models of job performance ratings: An examination of rate race, rate gender, and rater level effects. *Human Performance, 9*, 103-121.

Reynolds, D. H., Winter, J. L., & Scott, D. R. (1999, April). Development, validation, and translation of a professional level situational judgment inventory. In J. P. Clevenger (Chair), *The construct validity of the situational judgment inventory*. Symposium presented at the 14th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Reynolds, D. H., Sydell, E. J., Scott, D. R., & Winter, J. L. (2000, April). *Factors affecting situational judgment test characteristics*. Paper presented at the 15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Rosen, N. A. (1961). How supervise? *Personnel Psychology, 14*, 87-99.

Sacco, J. M., Scheu, C. R., Ryan, A. M., Schmitt, N., Schmidt, D. B., & Rogg, K. L. (2000, April). *Reading level and verbal test scores as predictions of subgroup differences and validities of situational judgment tests*. Paper presented at the 15th annual conference of the Society for Industrial and Organizational Psychology. New Orleans, LA.

Sackett, P. R. & Larson, J. R. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 419-490). Palo Alto, CA: Consulting Psychologists Press.

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

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

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job sample vs. paper and pencil trades and technical tests: Adverse impact and examinee attitudes. *Personnel Psychology, 30*, 187-197.

Schmidt, F. L. & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36*, 1128-1137.

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

Schmidt, F. L. & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199-223.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). The impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64*, 609-626.

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

Schmitt, N. & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests*. Mahwah, NJ: Lawrence Erlbaum & Associates.

Schmitt, N. & Mills, A. E. (2001). Traditional tests and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology, 3,* 451-458.

Shotland, A., Alliger, G. M., & Sales, T. (1998). Face validity in the context of personnel selection: A multimedia approach. *International Journal of Selection and Assessment, 6,* 124-130.

Simoneit, M. (1938). *Principles of the psychological study of the officer-recruit relationship in the army*. Berlin: Bernard & Graefe.

Smith, P. C. (1976). Behavior, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (ed.), *Handbook of industrial and organizational psychology* (pp. 745-775). Chicago, IL: Rand-McNally.

Smith-Jentsch, K. A., Salas, E., & Brannick, M. T. (2001). To transfer or not to transfer? Investigating the combined effects of trainee characteristics, team leader support, and team climate. *Journal of Applied Psychology*, *86*, 279-292.

Society for Industrial and Organizational Psychology. (2004). *Principles for the validation and use of personnel selection procedures*. Bowling Green, OH: Society for Industrial and Organizational Psychology.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., Snook, S. A. & Grigorenko, E. L. (2000). Practical intelligence: An example from the military. *Practical intelligence in everyday life*. Cambridge, UK: Cambridge Press.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*, 334-344.

Stone, E. F., Austin, J. T., & Shetzer, L. (1986). Moderated regression versus subgrouping strategies for detecting moderating effects. Paper presented at the meeting of the American Psychological Association.

Strong, M. H. & Najor, M. J. (1999, April). *Situational judgment versus cognitive ability tests: Adverse impact and validity*. Paper presented at the 14th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Tabachnick, B. G. & Fidell, L. S. (2001). Using multivariate statistics. (4th edition). New York: Allyn & Bacon.

Thurstone, L. L. (1947). *Multiple-factor analysis: A development and expansion of the vectors of mind.* Chicago, IL: University of Chicago Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Truxillo, D. M. & Hunthausen, J. M. (1999). Reactions of African-American and White applicants to written and video-based police selection tests. *Journal of Social Behavior & Personality, 14*, 101-112.

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

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

Weekley, J. A. & Ployhart, R. E. (2002, April). *Situational judgment and training experience: Antecedents and correlates*. Symposium presented at the 17th annual conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario.

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

Weekley, J. A. & Ployhart, R.E. (2006 a). (Eds.), *Situational judgment tests*. Mahwah, NJ: Lawrence Erlbaum & Associates.

Weekley, J. A. & Ployhart, R. E. (2006 b). Introduction and history of SJTs. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests*. Mahwah, NJ: Lawrence Erlbaum & Associates.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests*. Mahwah, NJ: Lawrence Erlbaum & Associates.

Wernimont, P. & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372-376.