Video-Based Situational Judgment Test Characteristics: Multidimensionality at the Item Level and Impact of Situational Variables

Carl J. Swander

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Psychology

Robert J. Harvey, Chair
Danny K. Axsom
Kevin D. Carlson
John J. Donovan
Morrell E. Mullins

May 3, 2001
Blacksburg, Virginia

Keywords: Video-Based Situational Judgment Tests, SJT, VBSJT, Situational Characteristics

Video-Based Situational Judgment Test Characteristics: Multidimensionality at the Item Level and Impact of Situational Variables

Carl J. Swander

(ABSTRACT)

A new approach was taken to identify a specific construct or dimension being measured by a video-based situational judgment test (VBSJT). Appropriate exertion of control was specifically explored in relation to a VBSJT test designed for entry-level selection of law enforcement officers. Ratings from ten law enforcement experts were utilized to identify this construct. The VBSJT items scored toward overexertion of control were significantly related to performance ($\underline{r}$ = .23) in a sample of 334 incumbent police officers, capturing a large portion of the effective variance of the test which had an overall validity of $\underline{r}$ = .34.

Situational variables within the items were then compared to ratings of exertion of control within a sample of 5426 applicants. General provocation toward overexertion of control and ethnicity significantly affected appropriate exertion of control. Gender and likeability also had significant impact on appropriate exertion, but the practical significance was limited. Specific character manipulations (i.e., rudeness, aggressiveness, pleasantness, cooperativeness, sympathy, and suspiciousness) also had a significant impact on appropriate exertion of control. Specific information manipulations (i.e., warrants, complaints, contemptible crimes and laws being broken) also had an impact on appropriate exertion of control. Some unexpected findings suggest that the character manipulations may actually override the effect of other provocation.

The overexertion of control scale was also applied to test hypotheses about the likely behavior of police officers. It was found that the location of the organization had an affect on overexertion of control. Contrary to the hypothesis, suburban locations had more overexertion of control than urban locations. Length of tenure for police officers did not have an effect on overexertion of control. This difference did not affect validity across organizations. Implications and further research are discussed.

ACKNOWLEDGMENTS

I would like to thank my thesis committee John Donovan, Kevin Carlson, Danny Axsom and Morrie Mullins for taking the time to help with this project. The committee took a real interest in helping me to conduct the best research possible. Also, thanks to my committee chairman, Robert J. Harvey, who served as my advisor throughout graduate school. Thanks also to RJ and Dr. Finney for looking beyond the rules and understand my individual circumstances.

I would also like to thank Cheryl Indehar. Not only did she efficiently collect the crucial portion of my data but she was a continual supporter. Without Cheryl this project could have easily taken another year. I would also like to thank all of the officers that took time out of their busy schedules to voluntarily complete two hours of ratings. I would also like to thank Gayle Kennedy. She is continually providing support and needed resources to all I/O graduate students.

I also want to say thanks to John Donovan for being such a good friend. It was nice to have someone around with so many common interests, which helped to make Blacksburg a bearable place to live. John also instilled a sense of calmness about the graduate school process that brought about confidence in my ability. Without this confidence I would not have been able to finish in four years.

I would also like to thank my entire family. My parents have been an incredible help. Not only have they helped me to get where I am but they continue to help me get where I want to be. Their suggestions and contributions throughout graduate school have been invaluable. Most of all, the level of support and love from my parents and grandparents has made me the successful person I am today. I would also like to thank my wife, Megan, for her support throughout the entire process. Thanks also for putting up with my extreme work habits during the crunch times.

TABLE OF CONTENTS

TABLES AND FIGURES

INTRODUCTION

Situational judgment tests (SJTs) have become popular predictors of job performance because they typically have substantial validities and smaller adverse impact than many other selection tests (e.g., Motowidlo, Dunnette, & Carter, 1990). However, unlike other selection measures, SJTs are not well understood. It is not clear what constructs are measured with a SJT or how these constructs are measured. Recent research has attempted to construct validate SJTs but the results obtained have been fairly ambiguous (e.g., Smith & McDaniel, 1998). This has led to poor understanding of how SJTs effectively measure job performance.

The key to understanding SJTs is to appropriately account for their multidimensionality. SJTs are multidimensional at both the item and test level (e.g., Jones, Dwight, and Nouryan, 1999). Multidimensionality suggests that multiple constructs related to job performance are measured with SJTs. However, variables such as personality and cognitive ability, which are commonly used to predict job performance, have only sporadically correlated with SJTs. Furthermore, commonly used categorization techniques, such as factor analysis, typically yield no evidence of specific factors. Therefore, multidimensionality may also be at the item level. Multidimensionality at the item level suggests that the personality or other constructs which influence one wrong answer choice on an item may be different from the construct affecting another response choice.

Thus, it is necessary to approach understanding of SJTs from a nontraditional perspective. This perspective requires consideration of multidimensionality at the item level and the fact that SJTs are usually designed to assess performance on complex jobs, which may involve many different constructs. SJTs are designed for specific jobs, which suggests that the underlying constructs of a SJT will differ depending on the job for which it was designed.

One purpose of the current study is to demonstrate that item responses in a video-based SJT (VBSJT) designed for law enforcement can be scored to identify a particular construct which was a major focus of the VBSJT by taking into account item multidimensionality. Based on meetings with over 100 subject matter experts, appropriate exertion of control on the part of police officers is a construct that is of great concern to law enforcement managers and was an important consideration in the development of the VBSJT.

This is not a construct validation study in the typical sense but a method for identifying a specific construct within a multidimensional VBSJT. The method requires utilizing

homogeneous characteristics of response options to identify appropriate exertion of control within this VBSJT. Therefore, each item and answers to the item will be rated in relation to appropriate exertion of control. This should provide a useful method for identifying a unidimensional construct that is an important component of this multidimensional VBSJT.

However, the primary goal of this paper is to add to the understanding of how SJTs measure job performance. Addressing the issue of understanding would not be complete with the simple identification of a unidimensional construct. This VBSJT measures appropriate exertion of control in an environment similar to that faced on the job. Although SJTs are regarded as "low fidelity" simulations, no research has actually evaluated the assumption that situational characteristics within the test actually have an impact on examinee responses. Therefore, VBSJT items will be explored to determine the effectiveness of certain situational characteristics in terms of eliciting over- or underexertion of control responses and thereby contributing to the effective measurement of this construct.

General provocation toward under- or overexertion of control will be explored for impact on the construct of appropriate exertion of control. Furthermore, individual situational characteristics that should contribute to general provocation (e.g., rude characters, pleasant characters, warrants being enforced) will be explored. Focal character ethnicity, gender, and likeability will also be explored as situational characteristics that affect applicants' exertion of control responses. Identifying effective manipulations of situations that affect appropriate exertion of control will further understanding of VBSJT functioning and identify important situational components to be considered when developing VBSJTs. Thus, the second purpose of this study is to confirm the assumption that SJT situations contribute to the effectiveness of the test and identify specific situational characteristics that are most effective in manipulating exertion of control. The distinction between VBSJTs and written SJTs (WSJTs) can be directly linked to the effectiveness of such situational stimuli. The use of situational variables included in this study can demonstrate effective manipulations that are only possible with the use of dramatic video.

Another purpose of this study is to demonstrate the utility of this new framework for identifying appropriate exertion of control within a SJT to test hypotheses about appropriate exertion of control on the job. The measurement of appropriate exertion of control with simulated situational characteristics should provide useful information related to the actual

behavior of police officers on the job. Appropriate exertion of control of experienced police officers can be compared to that of applicants, to determine if there are differences between the two groups in exertion of control, as would be expected from the law enforcement literature (e.g., Beutler, Nussbaum, & Meredith, 1988; Burgin, 1978; Sterling, 1972). The following study will explore appropriate exertion of control to determine if experienced police officers show greater overexertion of control than applicants. These findings can help determine if proven negative side effects of experience as a police officer actually translate to more negative interactions with the public.

The final consideration of this study is to demonstrate other ways to utilize SJTs on the job. Appropriate exertion of control will also be applied to test the hypothesis that location of an organization impacts the likely behavior of the police officer, as would be expected from the law enforcement literature (e.g., Beutler, Storm, Kirkish, Scogin, & Gaines, 1985). Overexertion of control displayed by officers taking the test may also be a function of organizational culture, which may reinforce these controlling behaviors as appropriate. Therefore, differential prediction of appropriate exertion of control will also be explored across locations.

The purpose of this study is to provide insight into SJT functioning that has not been established in previous research. Knowledge of multidimensionality at the item and test level will provide the key framework for understanding SJTs. This paper will hopefully provide the beginnings of a theoretical framework for understanding SJTs and a pathway for future research.

History of Situational Judgment Tests

Situational judgment tests (SJTs) have been used for personnel selection for over fifty years (Clevenger, Jockin & Morris, 1999). The How Supervise? (File, 1945) and the Cardall Practical Judgment Test were the first measures of situational judgment to be documented in the literature. Mandell (1953) and Rosen (1961) provided validation for the use of the How Supervise?. The Cardall Test of Practical Judgment and executive ratings of supervisory performance correlated at $r = .31$, demonstrating the effectiveness of this test (Dulsky & Krout, 1950). Bruce and Learner (1958) followed closely with the introduction of the Supervisory Practices Test. The Supervisory Practices Test was used to identify supervisors who could make good decisions in work related situations. This test effectively identified supervisors from non-supervisors (Bruce and Learner, 1958). Tenopyr (1968) demonstrated the usefulness of the Leadership Evaluation and Development Scale (LEADS), another example of an early SJT.

Although these SJTs were shown to be effective predictors of performance, SJTs did not gain a significant amount of attention in the literature until the early 1990's.

Motowidlo, Dunnette, and Carter (1990) reintroduced the situational judgment test to current literature. In exploring the situational judgment test, Motowidlo et al. (1990) distinguished the SJT from a traditional job simulation. Motowidlo et al. (1990) use the terms "high fidelity" and "low fidelity" to refer to the differences. They defined "high fidelity" measures of situational judgment as those that place the examinee in simulated situations requiring the examinee to behaviorally respond in the same manner as on the job. Assessment centers, work samples, and other direct simulations are referred to as "high fidelity". "Low fidelity" measures of situational judgment are defined as those that describe job situations requiring the examinee to make a response that would be typical of their job related behavior. Motowidlo et al. (1990) used two different samples to validate their SJT. They found validity estimates of $\underline{r} = .30$ and $\underline{r} = .32$ for overall effectiveness of the samples. Their exploration of these formats led to the conclusion that less expensive "low fidelity" measures of situational judgment can predict as well as more expensive "high fidelity" measures (Motowidlo, Dunnette & Carter, 1990). Although the usefulness of these tests was demonstrated by early research (e.g., Bruce and Learner, 1958; Rosen, 1961), Motowidlo et al.'s (1990) work has led to the recent popularity of "low fidelity" measures of situational judgment including VBSJTs (e.g., Dalessio, 1994, Jones & DeCotiis, 1986, Weekley & Jones, 1997) and WSJTS (e.g., Weekley & Jones, 1999).

Furthermore, recent literature has promoted the fact that SJTs have provided useful advantages over traditional selection instruments such as cognitive ability and personality tests. Some of the documented weaknesses of traditional measures are overcome by the use of SJTs. Although cognitive ability tests are considered to be among the most valid of tests for nearly all jobs (Schmidt & Hunter, 1998; Schmidt & Hunter, 1981; Hunter & Hunter, 1984), these tests often produce large mean differences between racial subgroups. Specifically, the difference is commonly found to be one standard deviation between African Americans and Caucasians (Chan & Schmitt, 1997; Hunter & Hunter, 1984). Although these tests typically do not yield differential prediction, the large subgroup differences cannot be overlooked (Chan & Schmitt, 1997; Cleary, 1968). Attempts have been made to reduce these differences by examining and modifying the tests (DeShon, Smith, Chan, & Schmitt, 1998). However, changing the tests does

not typically alleviate subgroup differences. These failed attempts have led researchers to examine the possibilities of using multiple predictors to reduce subgroup differences of cognitive ability tests (Ryan, Polyhart & Friedel, 1998; Sackett & Ellingson, 1997; Sacket & Roth, 1996; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). These results indicate that mean differences will be present unless alternative predictors, which typically have lower validities, are given significantly higher weights than cognitive ability (Ryan et al., 1998; Schmitt et al., 1997).

Personality inventories are perhaps the most commonly explored alternative to cognitive ability tests. In recent literature, personality inventories as personnel selection instruments have been thoroughly explored (Barrick & Mount, 1991; Hough, 1998; Ryan, Polyhart, & Friedel, 1998; Schmidt & Hunter, 1998). This literature suggests that personality as a predictor of job performance is far less valid than cognitive ability and that the use of such inventories leads to concern over response distortion (Elliot, Lawty-Jones, & Jackson, 1996, Stanley & Stokes, 1998). However, the small subgroup differences found for personality inventories have supported interest in personality testing (Schmidt & Hunter, 1998).

SJTs have become popular predictors of performance because they demonstrate smaller mean differences than cognitive ability without sacrificing validity (Motowidlo & Tippins, 1993; Robertson & Kandola, 1982). Typically, the standardized difference between African American and Caucasians is one third of a standard deviation (e.g., Clevenger, Jockin, Morris, & Anselmi, 1999; Motowidlo & Tippins, 1993; Strong & Najar, 1999). These measures comply with EEOC guidelines directing that alternative methods of selection be used when they demonstrate less adverse impact and have comparable validities (Strong & Najar, 1999). Furthermore, supporting Motowidlo et al. (1990), SJTs have been demonstrated to have similar validities to traditional measures of cognitive ability (Pereira & Harvey, 1999; Weekley & Jones, 1999). This finding has also been demonstrated in a recent meta-analysis of the validity of SJTs. McDaniel, Finnegan, Morgeson, Campion, and Braverman (1997) found the validity for situational judgment tests to be $\underline{\rho} = .56$. This is comparable to Hunter and Schmidt's (1998) validity finding of $\underline{\rho} = .51$ for cognitive ability. Further studies on situational judgment tests have consistently demonstrated strong relationships with job performance. The results of the studies subsequent to the meta-analysis demonstrate the validation of these tests but are smaller than the population estimate because corrections have not been made.

The criterion validity of WSJTs has been the most thoroughly demonstrated (Bruce & Learner, 1958; Clevenger, Jockin & Morris, 1999; McDaniel, Finnegan, Morgeson, Campion & Braverman, 1997; Motowidlo, Dunnette, & Carter, 1990; Motowidlo & Tippins, 1993; Pereira & Harvey, 1999; Pulakos & Schmitt, 1996; Robertson & Kandola, 1982; Smith & McDaniel, 1998; Strong and Najar, 1999; Tenopyr, 1969; Weekley & Jones, 1999). Motowidlo et al. (1990) report validities of $\underline{r}$ = .28 to $\underline{r}$ = .37. Motowidlo and Tippins (1993) report validities with overall job performance of $\underline{r}$ = .31 and with communication effectiveness of $\underline{r}$ = .33. In a second study using ratings of performance with incumbents in a marketing position, they report validities ranging from $\underline{r}$ = .14 to $\underline{r}$ = .29. They found a predictive validity of $\underline{r}$ = .25 and a concurrent validity of $\underline{r}$ = .20 with supervisory ratings of overall performance. Pulakos and Schmitt (1996) also found an overall validity of $\underline{r}$ = .24. McDaniel et al. (1997) further supported this strong evidence for the use of WSJTs with their meta-analysis across 95 studies.

Recent studies have continued to demonstrate the effectiveness of these predictors. Weekley and Jones (1999) found an average weighted correlation with supervisory rated performance of $\underline{r}$ = .19 across two studies. Pereira and Harvey (1999) found situational judgment to have a significant validity, $\underline{r}$ = .18, as measured with supervisory ratings. Smith & McDaniel (1998) found validity coefficients of $\underline{r}$ = .31 and $\underline{r}$ = .30. Strong and Najar (1999) found that the situational judgment test had an overall validity of $\underline{r}$ = .25. Clevenger, Jockin & Morris (1999) found an average correlation of $\underline{r}$ = .20 with nine supervisory rated dimensions.

There is also literature that suggests that VBSJTs have comparable criterion validities (Jones & DeCotiis, 1986; Swander & Spurlin, 1993; Swander & Spurlin, 1995; Swander & Spurlin, 1997a; Swander & Spurlin, 1997b; Swander & Spurlin, 1998; Weekley & Jones, 1997). Jones and DeCotiis (1986) found an overall test validity of $\underline{r}$ = .38 ($\underline{r}$ = .55, corrected for measurement errors) for their guest relations VBSJT. Weekley and Jones (1997) obtained a validity coefficient of $\underline{r}$ = .33 using a developmental sample and $\underline{r}$ = .22 ($\underline{r}$ = .34, corrected for criterion unreliability) in a cross validation sample of recently hired employees. In a second study, Weekley and Jones (1997) found a validity of $\underline{r}$ = .18 ($\underline{r}$ = .28, corrected for criterion unreliability).

Further evidence of criterion validity has been demonstrated by Ergometrics. Unpublished validation reports from five different VBSJTs indicate validity coefficients ranging from $\underline{r}$ = .33 ($\underline{r}$ = .49, corrected for criterion unreliability) to $\underline{r}$ = .48 ($\underline{r}$ = .56, corrected for

criterion unreliability). These validities ranged over five tests and 20 samples from different organizations (Swander & Spurlin, 1993; Swander & Spurlin, 1995; Swander & Spurlin, 1997a; Swander & Spurlin, 1997b; Swander & Spurlin, 1998).

Although SJTs have only recently regained interest, the body of empirical literature has been increasing rapidly. The current literature has clearly demonstrated that SJTs (video or written) are good predictors of job performance. However, not only is it important for a test to predict performance but it is also important to understand how the test predicts performance. Many researchers have explored the possible underlying constructs of SJTs. The two most commonly explored constructs are cognitive ability and personality. Also, variables such as job experience and practical and emotional intelligence have been explored. Following is a summary of results reported in the literature relating SJTs to specific constructs.

General Cognitive Ability

General cognitive ability has been explored as an underlying construct of SJTs. Specifically, many authors have explored the relationship between WSJTs and cognitive ability (Bruce and Learner, 1958; Carrington, 1949; Chan and Schmitt, 1997; Dulsky & Krout, 1950; Jones, Dwight and Nouryan, 1999; McDaniel et al., 1997; Motowidlo et al., 1990; Mullins and Schmitt, 1998; Pereira & Harvey, 1999; Rosen, 1961; Smith and McDaniel, 1998; Weekley and Jones, 1999). McDaniel et al. (1997) summarized the work on SJTs with a meta-analysis. They found a relationship of $\underline{\rho} = .53$ ($\underline{\sigma_p} = .29$) between written measures of situational judgment and cognitive ability. More recent work with SJTs has also resulted in similar findings. For example, Smith and McDaniel (1998) found a correlation with cognitive ability of $\underline{r} = .35$. Weekley and Jones (1999) found a correlation of $\underline{r} = .42$ with cognitive ability. Jones, Dwight and Nouryan (1999) used a measure of cognitive ability (business reasoning assessment) and found that it correlated highly with total score on their WSJT ($\underline{r} = .30$).

A relationship between cognitive ability and situational judgment has also been found for VBSJTs (Schmiderle et al., 1994; Weekley & Jones, 1997). Weekley and Jones (1997) found that their VBSJT was significantly related to cognitive ability in two different samples. The correlations were $\underline{r} = .33$ and $\underline{r} = .29$. Schmiderle et al. (1994) found a significant correlation between cognitive ability and the Seattle Metro Video Test of $\underline{r} = .24$.

Although these findings suggest that a large component of SJTs is cognitive ability, the relationship has not been demonstrated in other studies (Motowidlo et al., 1990; Mullins and Schmitt 1998; Smith and McDaniel, 1998; Swander, 2000). This inconsistent pattern appears with both formats of SJTs; suggesting that there might be other variables moderating this relationship. Furthermore, McDaniel et al. (1997) found that the estimated population variance within the meta-analysis was quite large, suggesting that this correlation varied significantly depending on the SJT. These variables could arise from the nature of the job, the sample used or any other number of job related variables.

Reading Comprehension

Reading comprehension has been identified as an important component in the relationship between cognitive ability and SJTs. Chan and Schmitt (1997) identified this relationship when they examined the differences between a WSJT and a VBSJT. They found that reading comprehension was related to the WSJT but not the VBSJT. Swander (2000) further supported this relationship. Pereira and Harvey (1999) found low correlations between cognitive ability and their SJT across two samples (r = .12 and .12). This relationship may have contributed to no relationship between reading comprehension and the SJT (r = .02 and .04). This finding suggests that reading comprehension may be an important moderator in the relationship between cognitive ability and SJTs.

Sacco, Scheu, Ryan, Schmitt, Schmidt, and Rogg (2000) further explored this relationship by examining readability statistics of WSJTs. They found that reading level of the WSJT was related to subgroup differences and validity. Furthermore, the harder a SJT is to read the more likely it is to be more highly correlated with general measures of cognitive ability.

Although reading comprehension has been identified as an important component of the relationship between WSJTs and cognitive ability, it does not apply to VBSJTs. However, the validity of VBSJTs is comparable to that of WSJTs. Added situational variables that can be included with the use of video may account for the validity of VBSJTs.

Personality

Personality is also a commonly explored underlying construct of SJTs (Carrington, 1949; Jones, Dwight & Nouryan 1999; Mullins & Schmitt, 1998; Pereira & Harvey, 1999; Smith & McDaniel, 1998; Swander, 2000). Rationale for this inquiry is based on mean differences

between ethnic subgroups. Mean differences for both SJTs and personality measures are much lower than for cognitive ability tests (Clevenger, Jockin, & Morris, 1999).

For example, Mullins and Schmitt (1998) found a relationship between a WSJT and conscientiousness ($\underline{r}$ = .26) and agreeableness ($\underline{r}$ = .22). Smith and McDaniel (1998) found conscientiousness ($\underline{r}$ = .32) and emotional stability ($\underline{r}$ = .22) to be correlated with their WSJT. Smith and McDaniel (1998) found that dependability ($\underline{r}$ = .32) and emotional stability ($\underline{r}$ = .22) scales from the HPI were correlated most highly with the WSJT. Pereira and Harvey (1999) found conscientiousness ($\underline{r}$ = .23, .21), caring ($\underline{r}$ = .31, .30), persuasiveness ($\underline{r}$ = .25, .22) and optimism ($\underline{r}$ = .27, .27) to have the strongest relationships with their WSJT.

Based on the results of these studies, conscientiousness appears to be the only stable personality correlate of WSJTs. This finding is not particularly surprising because of the big five personality dimensions, conscientiousness has been demonstrated to be the most predictive of job performance (Barrick & Mount, 1991). However, there have also been studies that have demonstrated no relationship between conscientiousness and SJTs (e.g., Swander, 2000). Furthermore, based on the Swander (2000) study, there are no significant differences in the measurement of personality constructs between WSJTs and VBSJTs.

Experience

Job related experience has also been explored in the literature on SJTs. Experience is a hypothesized underlying construct of SJTs because it is assumed if that people have experience dealing with situations similar to those presented in the test then they might have an advantage over those who have no experience. That is, examinees who have been placed in these situations know how they handled them in the past. Furthermore, examinees with experience could have evaluated various outcomes of their own decisions. However, some individuals may also have made wrong decisions in similar situations in the past and not had the ability or desire to see how outcomes were influenced by their decisions. The literature on this topic is primarily inconclusive, as with the other underlying constructs explored up to this point.

For example, Smith and McDaniel (1998) found length of job experience ($\underline{r}$ = .27) and working as a supervisor ($\underline{r}$ = .30) to be fairly large correlates of a SJT. Weekley and Jones (1997) also found work experience to correlate with their VBSJT ($\underline{r}$ = .26 and .16). These findings have led some to believe that SJTs are purely measures of skills gained through life and job experiences. However, Mullins and Schmitt (1998) did not find any relationship between

work experience and scores on the situational judgment test. Motowidlo and Tippins (1992), Jones, Dwight and Nouryan (1999), Mullins and Schmitt (1998), and Swander (2000) also did not find a relationship between SJTs and tenure. Pereira and Harvey (1999) found that their SJT had only small correlations with a measure of job-specific skills ($\underline{r}$ = .059 to .141). Thus, even with specific skills acquired through job experience, the examinee is not likely to be aided when answering SJT items. The inconclusive results of personality, cognitive ability, and job related experience have led some researchers to conclude that SJTs are measures of a unique construct. Specifically, that SJTs are measures of practical intelligence. The direct relationship between SJTs and practical intelligence has been explored.

Practical Intelligence

Practical Intelligence is knowledge gained through life experiences that is not openly stated (Wagner & Sternberg, 1985). Wagner and Sternberg (1985) developed a measure of practical intelligence for managers called the Tacit Knowledge Inventory for Managers (TKIM). It is contended that tacit knowledge is not a measure of intelligence or personality. It is suggested that tacit knowledge measures a unique construct (Sternberg & Wagner, 1985; Weekley & Jones, 1999). Motowidlo et al. (1990) implied that tacit knowledge is closely related to situational judgment. That is, SJTs measure this unique construct, not found in the other measures to which SJTs have sometimes been compared. However, when a SJT was directly compared with a test of tacit knowledge, the results did not support this hypothesis. Mullins and Schmitt (1998) found that their SJT and a measure of tacit knowledge for managers (TKIM; Sternberg & Wagner, 1985) were essentially unrelated ($\underline{r}$ = .08).

Emotional Accuracy

Another construct that has been studied in relations to SJTs is emotional accuracy. One key factor in interpreting the situations in a SJT may be emotional accuracy. That is, the examinee must understand the emotion of the characters in the SJT. Wrongful interpretation of emotion may lead the examinee to make inappropriate decisions about what to do in the situation. Swander (2000) found that a WSJT and a VBSJT were both related to emotional accuracy ($\underline{r}$ = .19 and .28, respectively) as measured with the Emotional Accuracy Research Scale (EARS; Mayer & Geher, 1996). Although this study demonstrated a relationship between emotional identification and a SJT, it is the only study to explore such a relationship. Thus, a

consistent relationship has not been identified and would need to be further explored to indicate that SJTs are consistent measures of emotional accuracy.

Although it is reasonable to assume that many of these constructs would be related to SJTs, conclusions based on these studies are ambiguous. Researchers have recently begun to explore other aspects of SJTs, rather than correlating them with specific constructs. Specifically, one explanation of the variance in correlates is the nature of the job knowledge the SJT is designed to measure.

Contextual Versus Task Knowledge

It is suggested that SJTs can be designed to measure different aspects of job knowledge (Clevenger, Pereira, Weichmann, Schmitt, & Schmidt-Harvey, in press). Specifically, SJTs may measure either contextual or task job related knowledge (Clevenger, Jockin, Morris, & Anselmi, 1999). Task knowledge is related to the specific behaviors needed to perform the job, while contextual knowledge is conceptualized as interactions in social situations with others, such as supervisors and customers. Knowledge of appropriate behavior within these situations is defined as contextual. Clevenger et al. (in press) specifically studied a test designed to measure contextual knowledge. Correlations with task knowledge were low ($\underline{r}$ = .09 and $\underline{r}$ = .18), indicating that the measure of situational judgment they designed was more of a measure of contextual job knowledge than task knowledge. It is hypothesized that the content of the SJT related to contextual and task knowledge will moderate the effects of cognitive ability and personality and clarify why the results have proven inconclusive. That is, contextual performance may relate more to personality while task performance relates more highly to cognitive ability. Although no research has specifically tested this hypothesis, it is likely that SJTs can be designed to measure both aspects of job performance.

This research may help to identify why SJTs designed to measure the same job often have different patterns of relationships with personality and cognitive ability. This may also help to explain why SJTs designed to measure the same job are not necessarily highly correlated. For example, Bruce and Learner (1958) demonstrated that SJTs designed to measure the same job were only correlated moderately. They found a correlation of $\underline{r}$ = .56 between the How Supervise? (File, 1945) and the Supervisory Practices Test, both of which are designed to measure supervisory performance. This indicated that there is overlap between the two tests, however, given that both tests were SJTs designed to measure supervisory potential, there is also

a fair amount of difference between the two tests. The different components of these tests may be related to the differences in focus of the tests, such as contextual or task job knowledge.

Contextual and task knowledge may serve as a useful categorization of some SJTs or items within a SJT. However, this classification may tend to blend together in many settings. For example, managers must be in contact and manage their subordinates on a daily basis. Questions designed to assess managerial performance could be based purely on interaction skills (i.e., contextual job knowledge) or they could be based purely on how they handle specific problems in a technical sense (i.e., task job knowledge). However, many questions in a SJT may contain a varying degree of both these factors. For example, a manager must be able to technically handle the problem well and interact effectively in the social situation presented. In order to correctly answer a question of this type the examinee must possess both contextual and task job knowledge. A wrong answer could be due to lack of either contextual or task knowledge.

As can be seen by the inconsistencies in the literature and the direction that the research is headed, there is a lack of understanding of the theoretical framework of SJTs but there is a desire to understand. Although making the distinction between SJT items as measures of contextual or task job knowledge may lead to further understanding of SJTs, it may be more appropriate to make classifications within each item. SJTs are multidimensional, which perhaps means that not only do they include items that measure different factors but that performance on specific items may be impacted by different factors. This knowledge may lend itself to more productive methods for understanding SJTs.

Multidimensionality of Situational Judgment Items and Tests

The fact that SJTs are typically designed for a particular jobs (e.g., police officers, managers, engineers) and the variety of situations and stimuli presented may be reasons why construct validity studies have found inconsistent results. SJTs are not designed as construct measures. Typical SJTs are designed to capture many aspects of job performance. Many researchers acknowledge that SJTs are indeed multidimensional, not only on the test level but also on the item level (e.g., Chan and Schmitt, 1997, Jones, Dwight, and Nouryan, 1999; Smith & McDaniel, 1998). Internal consistency reliability estimates of SJTs are commonly lower than traditional selection instruments. Low internal estimates of reliability are often a result of the multidimensionality of SJTs. McDaniel et al. (1997) compared the summarized literature on

SJTs to that of situational interviews and assessment centers, describing them as testing methods that generally tap a variety of constructs. Multidimensionality suggests that individual constructs may be identified within SJTs and that they are not measures of a single unique construct. However, the constructs measured by SJTs depend on the situations described by the items (Jones, Dwight & Nouryan, 1999) and the job. As Weekley and Jones (1999) contend, variations in item content of a SJT will lead to different correlates. They point out the need to develop SJTs that are designed to measure a single construct or identify homogenous subsets of items that might help to identify specific constructs. Jones, Dwight, and Nouryan (1999) also suggest that the design of the SJT is related to the constructs that are measured. They designed a SJT to measure judgment and decision-making skills of managers instead of interpersonal skills. They did find that the SJT was more strongly related to analytical ability, business reasoning, and decision-making. Instead of designing a SJT to measure a single construct, an alternative would be to identify constructs within the answers chosen. This will provide insight into how constructs are related to SJTs.

Multidimensionality does not necessarily mean that subsets of items can be identified that represent certain constructs. Factor analysis of a SJT has provided evidence that no meaningful factors can be identified (Swander & Spurlin, 1995). Swander & Spurlin (1995) found that their VBSJT with 54 items had 16 factors with eigenvalues between 1 and 1.3. Thus, no interpretable factors were identified. The inter-item correlations are also typically around $\underline{r} = .1$ (Swander & Spurlin, 1995). Multidimensionality at the item level may account for these results. Thus, SJT research needs to be conducted with the knowledge of multidimensionality at the item level. That is, research is needed to explore the items and identify constructs that may help to explain why people choose specific answers. Typically test item theory considers a correct answer to be a positive indicator of a particular construct, such as conscientiousness or reading comprehension, and a wrong answer to be a negative indicator of that construct. However with situational judgment items, wrong answers can be more indicative of different constructs than a correct answer. For example, in a public relations question where a customer is disputing the quality of service, the correct answer "I'm sorry, I'll see if I can get this fixed," is a good predictor of the public relations skills construct. However, the wrong answer, "I don't appreciate the way you are talking to me," may be more of an indicator of aggressiveness. Whereas, the wrong answer, "I'm sorry, but there is nothing I can do," may result from an incorrect perception

about organizational expectations. Thus, in understanding how situational judgment tests function, it may be important to consider alternative item responses separately and not just whether the question was answered right or wrong.

Purpose of Study

The general purpose of this study is to provide insight into SJT functioning that has not been established in previous research. Knowledge of multidimensionality at the item and test level will provide the key framework for understanding SJTs. First, this knowledge will be applied to demonstrate an effective method for identifying a unidimensional construct within a multidimensional SJT. Second, the situational characteristics within the items will be explored for their impact on this specific construct. The final consideration of this study is to demonstrate other ways to utilize SJTs on the job. This utility will be directly related to the proposed framework. All of these goals will hopefully provide insight into a theoretical framework for understanding SJTs and a pathway for future research and expand the utility of SJTs.

The current study will employ relevant items from a VBSJT designed for law enforcement officers. The construct that will be examined is appropriate exertion of control. The appropriate exertion of control construct will then be used to determine situational manipulations that are effective for capturing responses that are related to over- or underexertion of control. The measurement of appropriate exertion of control using a VBSJT will then be applied to test hypotheses about police officer exertion of control.

Appropriate Exertion of Control

Police officers have a high level of power that can be used to exert control in many situations (Conroy & Hess, 1992). Control for a police officer conveys the authority to detain, arrest, and interfere with the activities of others. Police work involves situations in which an officer must use police authority to exert control over others in order to accomplish the job. Appropriate exertion of control is dependent upon the situation (Desmedt, 1984). Using force is necessary in some situations and thus not a wrong response, but an appropriate job related behavior. Using unnecessary force is an obvious example of inappropriate exertion of control. Poor public relations that may result from using police authority to facilitate inappropriate interactions with citizens is also a concern. Using unnecessary force and other behaviors that represent overexertion of control constitute a serious problem for law enforcement organizations.

Over-exertion of control exceeds the control necessary to complete the police objective in that situation. Besides exerting more control than is necessary in a situation that actually requires some control, overexertion of control also includes interventions into non-police matters and harsh or rude behavior that is not necessary and would not be tolerated but for the officer's position. Endorsed behaviors that are unnecessary overreactions or misuse of authority will be defined as overexertion of control. Unwarranted attention and insensitivity to the effect of harsh actions on observers are also examples of overexertion of control in a police context. Therefore, overexertion of control in the context of the VBSJT is any behavior that is an unnecessary attention, unnecessary harshness, overreaction, or misuse of authority. This definition is supported throughout the test and job analysis.

Underexertion of control is also a police performance problem. Underexertion of control means exerting less control than is necessary to properly manage a situation that calls for police response. Underexertion of control includes reluctance to appropriately intervene, enforce the law or investigate suspicious situations.

Appropriate Exertion of Control in Law Enforcement

The importance of appropriate exertion of control has not only been demonstrated through a rigorous job analysis in the development of the VBSJT but the importance of this construct can also be related to the law enforcement literature. Particularly, the importance of stress and aggression as predictors of police performance can be seen (Beutler et al., 1985; Burkhart, 1980; Murphy, 1972). These variables have become recognized as important predictors of police officer performance because of behavioral outcomes that are related to appropriate exertion of control (Conroy and Hess, 1992).

Research on evaluation and selection of police officers has often been linked to psychological disorders. However, many authors suggested that police officers mainly fall within normal range on many pathological tests (e.g., Saccuzzo, Higgins, & Lewandowski, 1974; Saxe & Reiser, 1976). This limits the usefulness of pathological measurement scales to predict applicant performance. Although in the past the selection of law enforcement officers was primarily dependant on pathological scales, recent literature has identified the need to test for other relevant variables (Beutler, Nussbaum, & Meredith, 1988; Beutler, Storm, Kirkish, Scogin, & Gaines, 1985; Burkhart, 1980). In particular, the need for testing individual capacity to tolerate stress and demonstrate emotional stability was identified (Murphy, 1972). These

dimensions were based on their relationships to supervisors' ratings of stress tolerance and aggression (Beutler et al., 1985; Burkhart, 1980).

Stress experienced by police officers has been widely studied and documented. Police officer stress comes from many sources and is a consistent result of experience on the job. Much of the stress comes from confusing roles and expectations. Officers must sustain authority in potentially confrontational situations without overreacting or resorting to behaviors that would be considered police brutality. Officers are also expected to interact with appropriate courtesy in non-conflict types of situations. The lack of simple consistency in what is expected can lead to tensions and stress for police officers (Mihanovich, 1981). To accomplish their work, police must intrude in the lives of members of the public, sometimes in situations that are not criminal. Since all situations are different, officers are not given specific rules about how to handle every situation. They are expected to assess situations and make final decisions on actions to be taken (Desmedt, 1984). This is usually referred to as police discretion. Officers also often suffer from isolation and loneliness that can lead to feelings of stress (Conroy & Hess, 1992). Job specific stress can end in aggressive acts towards the public, whom the police are supposed to protect (Mihanovich, 1981).

Police officers face unique circumstances that demand the ability to cope with changing situations (Beutler, Nussbaum, & Meredith, 1988). They must be able to use judgment in the appropriate exertion of control. Police officers must not only be able to assert themselves in confrontational settings but also maintain good public relations skills in nonconfrontational settings. That is, they must be controlling in some situations but not in others. Some officers have difficulty with this and overreact or become involved in situations that do not require police attention. Discerning situational differences and responding appropriately is critical to police performance. Beutler et al. (1985) found that a measure of impulsivity was significantly related to reprimands related to using excessive force. Interpersonal insensitivity was also found to be related to excessive use of force (Beutler, et al., 1985). Officers who cannot distinguish among situations requiring different levels of intervention are likely to upset people and put themselves in danger. Brown (1981) identified two dimensions of policing style. The dimension he defines as aggressive is characterized by problems here referred to as overexertion of control. Aggressive policing style includes typically using force and asserting authority in non-criminal situations when intervention is not necessary. Police officers are commonly screened out on the

basis of aggression or hostility (Abernethy, 1995), demonstrating the seriousness with which law enforcement organizations seek to avoid problems associated with these dimensions. Poor stress tolerance and aggression, which can both lead to inappropriate exertion of control, have also been identified as important variables in the police literature (Sewell, 1984).

Conditional Reasoning

Although no research has attempted to identify a construct within a SJT by considering multidimensionality at the item level, a method similar to this has been established for the measurement of personality via conditional reasoning (James, 1998). James (1998) introduced the concept of conditional reasoning to the measurement of personality. Specifically, the constructs of achievement motivation (Burgess & James, 1998; James, 1998; Migetz, James, & Ladd) and aggression (Green, 1999; James, 1998; James, McIntyre, Glisson, Green, Patton, Mitchell, & Williams, 2000; McIntyre, 1995; Patton, 1998) have been studied using measures of conditional reasoning. Measures of conditional reasoning are developed to measure specific personality constructs without the problems of self-report measures. Not only are self-report personality measures susceptible to response distortion, but they have also been show to have less than moderate correlations with job performance. Measures of conditional reasoning are founded on the idea that people use justification mechanisms (James, 1998). Justification mechanisms serve as reasoning processes through which individuals rationalize their decisions and behaviors. Different justification mechanisms elicit different behavioral responses. Conditional reasoning tests are based on the measurement of these justification mechanisms. These measures utilize ambiguous reasoning items with which people are instructed to identify the answer that justifies the problem presented. James (1998) suggests that personality traits underlie justification mechanisms used. For example, a person with an aggressive personality will rationalize the use of hostility or aggressiveness (James et al., 2000). The theory holds that people will agree with evidence that supports their motives, while disagreeing with evidence supportive of motives with which they do not agree (Fisk & Taylor, 1999; James, 1998).

The theory of justification mechanisms and conditional reasoning can be directly applied to identifying constructs within SJTs. Although SJTs are not designed to measure specific personality characteristics, behavioral choices could be identified that represent specific constructs. Applying this method to a SJT is similar to doing so with a conditional reasoning test. That is, a single construct is identified and scored based on the identification of responses

that most likely represent a specific construct. An examinee must respond to the situations and from these responses the appropriateness of the examinee's exertion of control could be identified. However, SJTs are not designed to measure single constructs but are designed for specific occupations. This limits the constructs that may be identified within a SJT. As mentioned previously, it is important that the job the test is designed for is also considered when identifying constructs that may influence responses to the items.

The first goal of this study is to identify a method for the measurement of the appropriate exertion of control component within the VBSJT. James' method of measuring justification mechanisms will be applied to the VBSJT to identify over- and underexertion of control. Appropriate exertion of control is an important ability of police officers and was therefore one of the primary components incorporated in the development of the VBSJT. As mentioned previously, the importance of stress tolerance and aggression indicate that overexertion of control is the more prominent and problematic than underexertion of control. Tendency toward overexertion of control was considered to be a more important component to attempt to isolate prior to hiring as it represented a more difficult and frustrating problem for law enforcement managers to remedy. Overexertion of control was therefore the subject of greater focus than underexertion of control during the development of the VBSJT. It is expected that applying this method of identifying the unidimensional construct of overexertion of control should yield an important component of the VBSJT. This will be evaluated by testing the significance of the validity of this construct and identifying that it is one of the key dimensions within the VBSJT. Therefore, it is hypothesized that the overexertion of control scale (OCS) will have significant validity but the validity will be a portion of the original VBSJT.

H1: The OCS will have significant validity, which will be accounted for by the original VBSJT.

The results of this hypothesis will help determine if a VBSJT can be scored to identify one critical component of the job and the test. This will help to establish a method for identifying a unidimensional construct within a SJT. Underexertion of control will also be explored in relation to validity and all other hypotheses. However, no specific hypotheses will be made because of the limited focus on underexertion during test design. Underexertion of control may not be represented by enough items to make any formal hypotheses about the functioning of the scale.

The OCS can help to determine characteristics within the VBSJT items that are effective in eliciting answers that represent overexertion of control. That is, situational manipulations that are found to influence the overexertion of control responses can be considered useful characteristics of the situation. The underexertion of control scale (UCS) will also be used to explore situational characteristics that may have an influence on applicant underexertion of control responses.

Although this method for identifying a construct can be applied to a SJT, it should be noted that SJT items are highly dissimilar to measures of conditional reasoning. Whereas measures of conditional reasoning utilize ambiguous stimuli to identify the justification mechanism, SJTs use job related situations in which stimuli are included to elicit mistakes that would be made on the job.

Situational Characteristics

The situations represented in the VBSJT need to be explored to see if certain situations elicit more consistencies in response patterns. With all the research dedicated to finding variables that underlie SJTs, no articles have evaluated what makes situational judgment situational. SJTs are different from cognitive ability tests or personality tests because they present questions in the context of the particular job the test is being used to assess. The situations and the reactions to the situations are the important components of the question. The identification of specific situational characteristics that explain item responses will not only increase understanding of VBSJTs but aid test developers in choosing effective stimuli to include in item design.

Specifically, the situational characteristics will be related to both overexertion of control and underexertion of control. VBSJTs are designed to simulate real situations that would be encountered on the job. Thus, stimuli in the situations are purposefully included in order to simulate the actual job. Although these characteristics are important to the development of a VBSJT, no research has evaluated if the stimuli do affect item responses or how the item responses are affected. The assumption that the situational component of a SJT is the primary measurement mechanism is unproven. That is, we assume that SJTs more closely replicate the job than other selection instruments through the inclusion of situational variables but we have never proven that the included situational components have an effect. Through the identification of over- and underexertion of control responses the impact of certain situational variables can be

seen. These situational manipulations can be evaluated on their impact on appropriate exertion of control. Thus, it can be determined if the characteristics within the test are truly evoking over- or underexertion of control responses. Although it is useful to identify a specific construct embedded within a SJT, it is even more practically important to identify situational components that affect answer endorsement. There are probably strong relationships between environment, or situations, and the behavior elicited by the police officer (Burkhart, 1980). In other words, there are most likely situational "hot buttons" that tend to set off inappropriate exertion of control. Identifying key characteristics in VBSJTs that elicit over- or underexertion of control will help to identify the mechanisms that underlie the usefulness of SJTs and VBSJTs and will help to aid in SJT development. Furthermore, it can be helpful for identifying specific weaknesses of applicants and specific problem areas for incumbent police officers.

Situational factors have been recognized as important determinants of behavior. Strict use of trait theory of behavior has been criticized because of the malleability of behavior in similar situations (Mischel, 1968). The criticisms of solely trait-oriented behavior have come from both the social learning (Mischel, 1968) and ecological (Moos & Insel, 1974; Sells, 1966) perspectives. The state-trait argument within the personality literature has been widespread (Bowers, 1973; Epstein & O'Brien, 1985; Kenrick & Funder, 1991; Mischel, 1968; Mischel, 1990; Shoda, Mischel, & Wright, 1994). The situational approach to personality largely began in the 1960's (Kenrick & Funder, 1991). Many researchers began to explore the idea that environment plays a significant role in the personality demonstrated by the individual. There are several ideas that support the fact that situational variables may be important in the personality of an individual. One of the strongest arguments against personality traits is that correlations between personality scores and behaviors are rarely larger than $\underline{r} = .30$. Although many arguments have been made for a purely trait or environmental approach to personality, the person situation interaction is probably a more likely proposition. Bem and Funder (1978) explored how traits may only show in certain situations. Furthermore, some personality traits may be constrained during certain situations and are more likely to be present in other situations. For example, students in a classroom are less likely to demonstrate distinctive personality traits than those at a party. Personality traits may be more easily expressed in certain situations. Furthermore, differences in personality across situations may also be related to how an individual

interprets the situation (Mischel, 1990). Thus, situations interpreted as calling for exertion of control are likely to be different depending on the person and the situation.

With the construct of appropriate exertion of control being the major focus of this study, it is particularly relevant to consider the situational factors that contribute to the endorsement of over- and underexertion of control responses. It is not contended that the personality of the individual changes depending on the situation, but that the situation is interpreted as needing or not needing a controlling response. For example, rude characters within the test may evoke anger which may lead the examinee to choose a response that represents overexertion of control. This overexertion of control could be directly attributed to the fact that the character was rude. Not only is the classification of situations useful to understanding SJTs, it can also be helpful in identifying problem areas of individual police officers. Police are often exposed to anger provoking situations through interactions with citizens (Abernethy, 1995). Impulsive behavior, such as aggression, can result if the anger is not controlled in these situations (Hecker & Lunde, 1985). Patterns within item responses may suggest certain situations where officers tend to overreact or behave aggressively. This may help to determine training needs of the individual.

The SJT used in the current study is video-based. VBSJTs have some advantages over WSJTs, including more realistic situational characteristics. Demonstrating that situational characteristics of VBSJTs can influence the results will support the use of video. Furthermore, this can also help to explain why VBSJT validities are as large as WSJT validities without the reading comprehension component. Recently, video has become recognized as an effective medium to select employees (Smiderle, Perry, & Cronshaw, 1994). Video situations allow for movement, a richer environment, and the ability to fully demonstrate example situations (Weekley & Jones, 1997). VBSJTs offer an added component of reality with the use of video stimuli. Video provides visual information such as environmental details and facial expressions. It is clear that VBSJTs do offer higher fidelity simulations in which characters can be viewed more as they would be in the real world. Test takers are presented with facial and body expressions, tone of voice, physical environment, and dynamics (Jones & DeCotiis, 1986). Emotions of the characters in the test are more fully demonstrated, which can lead respondents to make more judgment errors based on overreaction. Furthermore, ethnicity and gender of the characters in the test are subtly incorporated, allowing for exploration of bias. Although research has explored the differences between written SJTs and VBSJTs, there have been no attempts to

understand how different stimuli affect performance on SJTs. There are situational characteristics that apply to both video and written SJTs, however, there are many components that are included in a VBSJTs that cannot be included in a WSJT. The use of video may help to identify specific situational factors important for eliciting over- or underexertion of control within law enforcement situations.

Provocation

First, the items will be classified according to overall provocation toward inappropriate exertion of control presented within the situation. Each situation will be explored for the presence of provoking stimuli that may underlie why a person may choose an inappropriate exertion of control response. Different stimuli are included in the VBSJT to evoke overexertion of control responses. These include character provocations such as the focal character being rude, aggressive, or behaving suspiciously. These also include informational manipulations such as the fact of a warrant, a complaint made about the focal character, or the crime in question being particularly offensive. They also include circumstances where an officer may assume resistance or escalation will occur, such as an arrest situation. Provocation toward overexertion of control is here defined as behaviors or circumstances that may cause greater temptation to respond in a way that is overly controlling. For purposes of identifying provoking stimuli, these characteristics will be defined as provocation toward overexertion of control. The VBSJT also includes stimuli that are attempts to elicit underexertion of control. These include character manipulations that represent the character as pleasant, cooperative, or sympathetic. These will be considered provocation toward underexertion of control. Although these provocations are specific to the type of inappropriate exertion of control, it is likely that they have an impact on both scales. That is, items with pleasant characters are likely to have better (i.e., lower) overexertion of control scores than those without pleasant characters and those with rude characters are likely to have worse (i.e., higher) overexertion of control scores than those without rude characters. All of these situational characteristics contribute to the overall provocation of the item. Overall provocation toward over-and under exertion of control will be assessed to determine if these variables affect scores on the OCS and UCS.

It is expected that overall provocation will affect appropriate exertion of control. If provocation toward overexertion of control is an effective manipulation then it would be expected that, on average, overexertion of control responses would be higher for the items with

provocation than for those items without provocation. Therefore, it is hypothesized that items with provocation toward overexertion of control will have worse (i.e., higher) scores on the OCS than items without provocation toward overexertion of control.

H2: Items with provocation toward overexertion of control will have worse (i.e., higher) scores on the OCS than items without provocation toward overexertion of control.

It is also expected that provocation toward underexertion of control will have an effect on the UCS. That is, items with provocation toward underexertion of control will have worse (i.e., higher) scores on the UCS than the items without provocation toward underexertion of control.

Although general provocation toward over- or underexertion will help identify potential effects that the situation may have on the answers to the test, it does not help identify specific contributing situational factors. Identifying specific situational characteristics can advance understanding of how to effectively develop SJTs and further provide key information about the differences between VBSJTs and WSJTs, as mentioned previously. Therefore, along with general provocation, individual components of provocation will be assessed. However, it is unknown at the onset of this research if the individual stimuli were effectively manipulated in enough items to make stable comparisons. Therefore general expectations will be stated with the knowledge that the specific comparison may not be possible.

The specific character manipulations that are expected to increase overexertion of control are rude, aggressive, or suspicious behaviors displayed by the focal character. Information expected to increase overexertion of control are crimes of a contemptible nature, complaints from other characters, warrants being enforced, laws being broken, and potential for others to become involved.

The specific character manipulations that are expected to increase underexertion of control are pleasant, cooperative and sympathy evoking characters. Although these situational characteristics are most closely related to the UCS it is likely that they may lead to less overexertion of control within the OCS. This is also expected to be the case when specific provocation toward overexertion of control is applied to the UCS.

Results from these comparisons will help demonstrate the importance of the situation itself for SJT items. Furthermore, exploring specific character manipulations will help to identify potential advantages of using video as compared to paper and pencil tests. Characters

are not realistically portrayed and manipulated on paper within the parameters of a test item, thus eliminating the possibility of including such stimuli within a WSJT. The effectiveness of character manipulations will provide very valuable information about the differences between these methods of SJT presentation.

Ethnicity

Another character manipulation that may affect appropriate exertion of control is ethnicity. The VBSJT allows for depiction of more realistic situations, which includes human diversity. The inclusion of diversity is particularly important due to the nature of police work. Diversity is directly related to racial profiling, an area of recent controversy in law enforcement (e.g., Drummond, 1999; Newport, 1999). Racial profiling is considered to exist when police question or detain citizens of certain ethnic groups in situations where they would not have detained others. Minorities are stopped because they are believed to be more likely than others to be committing crimes (Newport, 1999). Law enforcement in the United States is continuously under scrutiny for profiling with the suggestion that minorities are more likely to be detained. Police statistics are evaluated for the presence of racial profiling. A recent Gallup poll suggests that three out of four young African American men report being stopped by the police because of their skin color (Newport, 1999). These numbers suggest that racial profiling is an area of serious social concern. The media is dramatically increasing attention given to this subject matter. Although the VBSJT does not deal with the decision to stop people in vehicles, the most common form of profiling, there are many situations in which the target individual is a minority and the item responses involve choices on contact. Choosing answers that are overexertion of control in situations where a minority is the target may also indicate racial biases. Furthermore, if a respondent were to endorse overexertion of control toward minorities but not toward other citizens there would be even stronger evidence of bias. Because the recent literature suggests that police profiling is a common problem, and racial biases exist, it is hypothesized that items that include minorities as the focal character(s) in the situation will elicit higher overexertion of control scores than those items having a non-minority group member as the focal character. In other words, using minority characters will be an effective manipulation for identifying responders who may overexert control.

H3: Average overexertion of control scores across the items that include minority characters will be significantly worse (i.e., higher) than the average overexertion of control score across the items where the focal character is non-minority.

It is further expected that ethnicity of the focal character will also have an affect on the UCS. That is, average underexertion of control scores across items that include non-minority focal characters will be better (i.e., lower) than the average underexertion of control score across the items with minority focal characters.

Gender

Item responses can also be classified according to the gender of the characters most affected by the response. Gender may also be a control eliciting characteristic. Although police profiling is typically a concern for minorities, gender biases do exist. Furthermore, within law enforcement it is a concern that stereotyping of any type will lead to inappropriate exertion of control, especially in the direction of overexertion of control (Mihanovich, 1981). While biases could be present for either female or male characters within the test, it seems more likely that overexertion of control would be directed towards males. Males are physically larger than females and more likely to act out aggressive behavior and may be seen as more threatening. In addition, the vast majority of offenders are male with females representing only about 6% of correctional inmates (Swander & Spurlin, 1998). While there may be bias against women present in the interpretation of the situation it seems less likely that this would induce overexertion of control. Therefore, it will be hypothesized that gender will be an important determinant of overexertion of control, and that some respondents will choose overexertion of control responses when the character is male.

H4: Average overexertion of control scores will be worse (i.e., higher) across items where the focal character is male than where the focal character is female.

It is also expected that this relationship would be inverse for the UCS scale. Female focal characters are less threatening and more likely to influence answers on the VBSJT in the underexertion of control direction. It is expected that the average underexertion of control score will be better (i.e., lower) across items with male focal characters than where the focal character is female.

Likeability

Although these more salient features of the situation are likely to affect exertion of control demonstrated by the test takers, overexertion of control could also be a function of the likeability of the focal characters. That is, purely emotional reaction to the characters in the test may also contribute to an overly controlling response. Behavior of the characters in the VBSJT is specifically manipulated to affect the responses of the test takers. VBSJTs have been demonstrated to elicit more stereotypes towards the characters than the same WSJT (Swander, 2000). That is, stereotypes, or opinions made with no supporting information, were formed about the characters in the VBSJT based on appearance and demeanor demonstrated through the use of dramatic video. Therefore, it is likely that these components will contribute to overexertion of control responding patterns. For example, thinking that "mean" people deserve a more controlling initial approach than "nice" people is a common public relations error. Therefore, it is hypothesized that items with characters viewed as dislikeable will elicit more answers that are overexertion of control than the items with characters that are viewed as neutral or likeable.

H5: Average overexertion of control scores will be worse (i.e., higher) across items with characters viewed as dislikeable than items with characters viewed as neutral or likeable.

Likeable characters in the test could also affect responses in the opposite direction. That is, characters in the test who are viewed as more likeable might elicit responses that are more lenient or less controlling. Therefore, it is expected that the items with likeable characters will elicit more underexertion of control responses than those items where the focal character is neutral or dislikeable.

Differences Between Applicants and Police Officers

Identification of a single construct can also yield new ways to utilize SJTs. Specifically, the OCS can be used to test hypotheses about police officer behavior on the job. It is a concern for law enforcement organizations and the public if police officers are not appropriately exerting control. Overexertion of control by police officers is extremely relevant to modern policing according to the literature (presented below). This is additionally attested to by the intensity of media attention to this aspect of policing. This research could yield a potentially effective way of identifying overexertion of control among police officers. That is, overexertion of control that

is likely to be displayed on the job can be measured with the SJT. Although correlations between experience and SJTs have differed among studies, it is hypothesized that experience will be related to the overexertion of control component, as measured with a SJT. Although the validation report demonstrates that incumbent officers score slightly higher than applicants in terms of total score on the VBSJT used in this study (Swander & Spurlin, 1995), it is hypothesized that the OCS will trend in the opposite direction. The literature demonstrates that officers' personality profiles change, with an increase on aggressiveness scales, as a function of being on the job. The demonstrated conformance and cohesiveness common among law enforcement officers may account for the fact that these changes usually take place within two years (Beutler et al., 1988), a relatively short span of time. Although experience is ordinarily regarded as being associated with improved job performance, the aforementioned documented changes suggest that experience as a police officer can have negative effects (Conroy & Hess, 1992).

Cohesiveness and conformance are characteristic of police organizations. The literature on law enforcement suggests that police officers form highly cohesive groups (Sterling, 1972). Burkhart (1980) discusses reasons for this. First, police officers can be faced with life and death situations where they must be able to count on each other for assistance. Second, with limited financial and promotional rewards, dedication relies on group identification. Third, police officers are often isolated from and feel at odds with the rest of the community. Police officers consistently see the worst side of society which contributes to negative feelings they develop towards the community. Furthermore, they are recipients of injurious intent by the public, contributing to their isolation and reliance on other officers for support and camaraderie. Suspicion and lack of regard towards the public may result from this social alienation (Burhart, 1980; Fortier, 1972). Cohesiveness among police officers leads to an accepted level of behavior that is likely to be upheld by most, if not all, of the officers within an organization (Conroy & Hess, 1992). Inexperienced officers may be subject to internal and external pressures that cause them to join the social ranks of the already established police officers. The strong social climate within police organizations can also cause negative changes to candidates who enter the department without these negative attitudes. (Bem & Allen, 1974; Bowers, 1973). If the standard among officers within the organization is to be overly controlling then it is likely that new officers will change to meet the standards of more senior officers. The need to identify

accepted behaviors is critical for a police organization. Appropriate behavioral responses to situations within the community are critical to maintaining an organization that is affective and supported by the community.

Not only do they conform to demonstrated organizational ethics and form highly cohesive groups, but the literature suggests that police officers do indeed change in terms of personality profiles as a factor of being on the job. Sterling (1972) demonstrated that new police officer personality profiles became more similar to experienced officer profiles after time spent in the position. Beutler et al. (1988) also demonstrated that personality patterns of police officers change over time. They found changes in both a two year and a four year period. Some changes suggested that stress, as manifested by such things as alcoholism, increases as police officers remain on the job. Burgin (1978) summarized the effects of stress experienced by police officers suggesting that stress not only leads to physical harm to the officer but also results in hostility towards the public. Others have demonstrated that attitudes of applicants change after becoming officers (e.g., Banton, 1967). Cynicism was a major attitude that developed after becoming a police officer. This variable has been shown to increase with time spent on the job (Sherrid, 1979). Although applicants may be screened for aggressiveness before they are selected, their values and attitudes may change as a function of being on the job.

It has been demonstrated through use of the MMPI that experience as a police officer can lead to negative changes in personality profile (Beutler et al., 1988), but the research does not directly test the hypothesis that these changes lead to overexertion of control types of behavior. The MMPI is now one of the most commonly reported measures of personality used within law enforcement (Henderson, 1979; Inwald & Shusman, 1984 Shusman, Inwald & Landa, 1984). Police officer behavior is primarily derived from measured variables of personality and other dimensions included within the MMPI. Police are not directly supervised, thus making it impossible to actually observe and study their behaviors on the job. While SJTs do not measure actual behavior, they are simulations that can identify endorsed behavioral solutions to real life situations rather than mental states that are assumed to be associated with those responses.

Overexertion of control among police officers may be a combination of factors such as cynicism, cohesion within the group, or stress response. Although experience may have beneficial properties, overexertion of control appears to be one of the negative results of experience. Overexertion of control is evident throughout law enforcement with common

training components focusing on controlling anger or aggression in anger provoking situations (Abernethy, 1995) and overcoming stress (Conroy and Hess, 1992). Normally the impact of training and experience would be to raise expected performance on a VBSJT. In the current instrument incumbent officers scored two-tenths of a standard deviation higher than applicants on total test score. Nonetheless the impact of enculturation and negative experiences on the job mentioned above should lead experienced police officers to choose more responses that represent overexertion of control answers on the VBSJT. That is, these experience related variables associated with overexertion of control will cause an increase in overexertion of control. It is hypothesized that incumbent police officers will score worse on the OCS than applicants with no previous experience.

H6: Incumbent police officers will score worse (i.e., higher) on the OCS than applicants with no previous experience.

Furthermore, it is likely that cohesion will have a larger impact on police officer normative behavior after a few years of experience. Beutler et al. (1988) demonstrated that significant differences can be seen after two years on the job. Thus, it is also hypothesized that incumbent officers with two or more years of experience will score worse (i.e., higher) than new officers on the OCS.

H7: Incumbent officers with two or more years of experience will score higher on the SJT measure of overexertion of control than new officers.

It is not hypothesized that experience is purely a negative quality. Police officers, on average, score better on this VBSJT than applicants (Swander & Spurlin, 1995). Therefore, experience must be beneficial in making correct choices in the situations presented. This hypothesis would indicate that experience can have a negative relationship with specific constructs within the SJT while also having a positive overall relationship with job performance. This further supports the inconclusive results found in the previous studies and the fact that SJTs are multidimensional. A finding of this type would also support the potential utility of SJTs for organizational research and development, especially if it is possible to better understand the constructs actually measured by the instruments.

Differences Between Organizations

Another way to utilize a construct component of a VBSJT is to identify overexertion of control across organizations. Overexertion of control is likely to vary depending on the

organization. That is, location of the organization is likely to affect the attitudes of the officers. Finding different levels of overexertion of control will help to identify training needs across organizations. Cohesiveness of police officers should contribute to a strong pattern of similar behaviors within an organization. Different behavioral expectations within departments may also lead to different levels of overexertion of control. Beutler et al. (1985) found that location of the police department was responsible for differences among a variety of measures. Specifically, they found that inner-city police officers were more likely to be referred for stress related counseling. Furthermore, they found that college police officers had superior interpersonal skills. These findings may result from the different situations that the officers are faced with in these locations. Inner-city officers are likely to be faced with a wider variety of situations and people than those responsible for college campuses.

Differences in overexertion of control could be a function of the location of the organization (e.g., inner-city vs. rural), organizational climate (some departments take pride in the tough-on-crime image while others strive for a more community service image), or typical personality profile of those on the job, or a combination of factors.

H8: Average levels of overexertion of control will be different depending on the department where the officers are employed, with a higher level of overexertion of control found in departments with a greater volume of felony crime to confront (i.e., inner city urban vs. suburban or rural).

Furthermore, it is likely that overexertion of control within the organization may affect the relationship that overexertion of control has with the performance criteria. That is, above average overexertion of control throughout an organization may suggest that not only is the cohesiveness among officers strong, but also that the normative behavior within the organization is overly aggressive. In effect, behavior that is overexertion of control on the test would be accepted as normal on the part of some departments, lowering the correlation between the test (which may emphasize police norms for control) and supervisory ratings, which in this case may represent an inappropriate standard. Therefore, it is hypothesized that organizations with lower average overexertion of control scores will have stronger negative correlations between overexertion of control and performance than those with higher overexertion of control scores. That is, overexertion of control will be differentially predictive based on the normative expectations within the organization.

H9: Overexertion of control will be differentially predictive based on the level of overexertion of control within the organization, as determined by the OCS.

Again, the UCS will be applied to explore these relationships when possible.

Summary of Hypotheses

H1: The OCS will have significant validity, which will be accounted for by the original VBSJT.

H2: Items with provocation toward overexertion of control will have worse (i.e., higher) scores on the OCS than items without provocation toward overexertion of control.

H3: Average overexertion of control scores across the items that include minority characters will be significantly worse (i.e., higher) than the overexertion of control score across the items where a member of the non-minority group is the focal character.

H4: Average overexertion of control scores will be worse (i.e., higher) across items where the focal character is male than where the focal character is female.

H5: Average overexertion of control scores will be worse (i.e., higher) across items with characters viewed as dislikeable than items with characters viewed as neutral or likeable.

H6: Incumbent police officers will score worse (i.e., higher) on the OCS than applicants with no previous experience.

H7: Incumbent officers with two or more years of experience will score higher on the SJT measure of overexertion of control than new officers.

H8: Average levels of overexertion of control will be different depending on the department where the officers are employed, with a higher level of overexertion of control found in departments with a greater volume of felony crime to confront (i.e., inner city urban vs. suburban or rural).

H9: Overexertion of control will be differentially predictive based on the level of overexertion of control within the organization, as determined by the OCS.

METHOD

Subjects

The VBSJT was rated by ten law enforcement experts on the appropriateness of the exertion of control presented in each of the answer choices. They also rated each situation for the presence of the situational characteristics (see Appendix A). The sample of law enforcement experts consisted of highly experienced law enforcement professionals, with an average tenure of greater than twenty years (see Appendix B for descriptive statistics of the law enforcement experts). The sample consisted of four Commanders, four Sergeants, one Inspector, and one Officer. The group had an average of over nine years in a command position. Thus, the sample was not only highly experienced but possessed the necessary skills and abilities to progress upward in the organization.

The samples of examinees came from validation studies and applicant samples from the years 1995-2000. There are 334 police officers in the original validation sample and 5,426 respondents in the applicant sample. See Appendix B for descriptive statistics of all samples.

Measures

Frontline:™ Video Testing System for Law Enforcement Frontline is a video-based situational judgment multiple choice test developed for entry level law enforcement officers. Frontline was designed to be a comprehensive measure of human interaction skills, responding calmly to provocation, unbiased enforcement, situational judgment, ethics, social maturity, and handling authority. The test consists of 78 scenarios. The VBSJT includes typical, yet critical, situations that are faced by officers on the job. Development of the item content was similar to that of Motowidlo et al. (1990). Subject matter expert panels from eight organizations were used to identify critical incidents for law enforcement officers. These critical incidents were used as a guide to desired behaviors to be included in the test. The scripts for each of the items were then constructed by working closely with the subject matter experts. Once the scripts were developed the video was produced using a professional crew, experienced in video test production. The video was taped in various sites in three jurisdictions. Talent portraying police officers were exclusively police officers. Furthermore, most of the other talent used in the video were police officers. Talent were extensively rehearsed, coached, and directed in their portrayals of the various characters on tape. The test developers and subject matter experts reviewed each

performance for technical quality and adherence to the premise of the item. An expert panel was convened in each of the eight participating organizations to review the test items for relevancy and to weight each possible response. Close to 100 highly experienced law enforcement representatives reviewed each question and each possible response in detail prior to release of the exam.

Performance Evaluation A behaviorally based performance evaluation instrument was designed by Ergometrics specifically for validating this test. The performance appraisal instrument was based on behaviors identified through the job analysis. The critical behaviors were grouped into 13 dimensions. These include: Public Contact Situations – Communication Style, Maturity When Intervening in Stressful Situations, Interrogation/Investigation, Initiative in Handling and Resolving Situations, Officer Safety Orientation, Sensitivity to Diverse Groups, Situational Assessment and Analysis, Relations with Supervisors and Management, Relations with Co-workers, Work Habits, Professional Behavior and Bearing, Paperwork, Persistence in Learning and Keeping Up-to-Date, Physical Skill, Driving Skill, Weapons Skill, and Overall Evaluation. Each dimension was assessed with a seven-point scale. The highest response option was labeled "Outstanding, role model in this area. Extremely strong area. Most positive indicators would apply to this person." The lowest response option was labeled "An area of weakness for this employee. Many of the negative indicators would apply to this employee." Behavioral statements were used to provide specific indicators of good or poor performance relative to each dimension.

Expert Evaluator Rating Form The rating sheet contained explanations of situational variables and instructions for rating items and responses (Appendix A). There were 17 questions for each item taken from the VBSJT. This includes 11 specific provocation items, an overall provocation item, a likeability item and four appropriate exertion of control items, one for each response option. The specific provocations were rated on a Yes, No, or DNA scale. Degree of overall provocation within the item was assessed on a five-point scale: Strong Provocation Toward Underexertion of Control; Provocation Toward Underexertion of Control; No Provocation; Provocation Toward Overexertion of Control; Strong Provocation Toward Overexertion of Control. The likeability item was assessed on a five-point scale: Strongly Dislike, Dislike, Neutral, Like, and Strongly Like. Each of the multiple choice answers for the items were rated on appropriate exertion of control. These response options were rated on a six-

point scale: Extreme Underexertion of Control, Underexertion of control, Appropriate Exertion of Control, Overexertion of Control, Extreme Overexertion of Control, and Does Not Apply.

Procedure

The first step in the research process was to identify items that were related to appropriate exertion of control and drop items that were not related. The items were examined to make sure that content was related to appropriate exertion of control. The 78 item VBSJT was shortened to a 47 item version, with items having been eliminated from the study if their content was not related to appropriate exertion of control. For example, items that dealt with supervisory relations were not included. Ratings from ten law enforcement experts were collected for each of the 47 items. Each rater was given a rating sheet for each item in the shortened version of the test (Appendix A). The raters watched one test item at a time. Following each test item, they rated the 11 specific provocation items on the rating sheet The raters then rated the overall degree of provocation and likeability of the focal character within the item. Each of the multiple choice answers for the item was then rated on appropriate exertion of control. A scoring key was constructed based on the average exertion of control ratings provided by the experts for each response option. The scale was then validated using the original validation sample and comparing the results to those of the original VBSJT. The answer key was then applied to the entire sample of applicants to determine the factors that influence exertion of control.

Analyses

Interrater agreement was computed using Cronbach's alpha (α). Interrater agreement was computed for the appropriate exertion of control ratings across all items. Interrater agreement was further computed for the ratings of appropriate exertion of control on each item and for each situational characteristic item. Pearson product moment correlations within the validation sample were used to partially test Hypothesis 1 (Cohen & Cohen, 1983). Hierarchical regression was also used to partially test Hypothesis 1.

Dependent samples t-tests were used to test Hypotheses 2, 3, 4, and 5. An independent samples t-test was used to test Hypotheses 6, 7, and 8. Hierarchical regression was used to test Hypothesis 9. The validation sample was used to test Hypotheses 1, 6, 7, 8, and 9. The applicant sample was used to test Hypotheses 2, 3, 4, 5, and 6.

RESULTS

As mentioned previously, descriptive statistics for the expert raters, incumbents, and applicants are presented in Appendix A. Descriptive statistics and intercorrelations of the major variables in the study are presented in Tables 1, 2, and 3. The final number of test items included in the study was 43. Four items were dropped before any analyses were computed due to confusion reported by the raters about the focal character to be rated. The ratings of the 43 items were used to create scales based on appropriate exertion of control.

Reliability Analyses

The reliability of the judges was computed to determine if the judges were able to rate appropriate exertion of control and all the other situational variables consistently. As can be seen in Table 4, the average interrater reliability of the ratings of appropriate exertion of control is $\underline{\alpha}$ = .94. It does appear that the raters were able to consistently rate the appropriate exertion of control. The interrater reliability estimates within each item for appropriate exertion of control can also be seen in Table 4. The agreement between the raters is very high, ranging from $\underline{\alpha}$ = .82 to $\underline{\alpha}$ = 1.00. It is concluded that there was sufficient agreement between the raters to use all 43 items in the subsequent analyses. Interrater reliability was also computed for each of the rated situational variables. As can be seen in Table 5, interrater agreement was quite high across all situational characteristic items. Overall provocation had the lowest interrater agreement of $\underline{\alpha}$ = .80. This would be expected because of the complex nature of rating the overall provocation for an item. Percentage of agreement between the raters was also computed as a more conservative measure of agreement for the dichotomous situational characteristics (Cascio, 1998; Table 5). It is concluded that the ratings from the raters can be accurately used to test the hypotheses in this study.

Scale Creation and Scoring

The expert ratings of each item were aggregated to obtain an average exertion of control rating. The six point scale was coded as minus two for Extreme Underexertion of Control, minus one for Underexertion of control, zero for Appropriate Exertion of Control, one for Overexertion of Control, two for Extreme Overexertion of Control, and zero for Does Not Apply. The Does Not Apply option indicated that the answer was unrelated to appropriate exertion of control and

therefore given no weight in the appropriate exertion of control scale. The final scale consisted of 43 items that had answer ratings between minus two and positive two.

The scale was further dichotomized into separate scales based on under- or overexertion of control. Certain concerns made it necessary and more informative to separate the scale. First, there was a scoring issue with regard to having appropriate exertion of control be the mid point. It is possible that test takers could have a zero score by missing items in different directions. Examinees could have the same score and miss different items for different reasons. For example, if a person missed just as many items for overexertion of control as for underexertion of control, that person's score would be close to zero and would therefore appear to be a "good" score. Much of the information would be lost or misinterpreted if scored this way. Second, the VBSJT was designed with more scenarios and answer choices that were related to overexertion of control, therefore the underexertion portion of the scale might be underrepresented. Therefore, overexertion of control and underexertion of control were treated as separate scales. The rest of the analyses will include separate analyses for each scale.

The scales were created separately based on the aggregate ratings of the expert judges. These average ratings from the expert panel were used to score the scales. If less than half of the raters agreed that the answer was an inappropriate exertion of control then the score for that answer was keyed as zero. That is, the scores between 0 and .5 were scored as zero. The rest of the average weights were used because they added meaningful variance to the study. That is, the higher the rated over- or underexertion of control the more weight should be given to that answer choice. Each item included in the underexertion of control scale (UCS) had a scale ranging from 0 to –2 and each item included in the over exertion of control scale (OCS) ranged from 0 to 2.

The final consideration in creating the control scored scales was scoring of the keyed right answers on the exam. As mentioned above, over 100 subject matter experts were used in the design and evaluation of the VBSJT used in this experiment. Therefore, their decision on the right answer must also be considered when scoring the controlling errors in the test. The raters evaluated all choices to an item, including the "right" answer. In the event that answers were rating as showing some level of under- or overexertion of control by the panel but were actually rated as the right answer by the majority of law enforcement professionals, the answer was not used in the controlling scale. Therefore, when discrepancies existed in the right answer, the initial subject matter expert ratings were applied to the scale.

The final scales included 39 items in the overexertion of control scale (OCS) and 18 items in the underexertion of control scale (UCS). Four items in the OCS and 25 items in the UCS were dropped that had no variance in control because there were no answer choices that supported over- or underexertion of control, respectively. Table 1 presents descriptive statistics for each of the scales by sample. Tables 6 and 7 present intercorrelations of the raters' ratings for the OCS and UCS subsets of items, respectively.

Validity of OCS

The OCS was validated using law enforcement incumbent data from the original VBSJT validation. It was hypothesized (H1) that the OCS would have significant validity and would represent a significant portion of the overall correlation between police officer performance and the VBSJT. First, the validity of the OCS was significant ($r = -.234$; $p < .05$; see Table 8). The negative correlation indicates that the greater the overexertion of control score the lower the overall performance evaluation, as was expected. Therefore, the first part of Hypothesis 1 was supported.

The second part of Hypothesis 1 was evaluated using regression. Both measures, the VBSJT and the OCS, were entered into the equation to predict the overall performance rating. If the measure of overexertion of control is not significant when the VBSJT is entered in the equation then the high degree of multicolinearity would suggest that the OCS captures the same variance as the VBSJT, as predicted. As can be seen in Table 9, when both scores were entered into the regression equation the measure of overexertion of control did not add significant variance to the overall performance ($\Delta R^2 = .001$; $p > .05$). Hypothesis 1 was fully supported. Overexertion of control explains a significant amount of the variance in the relationship between the VBSJT and officer performance.

All dimensions of the performance evaluation were also correlated with the OCS. These correlations can be seen in Table 8. The correlations are primarily significant in the expected direction. Further, the correlations appear to be capturing a large portion of the variance in performance of all the expected dimensions. The full support of Hypothesis 1 indicates that an important variable of the VBSJT was identified with overexertion of control.

Validity of UCS

Although the majority of the test is focused on overexertion of control, underexertion of control was also explored in the current study. As can be seen in Table 8, the validity of the UCS was not significant (r = .054; p >.05). That is, this dimension of control was not an important factor in the ratings of performance. However, only 18 of the 43 items had any variance in the underexertion of control direction. This dimension was not considered as important as overexertion of control. Furthermore, the average number of responders to those answer choices was low, indicating that these were not very popular wrong answer choices within the validation sample. This may have contributed to the lack of a correlation along with the limited focus on this dimension in law enforcement. Although this UCS was not related to police officer performance, there were enough items and variance on those items to explore possible provoking stimuli that may be effective for capturing this construct for future reference. Therefore, the analyses with the UCS were computed when possible.

Provocation

The overall provocation hypothesis (H2) tested in this study was that the overexertion of control score across the items that include provoking stimuli toward overexertion of control would be significantly higher than the overexertion of control score across the items that do not include provoking stimuli. Each item was rated on provocation toward overexertion of control to test this hypothesis. The expert raters provided the ratings used for provocation. Each item was rated on a five-point scale ranging from Extreme Provocation for Underexertion of Control and Extreme Provocation for Overexertion of Control (see Appendix A). Average provocation ratings were then computed for each of the 39 items in the overexertion of control scale for the applicant sample (N = 5426). Items were then categorized into three distinct subgroups. If the majority of the raters agreed that there was some provocation then the item was categorized as having provocation toward either underexertion of control or overexertion of control. The rest of the items were categorized as not having any provocation toward overexertion of control. Although this categorization should yield three groups of items, there was only one item that was rated as having provocation toward underexertion of control. Therefore, general provocation toward underexertion of control was not assessed. Items with no provocation were compared with those with provocation toward overexertion of control. There were 22 items in the no

provocation subgroup and 16 items in the provocation subgroup. See Table 10 for descriptive statistics of provocation item subgroups. The sum of the overexertion of control ratings for each provocation group was then divided by the total number of items in each category to control for the unequal number of items in each group. A dependent samples t-test was used to compare the average overexertion of control within the two subsets of items. The mean difference was significant in the hypothesized direction ($\underline{M}_D = .031$; $\underline{p} < .05$; see Table 11). Given the scale used, this difference is not only statistically significant but it is practically significant as well. For example, the average overexertion of control score for all the applicants across all the items is .088 with a standard deviation of .061 (see Table 2). The items that contained provocation toward overexertion of control had, on average, a half standard deviation worse overexertion of control. This is a significant and meaningful. Thus, Hypothesis 2 was supported.

The effects of general provocation toward overexertion of control were also tested on the UCS. There were 10 items in the no provocation subgroup and 7 items in the provocation subgroup. A dependent samples t-test was used to compare the average underexertion of control within the two subsets of items. The mean difference was significant ($\underline{M}_D = -.083$; $\underline{p} < .05$; see Table 11). Thus, the items with provocation toward overexertion of control had significantly better underexertion of control scores than the items with no provocation toward overexertion of control, as might be expected.

Specific Provocation Items

Items related to specific provoking stimuli were also included on the rating form in an attempt to identify specific situational characteristics that effectively contribute to the results found with the overall provocation subgroups. There were no specific hypotheses made about the relationships because it was unclear at the onset of rating collection if there were enough items to conduct the analyses of the specific provocation within the applicant sample. However, based on the overall provocation hypothesis it was expected that the positive situational attributes (e.g., positive character) would lead to more underexertion of control and that negative situational attributes (e.g., rude character) would lead to more overexertion of control. The rating sheet contained 11 items that addressed specific situational characteristics that should contribute to the overall provocation perceived (see Appendix A). Each item was rated on a scale that included Yes, No and No Opinion or DNA (i.e., Does Not Apply). The items were scored to identify if the particular provocation was present. Therefore, the No and No Opinion

categories were both scored as zero and the Yes category scored as one. The ratings of each of the particular items ranged from zero to one. This was dichotomized by the rater agreement. If a majority of the raters agreed that the provocation factor was present then the item was included in the provocation subgroup and if a majority did not perceive provocation then the item was included in the no provocation subgroup. For example, item subgroups containing rude and non-rude focal characters were identified in this way. This method of categorization was conducted for each of the 11 specific provocation categories. This was done for both the OCS and the UCS. Most of the dimensions had enough items to make good comparisons. However, some subgroups had few or no items. Comparisons were computed for the subgroups with two or more items. Although results with this few items may be unstable, they may indicate trends that can be explored through further research.

Rude Focal Character

Ratings of rudeness of the focal character were collected for each of the 39 items. Of the 39 items in the OCS, 10 items had rude focal characters. The other 29 items were rated as not having a rude focal character. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are presented in Table 12. The dependent samples t-test between the subgroups was significant ($\underline{M}_D$ = .04; $\underline{t}$ = 28.1; $\underline{p}$ < .05; see Table 13). The mean difference is small but it is three-quarters standard deviation of the average control score across all items. That is, items with a rude focal character had average overexertion of control scores that were three-quarters standard deviation worse (i.e., higher on the OCS) than those without aggressive focal characters. This finding supports rudeness as a stimulus for overexertion of control.

The 18 items in the UCS included no items that had rude focal characters. Therefore, no analyses were computed.

Aggressive Focal Character

The 39 items in the OCS included 8 items with an aggressive focal character. The other 31 items were rated as not having an aggressive focal character. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are presented in Table 12. The dependent samples t-test between the subgroups was significant ($\underline{M}_D$ = .052; $\underline{t}$ = 31.9; $\underline{p}$ < .05; see Table 13). The mean difference is small but it is over four-fifths standard deviation of the average control score across all items. That is, items that had an aggressive focal character had average overexertion

of control scores that were four-fifths standard deviation worse than those without aggressive focal characters. This finding supports aggressiveness as a stimulus for overexertion of control.

The 18 items in the UCS included only one item that had an aggressive focal character. Therefore, no analyses were computed for the UCS.

Suspicious Focal Character

The 39 items in the OCS included 15 items with a suspicious focal character. The other 24 items were rated as not having a suspicious focal character. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are presented in Table 12. The dependent samples t-test between the subgroups was significant ($\underline{M_D}$ = -.026; $\underline{t}$ = 22.8; $\underline{p}$ < .05; see Table 13). The mean difference is over one-third standard deviation of the average overexertion of control score across all items. That is, the items that had a suspicious focal character had average overexertion of control scores that were over one-third standard deviation better (i.e., lower on the OCS) than those items without suspicious focal characters.

The 18 items in the UCS included 10 items with a suspicious focal character. The other 8 items were rated as not having a suspicious focal character. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are presented in Table 14. The dependent samples t-test between the subgroups was significant ($\underline{M_D}$ = .032; $\underline{t}$ = 16.1; $\underline{p}$ < .05; see Table 15). The mean difference is over one-third standard deviation of the average underexertion of control score across all items. That is, the items that had a suspicious focal character had average underexertion of control scores that were over one-third standard deviation better (i.e., lower or closer to zero on this negative UCS) than those without a suspicious focal character. Again the trend was for suspiciousness of character to yield less errors of underexertion of control.

Pleasant Focal Character

The 39 items in the OCS included 12 items with pleasant focal characters. The other 27 items were rated as not having a pleasant focal character. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are in Table 12. The dependent samples t-test between the subgroups was significant ($\underline{M_D}$ = -.031; $\underline{t}$ =26.1; $\underline{p}$ < .05; see Table 13). The mean difference is a half standard deviation difference of the average overexertion of control score across all items. That is, the items that had a pleasant focal character had average overexertion of control scores that were a half standard deviation better than those without pleasant focal characters.

The 18 items in the UCS included 7 items with pleasant focal characters. The other 11 items were rated as not having a pleasant focal character. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are in Table 14. The dependent samples t-test between the subgroups was significant ($\underline{M_D}$ = -.027; $\underline{t}$ =13.3; $\underline{p}$ < .05; see Table 15). The mean difference is one-third standard deviation difference of the average underexertion of control score across all items. That is, the items that had a pleasant focal character had average underexertion of control scores that were one-third standard deviation worse than those without a pleasant focal character.

Cooperative Focal Character

The 39 items in the OCS included 14 items with cooperative focal characters. The other 25 items were rated as not having a cooperative focal character. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are in Table 12. The dependent samples t-test between the subgroups was significant ($\underline{M_D}$ = -.044; $\underline{t}$ = 36.7; $\underline{p}$ < .05; see Table 13). The mean difference is over two-thirds standard deviation of the average overexertion of control score across all items. That is, items with a cooperative focal character had average overexertion of control scores that were over two-thirds standard deviation better than those without a cooperative focal character.

The 18 items in the UCS included 7 items with cooperative focal characters. The other 11 items were rated as not having a cooperative focal character. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are in Table 14. The dependent samples t-test between the subgroups was significant ($\underline{M_D}$ = -.027; $\underline{t}$ = 13.3; $\underline{p}$ < .05; see Table 15). The mean difference is one-third standard deviation difference of the average underexertion of control score across all items. That is, the items that had a cooperative focal character had average underexertion of control scores that were one-third standard deviation worse than those without cooperative focal characters. However, the results for the underexertion of control scale are the same as for pleasant because the same items were used in each of the subgroups.

Sympathetic Focal Character

Ratings of the sympathy of the focal character were collected for each of the 43 items. The 39 items in the OCS included 8 items with sympathetic focal characters. The other 31 items were rated as not having sympathetic focal characters. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are in Table 12. The dependent samples t-test between the subgroups was significant ($\underline{M_D}$ = -.062; $\underline{t}$ = 50.0; $\underline{p}$ < .05; see Table 13). The mean difference

is one standard deviation of the average overexertion of control score across all items. That is, the items that had a sympathetic focal character had average overexertion of control scores that were one standard deviation better than those without sympathetic focal characters.

The 18 items in the UCS included 4 items with sympathetic focal characters. The other 14 items were rated as not having a sympathetic focal character. The applicant (N = 5426) means and standard deviations for both subgroups are in Table 14. The dependent samples t-test between the subgroups was significant (M_D = .043; t = 20.34; p < .05; see Table 15). The mean difference is half of a standard deviation of the average underexertion of control score across all items. That is, the items that had a sympathetic focal character had average underexertion of control scores that were a half standard deviation better than those without sympathetic focal characters. For both OCS and UCS a sympathetic focal character led to fewer candidate errors.

Contemptible Crime

The 39 items in the OCS included 4 items with a contemptible crime associated with the focal character. The other 35 items did not have a contemptible crime. The applicant (N = 5426) means and standard deviations for both subgroups are in Table 12. The dependent samples t-test between the subgroups was significant (M_D = -.044; t = 27.6; p < .05; see Table 13). The mean difference is over two-thirds standard deviation of the average overexertion of control score across all items. That is, the items that had a contemptible crime had average overexertion of control scores that were over two-thirds standard deviation better than those without a contemptible crime.

The 18 items in the UCS included 3 items with contemptible crime associated with the focal character. The other 15 items did not have a contemptible crime. The applicant (N = 5426) means and standard deviations for both subgroups are in Table 14. The dependent samples t-test between the subgroups was significant (M_D = .008; t = 20.34; p < .05; see Table 15). The mean difference is one-tenth standard deviation of the average underexertion of control score across all items. That is, the items that had a contemptible crime had average underexertion of control scores that were one-tenth standard deviation better than those without contemptible crimes. Surprisingly, for OCS a contemptible crime led to fewer candidate errors, while for UCS a contemptible crime also led to fewer candidate errors.

Complaints About the Focal Character

The 39 items in the OCS included 9 items with a complaint about the focal character. That is the item portrayed that another character complaining about the actions of the focal character. The other 30 items did not have a complaint. The applicant (N = 5426) means and standard deviations for both subgroups are in Table 12. The dependent samples t-test between the subgroups was significant (M_D = -.027; t = 20.4; p < .05; see Table 13). The mean difference is over one-third standard deviation of the average overexertion of control score across all items. That is, the items that had a complaint about the focal character had average overexertion of control scores that were over one-third standard deviation better (i.e., lower) than those without a complaint.

The 18 items in the UCS included two items with a complaint about the focal character. The other 16 items did not have a complaint. The applicant (N = 5426) means and standard deviations for both subgroups are in Table 14. The dependent samples t-test between the subgroups was significant (M_D = -.046; t = 12.2; p < .05; see Table 15). The mean difference is over a half standard deviation of the average underexertion of control score across all items. That is, the items that had a complaint about the focal character had average underexertion of control scores that were over a half standard deviation worse than those without a complaint.

Warrants

The 39 items in the OCS included 5 items where officers had a warrant for arrest of a focal character. The other 34 items did not have a warrant. The applicant (N = 5426) means and standard deviations for both subgroups are in Table 12. The dependent samples t-test between the subgroups was significant (M_D = -.033; t = 17.9; p < .05; see Table 13). The mean difference is over a half standard deviation of the average overexertion of control score across all items. That is, the items that had a warrant for the focal character had average overexertion of control scores that were over a half standard deviation better than those without a warrant.

The 18 items in the UCS included 2 items with a warrant for arrest for focal character. The other 16 items did not have a warrant. The applicant (N = 5426) means and standard deviations for both subgroups are in Table 14. The dependent samples t-test between the subgroups was significant (M_D = -.074; t = 21.4; p < .05; see Table 15). The mean difference is over three-quarters standard deviation of the average underexertion of control score across all items. That is, the items that had a warrant for the focal character had average underexertion of

control scores that were over three-quarters standard deviation worse than those without a warrant.

Broken Laws

The 39 items in the OCS included 19 items where the focal character had broken a law. The other 20 item there was not a law broken. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are in Table 12. The dependent samples t-test between the subgroups was significant ($\underline{M}_D$ = -.049; $\underline{t}$ = 41.1; $\underline{p}$ < .05; see Table 13). The mean difference is over three-quarters standard deviation of the average overexertion of control score across all items. That is, the items that had a law broken had average overexertion of control scores that were over three-quarters standard deviation better than those without a law broken.

The 18 items in the UCS included 9 items where the focal character had broken a law. In the other 9 items there was not a law broken. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are in Table 14. The dependent samples t-test between the subgroups was significant ($\underline{M}_D$ = -.055; $\underline{t}$ = 228.4; $\underline{p}$ < .05; see Table 15). The mean difference is two-thirds standard deviation of the average underexertion of control score across all items. That is, the items that had a law broken had average underexertion of control scores that were two-thirds standard deviation worse than those without a law broken.

Potential Involvement

The 39 items in the OCS included 24 items where there was potential for others to become involved in the situation. In the other 15 items there was not a threat of others becoming involved. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are in Table 12. The dependent samples t-test between the subgroups was significant ($\underline{M}_D$ = -.007; $\underline{t}$ = 6.219; $\underline{p}$ < .05; see Table 13). The mean difference is over one-tenth standard deviation of the average overexertion of control score across all items. That is, items where there was potential for others to become involved had average overexertion of control scores that were over one-tenth standard deviation worse than those items where there wasn't a threat of others becoming involved.

The 18 items in the UCS included 9 items where there was potential for others to become involved. In the other 9 items there was not potential for others to become involved. The applicant ($\underline{N}$ = 5426) means and standard deviations for both subgroups are in Table 14. The dependent samples t-test between the subgroups was significant ($\underline{M}_D$ = -.060; $\underline{t}$ = 30.1; $\underline{p}$ < .05;

see Table 15). The mean difference is over two-thirds standard deviation of the average underexertion of control score across all items. That is, the items where there was potential for others to become involved had average underexertion of control scores that were over two-thirds standard deviation worse than those items where there wasn't a threat of others becoming involved.

Ethnicity

Ethnicity was also hypothesized (H3) to be an important situational characteristic that may help to explain variance in applicant responses on the OCS. It was hypothesized that questions with minority focal characters would elicit more overexertion of control than those questions with non-minority focal characters. Items were again categorized into subgroups according to the ethnicity of the focal character. The expert raters did not rate the items on this dimension due to possible contamination and hypothesis guessing. The first comparison that was made was between the items that had a minority focal character with those items that had a non-minority focal character on the OCS with the applicant data. There were 13 items in the minority subgroup and 21 items in the non-minority subgroup. Some items were not used in the analysis because there were more than one focal character and the ethnicity of the focal characters was mixed. The descriptive statistics for these subgroups can be seen in Table 16. The difference between the means of the two subgroups was statistically significant (M_D = -.018; t = -14.6; p < .05; see Table 17). However, the difference was not in the hypothesized direction. The non-minority subset of items had worse overexertion of control scores than the subset including the minority items. Therefore, Hypothesis 3 was not supported.

The same comparisons were also made to see if ethnicity affected underexertion of control. Therefore, groups were created the same as for the overexertion of control but the number of items in each subgroup was different. There were 10 items with non-minority focal characters and 7 items with minority focal characters. Means, standard deviations and other descriptive statistics can be seen in Table 16. The mean difference between the non-minority and minority subgroups was statistically significant (M_D = -.034; t = -60.105; p < .001; see Table 17). This is again in the opposite direction of what was expected. The items with minorities as the focal character had worse underexertion of control scores than those items with non-minority focal characters.

The same analysis was also computed for items with African American focal characters compared with those with Caucasian focal characters. This analysis was suggested by the national focus on racial profiling toward African Americans. Furthermore, the majority of the minority items had African American focal characters. There were large enough samples to compare a specific minority group as focal characters and further compare the scores across applicant subpopulations based on ethnic group status. Therefore, a dependent samples t-test was also used to specifically compare the items with African American focal characters against those with Caucasian focal characters. On the OCS 11 of the 13 items with minority focal characters had African American focal characters. The means can be seen in Table 16. The mean difference is statistically significant ($\underline{M}_D$ = -.034; $\underline{t}$ = -26.08; $\underline{p}$ < .05; see Table 17). Again, the mean difference is in the opposite direction as expected from the literature. It also appears that the difference is stronger between the African American and Caucasian items than when the other minority groups are included with the African American focal character items.

The mean difference between the Caucasian and African American subgroups on the UCS was also significant ($\underline{M}_D$ = -.119; $\underline{t}$ = -46.85; $\underline{p}$ < .001; see Table 17). The mean differences between the groups were also in the same direction as the overexertion of control scored items. That is, the items with African American or other minority groups as the focal character had on average worse underexertion of control scores than those with Caucasian focal characters. The results indicate that minority focal characters tend to elicit less overexertion of control and more underexertion of control than when the focal character is Caucasian.

Ethnicity Results Within Subgroups

Exploratory research was also conducted within the applicant sample to determine interaction with the racial diversity in the test. Table 18 summarizes the descriptive statistics for the African American and Caucasian samples independently. As can be seen, there are larger overexertion scores and underexertion scores for African American responders on all subsets of ethnicity items. Furthermore, the mean differences are larger for the African American sample than the Caucasians sample. That is, African American responders tended to have larger increases in overexertion of control within the Caucasian subset of items than the Caucasian responders. The African American responders also tended toward greater increases in underexertion of control for the African American subset of items. Therefore, the ethnicity of the focal character appeared to have larger effect for African American applicants than it did for

Caucasian applicants on the OCS and UCS. However, the differences were significant in the same direction for both groups (see Table 19).

Provocation Confound

Provocation was explored to determine if the different results found between minority and non-minority focal character items were due to different levels of overall provocation for those items. As can be seen in Table 6, overall provocation was correlated with the minority/non-minority focal character distinction ($\underline{r} = .24$). This correlation suggests that the items with minority focal characters were rated as having more provocation toward overexertion of control. Controlling for provocation would only strengthen the findings that the items with minority focal characters had better overexertion of control scores. The correlation between overall provocation and the minority/non-minority focal character distinction was also computed for the UCS. The correlation ($\underline{r} = .46$) also suggests that controlling for provocation would further separate the two groups in the direction found.

Gender

Gender was also a situational variable that was hypothesized to have an affect on the exertion of control used. Specifically, it was hypothesized (H4) that items with male focal characters would have higher overexertion scores, on average, than items with female focal characters. Items were categorized into subsets based on the gender of the focal character. There were a total of 25 items with a male focal character and 8 items with a female focal character. Descriptive statistics of the two subsets of items in the applicant sample can be seen in Table 20. Although the difference between means is small, it is statistically significant ($\underline{M}_D = -.004$; $\underline{t} = -2.86$; $\underline{p} < .05$; see Table 21). However, the difference is opposite the direction hypothesized. That is, on average, items with females as the focal character had slightly worse overexertion of control scores than those items with males as the focal character. Therefore, Hypothesis 4 was not supported.

Gender was also explored to see if it had an impact on the underexertion of control. The same subset of items was used as for overexertion of control but many of the items were removed due to no possible underexertion of control answer choice. There were 16 items with male focal characters and 2 items with female focal characters. As can be seen in Table 21, there is a large difference between the means of the items ($\underline{M}_D = -.117$). The difference is statistically significant ($\underline{t} = 87.3$; $\underline{p} < .05$). The results of the two analyses indicate that the items with female

focal characters were less likely to elicit underexertion of control reposes and more likely to elicit overexertion of control. However, as mentioned previously, an aggregate of only two items is very unstable. Including more items in this subgroup could significantly change the outcome of this portion of the study.

Gender Differences Within Subgroups

Exploratory research was also conducted within gender item subgroups to determine if focal character gender had any interaction with the gender of the test taker. Table 22 summarizes descriptive statistics for the male and female samples independently. While there were slight differences between subgroups, the differences did not constitute anything that was meaningful and were probably due to chance. The differences between groups were less than .01 on both the OCS and UCS (see Table 23). Therefore, it is determined that gender of the test taker did not influence exertion of control applied in gender specific items within the test.

Provocation Confound

Provocation was explored to determine if differences found between female and male focal character subsets were due to different levels of overall provocation for those subsets. As can be seen in Table 6, overall provocation with the OCS items was correlated with the female/male focal character distinction ($\underline{r} = -.41$). This correlation suggests that the items with male focal characters had more provocation toward overexertion of control. Controlling for provocation would only strengthen the findings that items with female focal characters had worse (i.e., higher) overexertion of control scores than the items with male focal characters. Correlation between overall provocation and the gender focal character distinction was also computed for the UCS. The correlation ($\underline{r} = -.40$) also suggests that controlling for provocation would further separate the two groups in the direction found.

Likeability

Likeability of focal characters within the items was also treated similarly to overall provocation for the OCS. Ratings of likeability were used to categorize items into three subsets. The first subset consisted of 6 items where the focal characters were viewed as likeable by the expert panel. The second subset contained 12 items that were viewed as having neutral characters. The third subset consisted of 21 items where the focal character was viewed as being dislikeable. First, a dependent samples t-test was used to test the hypothesis (H5) that the items with dislikeable characters would elicit more overexertion of control than the items with neutral

or likeable characters. The descriptive statistics for each subgroup of items in the applicant sample are presented in Table 24. The mean difference between the items with dislikeable versus neutral characters was significant (M_D = -.003; t = -2.02; p < .05; see Table 25). The mean difference between the items with likeable characters compared to those with dislikeable characters was significant (M_D = .003; t = 1.98; p < .05; see Table 25). Although both results are significant, the results are contradictory. Furthermore, the difference between items with likeable versus neutral focal characters was significant. However, likeable characters elicited worse overexertion of control scores. Although results of these t-tests are significant, the differences represent a small effect. Contradictory results coupled with limited practical significance suggest that Hypothesis 5 was not supported and that likeability of the characters does not have an effect on overexertion of control.

Results were also computed for items that were scored toward underexertion of control. There were three items in the like subset, six in the neutral subset, and nine in the dislike subset. Descriptive statistics are presented in Table 24. Again a series of dependent samples t-tests were used to assess the differences between average exertion of control scores. As can be seen in Table 25, results appear to be minimally significant in the opposite direction expected. Candidates made slightly more errors on items with dislikable and neutral characters than those items with likeable characters.

Effects of Tenure

Based on the law enforcement literature it was hypothesized that group cohesion and cynicism stemming from the nature of police work would affect the overexertion of control of police officers. It was specifically hypothesized (H6) that the experience would lead to more overexertion of control. The hypothesis was tested by comparing the mean scores on the OCS for the applicants against the validation sample. Table 1 summarizes the descriptive statistics of the different samples. Contrary to what the literature may suggest, the experienced police officers scored significantly better on the OCS than the applicants (M_D = .618; t = -4.62; p < .05; see Table 26). Therefore, Hypothesis 6 was not supported. Furthermore applicants were also significantly different on the UCS (M_D = .641; t = 7.7; p < .05; see Table 26). That is, police officers were also less likely to make underexertion of control errors (see Table 1). These results correspond with the finding that the experienced officers do better on the test as a whole.

Hypothesis 7 stated that officers with greater than two years of experience would have higher scores on the OCS that the officers with less than two years experience. Using the sample incumbent officers ($N = 334$), the descriptive statistics are presented in Table 27. Hypothesis 7 was not supported. Overexertion of control scores for officers with less than two years experience were not significantly different from those with greater than two years experience ($M_D = .245$; $t = 1.17$; $p > .05$; see Table 28). Therefore, both hypotheses (H6 and H7) about tenure were not supported. Differences in officer tenure were also tested on the UCS. The difference between the average scores on the UCS was significant ($M_D = .364$; $t = 2.40$; $p < .05$; see Table 28). The direction of the difference indicates that the officers with greater than two years experience were more likely to make errors in terms of underexertion of control.

Differences Between Organizations

Validation organizations were grouped based on location to test the hypothesis (H8) that the police officers who face more felony crime will have higher overexertion of control scores. Therefore, using the sample of incumbents ($N = 334$), organizations were grouped by whether they were suburban or urban. The validation sample included three urban organizations ($N = 138$) and four suburban/rural organizations ($N = 196$). The descriptive statistics for each group are presented in Table 29. The hypothesis was tested using the OCS. An independent samples t-test was used to test the difference between means on overexertion of control. The mean difference between the two groups is statistically significant ($M_D = .69$; $t = 3.308$; $p < .05$; see Table 30). However, the mean difference is opposite the direction hypothesized. That is, police officers in the suburban/rural locations score, on average, score worse on the OCS than the police officers in the urban locations. Therefore, Hypothesis 8 was not supported.

The difference between organizations was also explored using the UCS. The difference between locations on underexertion of control is not significant ($M_D = .062$; $t = .485$; $p > .05$; see Table 30). It is concluded that there is no difference between the groups on underexertion of control.

It was also hypothesized that the difference in provocation across locations would be related to different validities for the different groups (H9). That is, that the interaction between location of the organization and scores on the OCS would be significantly related to overall ratings of performance. Although Hypothesis 8 was not supported the difference was significant in the opposite direction. Table 29 indicates that there were systematic differences in

overexertion of control based on location. Therefore, Hypothesis 9 was tested to determine if overexertion of control is higher in the suburban locations because focus on it as an important component of performance is not as great as in the urban organizations.

The validity coefficients are presented in Table 31. The difference in validity coefficients can be seen in Table 32. Hierarchical regression analysis was used to determine if location of the organization and overexertion of control score interaction was significant. First, location of the organization (dummy coded; Urban = 1; Suburban = 2) and overexertion of control score were entered into the regression equation. As can be seen in Table 32, only the regression coefficient for the OCS was significant and both variables accounted for a total of 5.7% ($\underline{R}^2 = .057$; $\underline{p} < .05$) of the variance in overall job performance. The next step was to enter location of the organization by overexertion of control score interaction. This was computed by multiplying the dummy coded location of the organization and overexertion of control score. When this variable was entered into the regression equation the total variance accounted for ($\underline{R}^2$) increased by .7%. An $\underline{F}$ test for the change in the variance accounted for revealed that the increase was nonsignificant ($\underline{\Delta R}^2 = .007$; $\underline{\Delta df} = 1$; $\underline{F} = 2.449$; $\underline{p} > .05$). Therefore, Hypothesis 9 was not supported. There is no differential validity based on the average overexertion of control score within an organization.

DISCUSSION

The present study sought to explore response patterns of a VBSJT by identifying an integral construct and situational characteristics that contribute to the effectiveness of a VBSJT designed for the selection of entry level law enforcement officers. The VBSJT was scored to identify the construct of appropriate exertion of control. This scale was used to identify the test item characteristics that successfully impact whether a candidate chooses an appropriate exertion of control response. In addition, the scale was applied to real samples of police officers to demonstrate useful ways to utilize these tests for diagnostic purposes. The following discussion presents the main findings and conclusions of the study. Limitations and further research are discussed.

Appropriate Exertion of Control

One of the primary goals of this paper was to identify a single construct within a multidimensional VBSJT. Although appropriate exertion of control is not a typical construct (e.g., conscientiousness), it is a more singular, unidimensional conceptual variable than the VBSJT as a whole, and it fits quite well in situations simulating police decision making. Appropriate exertion of control was broken into both overexertion of control and underexertion of control because it is unclear whether to treat these as a linear scale and the VBSJT was clearly more focused on overexertion of control (Swander & Spurlin, 1995). While these two components were found within the VBSJT, the overexertion of control component was represented with more items and possible answer choices. Specific hypotheses were tested about the overexertion of control scale but the underexertion of control scale was included for purely exploratory reasons. Thus, specific hypothesis testing based on the results of this experiment may be warranted for a VBSJT that includes underexertion of control as a larger portion of the scale.

The importance of the overexertion of control component of the VBSJT was demonstrated by the significant validity coefficient obtained for the overexertion of control scale (OCS; $\underline{r} = -.234$). This correlation proved to be a large portion of the correlation between the VBSJT and job performance ($\underline{r} = .332$). The results support the hypothesis that overexertion of control is an important component within the field of law enforcement and that overexertion of control is an integral part of the VBSJT as a whole. One important component of the

multidimensional VBSJT was effectively captured utilizing expert subject matter ratings. Multidimensionality at the item level is clearly an important consideration that must be made when studying the underlying constructs of a SJT.

The construct of underexertion of control was also parceled out from the VBSJT. The validity of the underexertion of control (UCS) scale was near zero. The limited number of items that represented this construct (18) may account for the near zero correlation. Although underexertion of control was not viewed as an important component, the items that represented this construct were included in all possible analyses for exploration of the effects of situational characteristics on this construct. Results provided for this scale may not be important for law enforcement but may provide useful information for the development of SJTs for other jobs.

Capturing these constructs measured by the VBSJT will help provide understanding of test functioning as a whole, insight about inconclusive results, and a better framework for studying SJTs. Although this test was created specifically for entry level law enforcement selection, appropriate exertion of control could very readily generalize to any job that requires human interaction or customer relations. The results summarized below should generalize to VBSJT and SJTs design for other jobs that require extensive human interaction skills and in particular jobs where incumbents have discretion in their exertion of control over others.

Situational Characteristics

The identification of a unidimensional construct within the VBSJT and the answers that correspond to that construct made it possible to identify important situational characteristics of a VBSJT. The results from both the OCS and UCS provide information that will be useful in the future to appropriately manipulate these constructs within VBSJTs. The results also indicate that development of situational specifics in VBSJTs are critical to item responses. Further, the results also indicate the usefulness of VBSJTs in capturing a more realistic job preview of the applicants' performance.

Situational characteristics are inherently built into all types of SJT items. That is, the situation and the characters are described in an attempt to capture the situation and give examinees enough information to make appropriate judgments. However, there are many elements included within this study that pertain only to VBSJTs (e.g., rude or pleasant characters). Character descriptions are often not included, or limited to names, within a WSJT. Therefore, many of the results actually apply to the debate about the different characteristics of

WSJTs versus VBSJTs (see Chan & Schmitt, 1997). Not only can the results of this study be directly applied to the development of future VBSJTs but they also demonstrate effective situational characteristics presented in video that could not realistically be captured through traditional paper and pencil tests.

Overexertion of Control

The present study explored many different situational components in an attempt to identify the stimuli that contribute to the measurement of overexertion of control. Overall provocation represented within the situations was assessed to determine if the situational characteristics aimed at eliciting overexertion of control had an effect. Overall provocation was hypothesized to influence the responses of the examinees in the direction of choosing more controlling behaviors. Overall provocation was considered to be any stimuli that would potentially elicit an overexertion of control response from a test taker. These included such things as rude and aggressive characters, warrants, and laws being broken. The results demonstrated that higher overall provocation toward overexertion of control works to effectively elicit higher overexertion of control scores. Therefore, it is possible to manipulate situational variables to elicit responses that may not otherwise be chosen by test taker. This is a major presumption regarding SJTs that has not been proven until now. That is, it has been assumed that the situational context of the item significantly contributes to the effectiveness of the test (e.g., Motowidlo, Dunnette and Carter, 1990).

Although proving the presumption was an important step, identifying specific manipulations that were more or less effective for producing overexertion of control responses could lead to greater understanding and be directly applied to strengthen developmental efforts. These manipulations could be directly applied to other tests that are developed to measure overexertion of control (e.g., corrections, supervision, customer service, or any other human interaction instrument). Therefore, the specific manipulations that were to be included in the scale used to obtain overall ratings of provocation were also evaluated independently. The specific provocation items were broken into eleven different characteristics, but ten of the eleven specific provocation items can be categorized into two types of provocation: character manipulation and situational information manipulation.

Character dimensions manipulated by the use of dramatic video, included rude, aggressive, sympathetic, suspicious, pleasant and cooperative. It was expected that the more

negative characteristics (rude, aggressive, suspicious) would lead to worse overexertion of control scores. Rude and aggressive characters were particularly effective in eliciting overexertion of control responses. Each of these provocations had significantly worse overexertion of control scores across items than the items without rude and aggressive focal characters.

However, suspicion did not have the expected effect. That is, characters who were not suspicious were associated with more overexertion of control than those who were suspicious. The intercorrelations between suspicion and the other rated dimensions of provocation were explored for possible explanations (see Table 6). Correlations among items with suspicious characters were primarily in the expected direction. That is, characters viewed as suspicious were correlated with characters viewed as aggressive, not pleasant, not cooperative, and not sympathetic. There do not appear to be any confounding variables. An alternate explanation may come from the differences between the expert raters and the applicant sample. The expert raters were highly experienced law enforcement professionals and therefore may have seen more behaviors as suspicious, whereas the applicants may have been less attuned to suspicious behavior. Also, and perhaps more importantly, the expert raters were specifically asked to evaluate the items for suspiciousness; applicants simply viewed the items without having their attention directed at the absence or presence of suspiciousness. These explanations were partially supported by the quantity of items that were rated as having suspicious focal characters. The expert raters viewed 15 items as having suspicious focal characters, whereas in the original test development there were not 15 items where there was any scripted attempt to make the focal character suspicious. Due to effect of the experiment, the expert raters may have employed a lower threshold for suspiciousness than an unprompted viewer. Of course it could still be true that suspiciousness is not a 'hot button' affecting overexertion of control like rude and aggressive behaviors.

All of the specific provocations that were related to the characters and emotional responses were effective manipulations in the expected direction. Although it was deemed most likely that positive character attributes (i.e., pleasant, cooperative, sympathetic) would increase underexertion of control, it turned out that these items also had significantly better (i.e., lower) overexertion of control scores. These manipulations are useful and should be considered when designing a VBSJT.

As mentioned above, the character manipulations were only possible or at least realistically possible with the use of dramatic video (Jones & Decotiis, 1986; Weekley & Jones, 1997). The strong positive results from these situational characteristics provide evidence of the utility of using dramatic video to present situations. These findings also demonstrate the impact that one person's behavior can have on another person. This is a commonly overlooked component in testing as demonstrated by the relatively little use of video as compared to paper and pencil tests.

The second type of provocation evaluated was the information contained within the situation. This information within the item was related to overexertion of control. For example, warrants, laws being broken, complaints, and contemptible crimes were used to enhance the scenarios. Although it was expected that these situational characteristics would lead to more overexertion of control, the mean differences were in the opposite direction. Complaints, warrants, laws broken, and contemptible crimes were all associated with better overexertion of control scores. Although the results are different than expected they provide insight on how item characteristics affect item responses. The results warranted further exploration.

First, it is important to note that items with manipulations of situational information (e.g., warrants or complaints) actually had less overexertion of control than items that did not include these stimuli. Correlations between all of the rated variables within the study were computed to explore any systematic combinations of stimuli that may have affected the results (see Table 6). The intercorrelations of items in the OCS suggest that there were patterns of combined situational characteristics. For example, having a warrant was correlated $\underline{r} = .41$ and $\underline{r} = .35$ with characters that were pleasant and cooperative, respectively. This indicates that the warrants were commonly coupled with pleasant characters. However, it appears that the warrants were not as effective situational characteristics as the demeanor or emotional response of the character in the test. The test takers may have based their exertion of control on the pleasantness of the character and not the fact that they were wanted for having committed a crime. In this situation it appears that emotional reactions to dramatic video highly outweigh the physical, concrete evidence presented within the situation.

Other information introduced in the situation that affected item responses in the opposite direction hypothesized were complaints made by others in the situation and crimes that were of a contemptible nature. The intercorrelations among the variables suggest that these variables

should have produced results in the expected direction. That is, these characters were generally rated as having less positive characteristics and more negative characteristics. A simple explanation could be that these manipulations were too obvious. Although true to the job of a police officer, test takers may have assumed that the information included in the item was an attempt to get them to overreact and therefore they did not endorse answers that represented overexertion of control.

The final characteristic explored on the OCS was the potential for others to get involved. This situational characteristic significantly affected the response patterns of the applicants. However, the mean difference between the groups was very small and not in the expected direction. As with the suspicious focal characters, there was most likely a perceptual difference between experienced law enforcement professionals directed to pay attention to this variable and the applicants. The expert panel rated almost all items that had other people in the situation as potential for others to become involved. Although it is possible that this could happen, the potential for others to become involved was designed to be manipulated within the test by adding characters in the situations who showed emotional involvement with the focal character or situation, such as a family member.

Results from the OCS are very informative. The emotional character manipulations appeared to be the most effective manipulations for eliciting overexertion of control. These dimensions appeared to outweigh some of the more concrete situational characteristics. Furthermore, it appears that some of the concrete stimuli may actually elicit responses in the opposite direction expected. Although many of these situational characteristics add to the realness of the video, the use of such stimuli should be considered. For example, the results from the overall provocation were not as strong as the results from the character manipulations, suggesting that the other informative situational characteristics "washed out" some of the effect of the character manipulations. Therefore, combinations of stimuli must be considered. Overall, the information provided for the provocation of overexertion of control is very favorable. Situational characteristics do have an impact on the answers chosen by the applicants and the characteristics that have the most impact are those that cannot be well simulated by a WSJT. In a WSJT emotional manipulations must be identified and labeled, making them more obvious and this research demonstrates that test takers may respond in the opposite manner as anticipated when manipulations are obvious.

Underexertion of Control

Overall provocation toward overexertion of control also had significant effects on the answer choice of the underexertion of control responses. That is, the more the overall provocation toward overexertion of control, the less underexertion of control that was demonstrated. Therefore, it appears that the same provocation that elicits overexertion of control also has the affect of moderating underexertion of control in a direction that is more appropriate to the situation. Although only one item was rated as having provocation toward underexertion of control, the items with less provocation toward overexertion of control had worse underexertion of control scores. The specific provocation subgroups were also explored within the UCS. Similar to the OCS, the character manipulations of pleasantness, cooperativeness, and sympathy all had higher underexertion of control scores, as expected. Aggression and rudeness as provocation were not represented within the UCS subgroups.

Contradictory to the OCS, sympathetic focal characters elicited better scores on the UCS. Thus, the results appear to indicate that sympathetic focal characters tend to lead responders into answering correctly or with the appropriate exertion of control. That is, items with sympathetic focal characters had better scores for both over- and underexertion of control. The same phenomenon was true for the items with suspicious focal characters.

The results indicated that suspicious characters elicited less extreme overexertion of control responses. As mentioned above, this may be a result of the perceptual differences between the raters and the applicants who were not instructed to focus on suspiciousness. However, based on the results, it appears that items with characters who were more suspicious led people to making the right choice. That is, they were less likely to make a mistake in either direction of exertion of control. This may have been a situational characteristic that inhibited mistakes in exertion of control and led applicants to making the right choices. In other words, sympathetic or suspicious characters provide good clues about the best way to handle the situation and these factors were not as strong of distracters as pleasant or aggressive behavior.

Crimes of a contemptible nature, complaints, warrants, and laws being broken all lead to more extreme underexertion of control scores. This is the same pattern that was found for the OCS where these information variables were also associated with less extreme overexertion of control scores. The intercorrelations within the items in the UCS (see Table 7) also suggest that warrants and broken laws were coupled with pleasant, cooperative and nonaggressive focal

characters. Again, it appears that information variables make the questions easier but the emotional characteristics outweigh the information situational characteristics. Within the UCS item subset, the expert rating intercorrelations do not explain the effect that crimes of a contemptible nature and complaints have on applicant responses. Although these may be effective manipulations, they appear to be more transparent within the test items and work opposite to what would be expected. These situational characteristics need to be carefully considered when developing a VBSJT.

Ethnicity

The hypothesis that a minority focal character would lead to more extreme overexertion of control scores was not supported. In fact, the results suggest that the pattern of responses is in the opposite direction. That is, items with minority focal characters had less extreme overexertion of control scores than those with non-minority focal characters. Intercorrelations were explored for possible explanations. Rater intercorrelations suggest that the items with minorities as the focal characters were more provoking in general, ruder, more suspicious, less cooperative and less sympathetic. Although the provocation should lead to more extreme overexertion of control, the ethnicity of the focal characters appeared to have an affect. Controlling for overall provocation would actually lead to larger differences in the direction found. The same pattern was found for the UCS. That is, items with minority focal characters also had more extreme underexertion of control scores than those items with non-minority focal characters.

A possible explanation is that the test takers were sensitive to the ethnicity of the focal characters. With increased exposure on the part of the general public to problems with bias and racial profiling (e.g., Drummond, 1999; Newport, 1999), it is possible that test takers consider the ethnicity of the focal characters and are cautious about overexerting control or respond by underexerting control because of this awareness.

Further exploration was made between the items with African American focal characters and the items with Caucasian focal characters. The results were also opposite to that expected. That is, less overexertion of control and more underexertion of control for the items with African American focal characters than those with Caucasian focal characters. The magnitude of difference was even larger between these groups than the comparison of Caucasian with the total minority group.

The difference between Caucasian and African American applicants was also explored for differences that might account for the results found. There were more extreme overexertion scores and underexertion scores for African American responders on all subsets of ethnicity items. Furthermore, the mean differences were larger for the African American sample than the Caucasian sample. That is, African American responders tended to have more extreme overexertion of control within the Caucasian subset of items than Caucasian test takers. The African American responders also tended toward more extreme underexertion of control for the African American subset of items. Therefore, the ethnicity of the focal character appeared to have larger effect for African Americans than it did for Caucasians on the OCS and UCS.

Gender

The hypothesis that male focal characters would elicit more extreme overexertion of control scores than female characters was not supported. In fact, there was a small but significant difference in the opposite direction. That is, the items with female focal characters elicited more extreme overexertion scores than the items with male characters. The practical significance of this statistical difference is small. However, the intercorrelations of the expert raters ratings suggest that controlling for overall provocation would enhance these results. Therefore, if a difference exists, it is in the opposite direction hypothesized. Although from this study the magnitude of the effect of the gender of the focal character is unclear, gender differences are a reality and diversity makes VBSJTs more realistic.

Only two items were used in the underexertion of control scale that had female focal characters. The small number of items on the UCS with female focal characters limits the conclusions that can be made on this portion of the scale. Exploratory research was also conducted within the gender item subgroups to determine if this had any interaction with the gender of the test taker. While there were slight differences between subgroups, the differences did not constitute anything that was meaningful. Therefore, it is determined that gender of the test taker did not influence exertion of control applied to gender specific items within the test.

Likeability

The likeability hypothesis was not supported in the study. Characters rated as either likeable or dislikeable elicited more extreme overexertion of control than the focal characters rated as neutral. Furthermore, items with likeable characters had more extreme overexertion of

control than those with dislikeable characters. There were also significant differences between items with likeable characters and items with neutral or dislikeable characters on the underexertion of control scale. Items with likeable characters had worse underexertion of control scores than the rest of the groups. Basically, test takers had more difficulty responding appropriately to likable characters than to neutral or dislikeable characters.

There are few considerations that must be made when interpreting the results of the likeability of the focal characters. First, the differences that were significant were small. Given the large sample, even small differences were significant. Second, the likeable ratings were highly correlated with general and specific provocations, most of which correlated in the expected direction for both the OCS and UCS items (see Tables 6 and 7). The OCS and UCS scale items had correlations between general provocation and likeability of $\underline{r}$ = -.61. and $\underline{r}$ = -.40, respectively. Furthermore, there were strong negative correlations between likeability and rudeness and aggressiveness and strong positive correlations between likeability and pleasantness, cooperativeness and sympathy, within the UCS and OCS item subsets.

Although likeability was correlated significantly with the specific provocation in the expected direction the effect was minimal. Correlation with the informational situational characteristics was also in the expected direction, which may partially explain why the differences were not very large. That is, the ratings of likeability were based on the combined effect of all the information, including feelings of dislike toward those that were associated with contemptible crimes or had complaints made against them. This would have counterbalanced the effect of the ratings of liking. However, ratings of likeability could also be highly subjective and idiosyncratic and not translate to generalizations across people.

Applications of OCS and UCS

The comparison of results between applicants and incumbents was a successful demonstration of the usefulness of a VBSJT. The literature suggests that police officers increase in aggression and cynicism toward citizens as a result of experience on the job (e.g., Beutler et al., 1985). Furthermore, the literature suggests that police officers who work in violent, high felony crime areas experience more stress. Therefore, there are two main factors that may contribute to police officer aggression and cynicism: experience and criminal interaction. However, the research presented in this paper goes against the traditional research literature. Experienced police officers were found to score better on the OCS and the UCS than applicants.

That is, they were less likely to overexert control or underexert control. Furthermore, police officers with greater than two years experience had lower overexertion of control scores and even had higher underexertion of control scores. This is directly in contradiction to the law enforcement literature that has demonstrated increased aggressiveness and stress scores on related dimensions of the MMPI (Banton, 1967; Beutler et al., 1998). The results are more consistent with police training models that emphasize using the minimum control necessary in a situation and to start interactions with citizens in a friendly manner.

This research is merely a demonstration of other ways of using a VBSJT. Police training and counseling, including in such areas as profiling, would benefit from a diagnostic VBSJT. Although it appears that these hypotheses were similar to those already explored in the literature, the VBSJT has a particular advantage over traditional measures used to explore personality and other police related constructs. This VBSJT included contextual information that provides the test taker with an on-the-job feel. This context not only initiates job relevant behaviors but also includes stimuli that have been demonstrated to manipulate the test taker responses. Therefore, the test measures likely behaviors that may actually be more related to how the officer would behave on the job than making assumptions based on related dimensions of the MMPI. There is also benefit in understanding how officers respond to simulated events because police officers work independently and their behavior is primarily unobserved by supervisors. Thus, this study is unique in that the measure used for appropriate exertion of control more closely simulated the job and thus manipulated situational characteristics as they would be in the real world.

The results of this study suggest that stress and cynicism in a neutral environment, such as when completing a personality questionnaire, may not translate to actual behavioral responses in given job situations. It is not concluded that the results of this study refute the evidence from the literature (e.g., Shusman, Inwald, & Landa, 1984) but that this study may lead to different conclusions about the actions taken by the officers. That is, job stress may not lead to increased negative behavioral responses. Experienced police officers may feel stress or cynicism, and this may affect their health or feeling of well-being, but they still understand the boundaries of appropriately exerting control as police officers. In fact this research demonstrates that police officers have more appropriate responses in terms of exertion of control than the general public as represented by job applicants. This research suggests that it is important to validate the outcomes of stress or other related variables that are assumed to affect behavioral responses.

Although this research does not include a measure of stress, cynicism or aggression, the tendency to have lower overexertion of control scores suggests that external factors such as stress and cynicism do not affect exertion of control.

Differences Across Organizations

The second way of utilizing the VBSJT was to explore the differences between organizational locations. The validation sample was broken into suburban or urban groups and compared on appropriate exertion of control. The hypothesis that the locations with more felony crime would have higher overexertion of control scores was not supported in the study. Despite the literature that demonstrates the negative impact of urban locations (e.g., Beutler et al, 1985), officer scores from these locations were not high on the OCS. In fact, they were significantly lower than the police officers in rural/suburban areas. Although directly contradictory to the hypothesis made in this study the results may be understandable. One conclusion may be that the police in urban areas experience more felony crime on a regular basis that desensitizes them in terms of overreaction to crimes of a less serious nature. The provocation within the test could have been seen as trivial to the officers in the urban areas.

Another conclusion could be that officers in urban areas are more aware of the negative consequences of overexertion of control. This awareness may stem from training or superiors focusing on the importance of this dimension of performance. This was partially tested in this study. The hypothesis was that the organizations with lower overexertion of control scores would have higher validities for this dimension. That is, the more important this dimension is seen throughout the organization the more likely the officers are to be careful not to overexert control. However, it was expected that the urban area would have higher scores on the OCS, and therefore, have a lower validity. Although the hypothesis was not supported, the hypothesis about equal validity across locations was tested. Although there were differences in the validity coefficients (Table 31) and the direction of the fit lines for each group appears to indicate an interaction (Figure 1), the interaction between location and the OCS was not significant.

Although neither of the hypotheses were supported, it is evident that there are differences between organizations and that they may be of concern more for the rural/suburban organizations that may need to focus more on issues of appropriate exertion of control. The results from applying the OCS to find differences between organizations were very interesting and again were contradictory to the literature. Overall, these results demonstrate alternate ways of utilizing a

VBSJT and also suggest the situational characteristics included within the test may give a more accurate idea of actual job behaviors. It should not be automatically assumed that, taken out of context, police officer feelings, attitudes, or personalities necessarily carry over to their job behaviors. Furthermore, it is apparent that selecting police officers with a test that is largely based on exertion of control in actual police encounters could help further reduce inappropriate exertion of control in real settings

In conclusion, the results of the hypotheses about police officer performance on the OCS indicated that experience may be a large determinant of performance on this dimension. Furthermore, all of these hypotheses were contradictory to the negative effects of experience that has been demonstrated in the literature. There are a few interesting conclusions that may be drawn from the findings. First, given the findings in the literature, negative effects of experience, such as stress and cynicism may not be related to overexertion of control. That is, these variables may be related to other negative behaviors such as police corruption but not to interactions with the public. Therefore, experience may be a positive indicator of public interaction. Experience as an important positive component also appears to be supported through the findings that police officers in urban locations actually scored better on the OCS than those in rural locations. It is quite possible that officers in urban areas are exposed to more situations on a daily basis and thus have higher levels of experience. Therefore, police interaction with the public may get better with experience.

Another possible conclusion is that the negative variables of stress and cynicism are related to overexertion of control behaviors but the VBSJT does not measure the actual behavior of the police officer. That is, experienced police officers may have the knowledge about the right response to the situation but may not actually be behaving in such a way. In other words, the experienced police officers may be "faking" the test. Although overexertion of control is significantly related to performance, experienced police officers may better understand what they are being evaluated on and how to effectively appear to be doing a good job than those officers with less experience.

Therefore, there are two completely different conclusions to drawn: experience is negatively related to job performance or experience is positively related to job performance. While the previous literature may indicate that experience is negatively related to police personality, these studies utilize measures that are independent of context. While it is not argued

that experience does not lead to cynicism or stress, it is more likely that these variables may not actually influence the exertion of control exhibited by the officers. Given the consistent findings across levels of experience and organizations within this study, it is concluded that knowledge about appropriate exertion of control is positively related to police performance. Although it is an empirical question weather the overexertion of control behavior is positively or negatively related to police performance, the findings from this study, using this VBSJT, indicate that police officer experience is positively related to police behavior. Police officer overexertion of control was significantly related to ratings of performance. Furthermore, it has been demonstrated that the impact of the situational variables included within the test help determine the answer choices. Thus, based on the results from this study it is concluded that experience is positively related to police officer human interaction, contrary to what might be expected from the previous literature. The contextual information provided within the VBSJT might be the critical difference between these findings and those assumed from the literature presented. While no empirical evidence is presented, it is argued that the context provided with the VBSJT provides a closer measure of actual job related behavior than measures from the MMPI or other personality measures.

Limitations

The results from this study are extremely beneficial to the understanding and subsequent development of SJTs. However, there are a few limitations within the current study that should be noted. First, the construct identified within the test does not come from a large body of literature. Appropriate exertion of control is a new construct that was applied to understand a VBSJT designed for use in law enforcement. However, this variable may come to be an important component of many SJTs because of its direct relationship with public interaction. That is, employees who interact with others as a part of their job must use appropriate exertion of control, although the appropriate level would have to be defined in the context of the job. Establishing appropriate exertion of control as an important unique construct will take more empirical work. Appropriate exertion of control was also the only construct taken from the VBSJT. Therefore, it is unclear what other important components are included in the VBSJT. Identifying the true construct validity of this test would require that all important components are identified within the VBSJT. Furthermore, including external measures of these components may help to understand the nature of these relationships.

Second, this study utilized ratings from the ten law enforcement experts to generalize to the feelings and attitudes of the applicant pool. That is, questions about liking, rudeness and pleasantness were more based on feelings of each individual rating the items. Although it is believed that the ratings of the emotional characteristics were fairly accurate and a good estimate of population averages, it would be interesting to actually explore test taker reactions to the characters and their responses to the questions. However, this would most likely severely contaminate test taker responses. Asking people to identify their feelings about the test questions would bring this thought to the forefront of consciousness and most likely influence answers to the items.

Third, the sample of ten law enforcement experts may have been a limitation in the current study. All of the police officers were from a large Midwest organization. Performance expectations within this organization may have been different from other locations. The narrow sample may have lead to systematic differences in ratings. These systematic differences could have confounded results of this study. Further, using these ratings to score the appropriate exertion of control scale was difficult when expert ratings of inappropriate exertion of control did not match the right answers assigned by the large sample of subject matter experts used to develop and score the original test. While these differences could have been due to the different focus when rating the items, it appeared that the differences were due to systematic differences between this panel and the original subject matter expert panel. However, disagreement between these groups did not occur frequently and did not appear to significantly affect the results.

Finally, the complex nature of the situations within a SJT make it hard to truly identify any particular situational characteristic. The large correlations between the situational components further confounded this issue. Although averages across many items help to remove confounding situational characteristics, there were certainly characteristics that were not considered. With all the situational variables that confound the findings it is hard to truly explain the differences. Furthermore, some of the comparisons were based on a small number of items. Therefore, one should be careful about drawing strong conclusions based on the results with only a few items within the subset.

Future Research

This research has demonstrated a new way of exploring a VBSJT. This method was useful for understanding important situational variables that had an impact on test taker

responses. Furthermore, the test taker responses revealed more information on how VBSJTs work. This new research also created new ways to use VBSJTs.

Although this paper has provided much useful information, the research is just a small step in the right direction. The primary purpose of this paper was to demonstrate a new method for understanding SJTs and how they relate to job performance. Approaches used for analyzing traditional, unidimensional tests are insufficient. SJTs are multidimensional at the item and test level. It is necessary to take a new approach to them. Research challenges aside, it is the multidimensionality itself, as a replication of reality, which seems to account for the high validity of these tests. This creates the need to approach understanding nontraditionally. Therefore, an abundance of research can stem from this methodology.

First, it is apparent that SJTs and VBSJTs are measurement techniques that can be used to select people for a variety of jobs. These tests range from public safety occupations to customer service jobs. This needs to be considered when exploring the different constructs included within the SJTs. Each SJT must be approached individually from a design and research position. Future research could attempt to identify consistent patterns of constructs or components across similar types of jobs, or similar skills and abilities across jobs. That is, there are different types of SJTs but many would probably fall in the same categories. For example, customer service is a common underlying dimension of SJTs. Although customer service could be as a bank teller or DMV employee, the constructs related to customer service are very similar. Therefore, it would be valuable to identify the important components that influence customer service.

Second, once the effective components are identified, situational characteristics could be identified that are effective for measuring these important constructs. This process should include clear definition of provocations commonly encountered by employees that are strong triggers for poor performance. Not only will this help to understand SJTs and what and how we are measuring performance but treat the development of SJTs as more of a science.

Third, VBSJTs are simulations. These tests present real aspects of the job and measure reactions of test takers that are influenced by their ability to solve real problems with multiple situational influences. While these tests have been used primarily for selection, it is possible to use them with a more diagnostic approach. Identifying specific constructs within the test, or designing a test with the knowledge of the components included would lend itself to diagnostic information available about a specific test taker or group of test takers. This research has

provided very useful information with regard to police officers and their interaction with citizens. It would be possible to give scores on the differently scored subsections of the test and also break those scores by the situational characteristics. Therefore, diagnostic scores could be given to actual police officers about areas where they should seek assistance and areas that represent strengths for them. Not only would it be helpful to let them know that they may have an approach that shows overexertion of control but also to inform them of the main factors that contributed to them missing the answers.

Finally, exploring the situational characteristics that contribute to the functioning of SJTs has an inherent weakness. That is, each situation is different and contains different stimuli. The stimuli that were evaluated in this study were only used because they were considered to be important to appropriate exertion of control and appeared across situations. This allowed for aggregate level exploration. However, there could have been confounding stimuli that were not considered. In order to account for this, SJT exploration would benefit from direct manipulation of specific situational characteristics while holding all other variables constant. For example, the exact same situation with different focal characters, either manipulating gender or ethnicity. However, this may not be feasible. Because SJTs are framed in a real world setting, test takers may recognize similar ideas and catch on to simple changes. This is a question that will have to be answered with further research.

In spite of their high validity, SJTs in general are often criticized for lack of a strong theoretical background. This is an appropriate criticism and needs to be addressed. SJTs are complex instruments that can only be understood by approaching their complexity as the driving component in their utility. The theoretical framework of VBSJTs is only just beginning to take shape but hopefully this research provides a methodology and example of how to strengthen understanding of these tests. The evolution of technology creates new avenues for simulation and a new generation of more sophisticated and powerful testing and diagnostic instruments. Building strong theoretical framework for simulations will allow our field to advance in conjunction with other fields and make appropriate use of technological resources.

CONCLUSION

This study provides a useful method for identifying a unidimensional construct within a multidimensional VBSJT. The unidimensional construct of appropriate exertion of control was found to be a very important component of the VBSJT designed for entry level law enforcement. The study also provides strong support for the utility of applying this method to understanding VBSJTs. Situational characteristics are important determinants of test taker responses and should be seriously considered. Character manipulations provided the strongest effects on appropriate exertion of control, thus providing support for the use of VBSJTs. With the use of dramatic video it was possible to realistically and effectively manipulate the behavior of characters to elicit reactions from the examinees. VBSJT design and development could benefit considerably from the knowledge of effective stimuli to include within the item stem. Therefore, literature should apply this methodology to other components within VBSJTs, or SJTs, to develop a sound theoretical framework to guide development. The results should be applied to increase knowledge about how to effectively manipulate situational characteristics. Certain types of situational characteristics will always need to be included but should be done so intentionally with understanding of what will be gained and lost in item functioning. For example, gender may not have had a large effect, indicating that it is not an important manipulation to appropriate exertion of control but it may contribute to other dimensions of the VBSJT or at least contribute to the realistic feeling of the test.

This study also provides other useful, diagnostic ways to utilize SJTs besides in personnel selection. Because VBSJTs are designed to simulate the job the answer choices that the test takers choose can be used to diagnose different aspects of performance. Therefore, one way to utilize these tests is to identify discreet dimensions within the realm of the multidimensional test. This dimensional information can be used to specifically address underlying strengths and weaknesses of employees or organizations.

Although the effectiveness of SJTs has been well documented, the research attempting to explain their effectiveness is limited. This paper provides a new approach to understanding, developing, and utilizing SJTs that can help advance the literature in this area. The use of these tests will continue and applying this new approach will help to advance understanding of SJTs and maximize their effectiveness.

References

Abernethy, A. D. (1995). The development of an anger management training program for law enforcement personnel. In L. R. Murphy, J. L. Hurrell, S. L. Sauter, & G. P. Keita (Eds.), Job Stress Interventions (pp. 21-30). Washington D. C.: American Psychological Association.

Banton, M. (1967). The Policeman in the Community. New York: Basic Books.

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

Bem, D. J. & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. Psychological Review, 81(6), 506-520.

Bem, D. J. & Funder, D. C. (1978). Predicting more of the people more of the time: Assessing the personality of situations. Psychological Review, 85, 485-501.

Beutler, L. E., Nussbaum, P. D., & Meredith, K. E. (1988). Changing personality patterns of police officers. Professional Psychology: Research and Practice, 19, 503-507.

Beutler, L. E., Storm, A., Kirkish, P., Scogin, F., & Gaines, J. A. (1985). Parameters in the prediction of police officer performance. Professional Psychology, 16(2), 324-335.

Bowers, K. S. (1973). Situationism in psychology: An analysis and a critique. Psychological Review, 80, 307-336.

Brown, M. K. (1981). Working the street: Police discretion and the dilemmas of reform. New York: Russell Sage Foundation.

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

Burgess, J. R. D. & James, L. R. (1998). A comparison of test validity: A self-report format vs. a reasoning format. Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Burkhart, B. R. (1980). Conceptual issues in the development of police selection procedures. Professional psychology, 11, 121-129.

Burgin, A. L. (1978). The management of stress in policing. Police Chief, 45, 53-54.

Buss, A. H. (1961). The psychology of aggression. New York: Wiley.

Carrington, D. H. (1949). Note on the cardall practical judgment test. Journal of Applied Psychology, 33, 29-30.

Cascio, W. F. (1998). Applied psychology in human resource management (fifth edition). New Jersey: Prentice Hall.

Chan, D. & Schmitt, N. (1997). Video-based paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions. Journal of Applied Psychology, 82(1), 143-159.

Chan, D., Schmitt, N., Sacco, J. M., and DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. Journal of Applied Psychology, 83, 471-485.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Clevenger, J., Jockin, T., Morris, S., & Anselmi (1999). 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., Weichmann, D., Schmitt, N., & Schmidt-Harvey, V. (in press). The situational judgment inventory as a measure of contextual job knowledge.

Cohen, J. & Cohen, P. (1983). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New Jersey: Laurence Earlbaum.

Conroy, D. L. & Hess, K. M. (1992). Officers at Risk: How to Identify and Cope with Stress. Placerville: Custom Publishing.

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

DeShon, Smith, Chan, & Schmitt (1998). Can racial differences in cognitive test performance be reduced by presenting problems in a social context? Journal of Applied Psychology, 83(3), 438-451.

Desmedt, J. C. (1984). Use of force paradigm for law enforcement. Journal of Police Science and Administration, 12(2), 170-176.

Drummond, T. (1999). It's not just New Jersey: Cops across the U.S. often search people because of their race, a study says. Time, June 14.

Dulsky, S. G. & Krout, M. H. (1950). Predicting promotion potential on the basis of psychological tests. Personnel Psychology, 3, 345-351.

Elliot, S., Lawty-Jones, M., & Jackson, C. (1996). Dissimulation on self-report and objective measures of personality. Personality and Individual Differences, 21(3), 335-343.

Epstein, S. & O'Brien, E. J. (1985). The person-situation debate in historical and current perspective. Psychological Bulletin, 98(3), 513-537.

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

Fiske, S. T. & Taylor, S. E. (1991). Social Cognition (2nd ed.). New York: McGraw-Hill.

Folger, R., & Baron, R. A. (1996). Violence and hostility at work: A model of reactions to perceived injustice. In G. R. VanderBos & E. Q. Bulatao, Violence on the job: Identifying risks and developing solutions (pp. 51-85). Washington, D. C.: American Psychological Association.

Forier, K. (1972). The police culture: Its effects on sound police-community relations. The Police Chief, 1, 33-36.

Furcon, J. (1979). An overview of police selection: Some issues, questions, and challenges. In C. D. Spielberger (Ed.), Police Selection and Evaluation: Issues and Techniques (pp. 3-10). New York: Praeger.

Green, P. D. (1999). The visual-oral conditional reasoning test: Predicting scholastic misconduct and deception. Unpublished Ph.D. dissertation, The University of Tennessee, Knoxville.

Hecker, H. L. & Lunde, D. T. (1985). On the diagnosis and treatment of chronically hostile individuals. In M. A. Chesney & R. H. Rosenman (Eds), Anger and Hostility in Cardiovascular and Behavioral Disorders (pp. 227-240). Washington D. C.: Taylor & Francis.

Hemmelgarn, A. L. (1996). Measuring achievement motives: The process of conditional reasoning through the reading comprehension and inference test. Unpublished Ph.D. dissertation, The University of Tennessee, Knoxville.

Henderson, N. D. (1979). Criterion-Related validity of personality and aptitude scales: A comparison of validation results under voluntary and actual test conditions. In C. D.

Spielberger (Ed.), Police Selection and Evaluation: Issues and Techniques (pp. 179-195). New York: Praeger.

Hough, L. M. (1998). Personality at work: Issues and evidence. In M. Hakel (Ed.), Beyond Multiple Choice: Evaluating Alternatives to Traditional Testing for Selection (pp. 131-166). Hillsdale, NJ: Earlbaum.

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

Inwald, R. E. & Shusman, E. J. (1984). The IPI and MMPI as predictors of academy performance for police recruits. Journal of Police Science and Administration, 12(1), 1-11.

James, L. R. (1998). Measurement of Personality via Conditional Reasoning. Organizational Research Methods, 1(2), 131-163.

James, L. R., McIntyre, M.D., Glisson, C. A., Green, P. D., Patton, T. W., Mitchell, T. R., & Williams, L. J. (2000). Use of conditional reasoning. Manuscript submitted for publication.

Jones, C. & DeCotiis, T., A. (1986). Video-assisted selection of hospitality employees. The Cornell H.R.A. Quarterly,

Jones, M. W., Dwight, S. A. & Nouryan, T. R. (1999). Exploration of the construct validity of a situational judgment test used for managerial assessment. Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Kenrick, D. T., & Funder, D. C. (1991). The person-situation debate: Do personality traits really exist? In V. Derlega, B. Winstead, & W. Jones (eds.), Personality: Contemporary Theory and Research (pp. 149-174). Chicago, IL: Nelson-Hall.

Kessler, D. A. (1989). Improving policing: The impact of neighborhood-oriented policing. Unpublished Ph.D. dissertation, Indiana University.

Laursen, B. & Collins, W. A. (1994). Interpersonal conflict during adolescence. Psychological Bulletin, 115, 197-209.

Mandell, M. M. (1953). How Supervise? In O. K. Buros (Ed.), The fourth mental measurements yearbook (pp.774-775). Highland Park, NJ: The Gryphon Press.

Mayer, J. D. & Geher, G. (1996). Emotional intelligence and the identification of emotion. Intelligence, 22, 89-113.

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

McIntyre, M. D. (1995). A feasibility study of a conditional reasoning measure of aggressiveness and prosocialability. Unpublished Ph.D. dissertation, The University of Tennessee, Knoxville.

Mischel, W. (1968). Personality and Assessment. New York: Wiley.

Mischel, W. (1990). Personality dispositions revisited and revised: A view after three decades. In L. A. Pervin (Ed.), Handbook of personality: theory and research (pp. 111-134). New York: Guilford.

Migetz, D. Z., James, L. R., & Ladd, T. (1999). A validation of the conditional reasoning measure of achievement motivation and fear of failure. Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Mihanovich, C. S. (1981). The blue pressure cooker. Criminal Justice, 81, 101-102.

Moos, R. H. & Insel, P. M. (1974). Issues in Social Ecology. Palo Alto, National Press.

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

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

Murphy, J. J. (1972). Current practices in the use of the psychological testing by police agency. Journal of Criminal Law, Criminology, and Political Science, 63, 570-576.

Murray, H. A. (1938). Explorations in Personality. New York: Oxford University.

Newport, F. (1999). Racial profiling is seen as widespread, particularly among young black men. Gallup Poll Releases, December 9.

O'Leary-Kelly, A. M., Griffen, R. W., & Glew, D. J. (1996). Organization-motivated aggression: A research framework. Academy of Management Review, 21, 225-253.

Patton, T. W. (1998). Measuring personal reliability via conditional reasoning: Identifying people who will work reliably. Unpublished Ph.D. dissertation, The University of Tennessee, Knoxville.

Pereira, G. M. & Harvey, V. S. (1999). Situational judgment tests: do they measure ability, personality or both? Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

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(3), 241-258.

Robertson, I. T. & Kandola, R. S. (1982). Work sample tests: validity, adverse impact and applicant reaction. Journal of Occupational Psychology, 55, 171-183.

Rokeach, M., Miller, M. G., & Snyder, J. A. (1971). The value gap between police and policed. Journal of Social Issues, 27(2), 155-171.

Rosen, N. A. (1961). How Supervise? – 1943-1960. Personnel Psychology, 14, 87-99.

Ryan, Ployhart & Friedel (1998). Using personality testing to reduce adverse impact: a cautionary note. Journal of Applied Psychology, 83(2), 298-307.

Sacco, J. M., Scheu, C. R., Ryan, A. M., Schmitt, N., Schmidt, D. B., and Rogg, K. L. (2000). Reading level and verbal test scores as predictors of subgroup differences and validities of situational judgment tests. Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Saccuzzo, D. P.; Higgins, G.; Lewandowski, D. (1974). Program for psychological assessment of law enforcement officers: Initial evaluation. Psychological Reports, 35, 651-654.

Sacket, P. R. & Roth, L. (1996). Multi-stage selection strategies: A monte carlo investigation of effects on performance and minority hiring. Personnel Psychology, 49, 549-572.

Sackett, P. R. & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. Personnel Psychology, 50, 707-722.

Saxe, S. J., & Reiser, M. (1976). A comparison of three police applicant groups using the MMPI. Journal of Police Science and Administration, 4, 419-425.

Sells, S. B. (1966). Ecology and science of psychology. Multivariate Behavioral Research, 1, 131-144.

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. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research. Psychological Bulletin, 124(2), 262-274.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. Journal of Applied Psychology, 82, 719-730.

Sewell, J. D. (1984). Stress in university law enforcement. Journal of Higher Education, 55(4), 515-523.

Shealy, A. E. (1979). Police corruption: Screening out high-risk applicants. In C. D. Spielberger (Ed.), Police Selection and Evaluation: Issues and Techniques (pp. 197-210). New York: Praeger.

Sherrid, S. D. (1979). Changing police values. In C. D. Spielberger (Ed.), Police Selection and Evaluation: Issues and Techniques (pp. 167-176). New York: Praeger.

Shoda, Y., Mischel, W. & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into idiographic analysis of personality. Journal of Personality and Social Psychology, 67(4), 674-687.

Shusman, E. J., Inwald, R. E. & Landa, B. (1984). Correction Officer Job Performance as Predicted by the IPI and MMPI: A Validation ad Cross-Validation Study. Criminal Justice and Behavior, 11(3), 309-329.

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

Smith, D. H. F. (1996). A conditional reasoning approach to measuring relative achievement motivation: Validation results from an applied setting. Unpublished Ph.D. dissertation, The University of Tennessee, Knoxville.

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

Spielberger, C. D., Jacobs, G., Russell, S., & Crane, R. S. (1983). Assessment of anger: The state-trait anger scale. In J. N. Butcher & C. D. Spielberger (Eds), Advances in Personality Assessment (Volume 2), (pp. 161-189). New Jersey: Lawrence Erlbaum.

Stanley, S. A., & Stokes, G. S. (1998). Controlling faking with test format: An examination. Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Sterling, J. W. (1971). Changing in role concepts of police officers. Gaithersberg: International Association of Chief of Police.

Strong, M. H. & Najar, M. J. (1999). 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.

Swander, C. (2000). Video-based and written situational judgment tests: method effects, criterion validity, and construct validity. Unpublished manuscript.

Swander, C. & Spurlin, O. (1993). Validation of the Diplomat™ Video Test. unpublished Technical Report, Ergometrics, Edmonds, WA.

Swander, C. & Spurlin, O. (1995). Frontline™ Validation Report. unpublished Technical Report, Ergometrics, Edmonds, WA.

Swander, C. & Spurlin, O. (1997a). Diplomat Phone Skills Test™ Validation Report. unpublished Technical Report, Ergometrics, Edmonds, WA.

Swander, C. & Spurlin, O. (1997b). TellerTest™ Validation Report. unpublished Technical Report, Ergometrics, Edmonds, WA.

Swander, C. & Spurlin, O. (1998). Correction Officer Video Test™ Validation Report. unpublished Technical Report, Ergometrics, Edmonds, WA.

Tenopyr, M. L. (1969). The comparative validity of selected leadership scales relative to success in production management. Personnel Selection, 22, 77-85.

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

Washington, L. (2000). Police prejudice promotes profiling. The Philadelphia Tribune.

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(3), 679-700.

Table 1

Descriptive Statistics of OCS, UCS, and VBSJT

	Scale	Items	N	Min	Max	M	SD
Applicants	OCS	39	5426	0	16	3.45	2.40
	UCS	18	5426	-10.80	0	-1.91	1.49
	VBSJT	78	5426	96	293	223.55	22.97
Incumbents	OCS	39	334	0	10.80	2.83	1.90
	UCS	18	334	-6.80	0	-1.27	1.15
	VBSJT	78	334	178	277	242.43	17.59

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. VBSJT = Frontline™ video-based situational judgment test.

Table 2

Descriptive Statistics of Average Item Responses on OCS and UCS

	N	Min	Max	M	SD
OCS	5426	.00	.41	.088	.061
UCS	5426	-.60	.00	-.106	.083

Note. Numbers represent the average for each item on both scales across the applicant sample. OCS = Overexertion of control scale. UCS = Underexertion of control scale.

Table 3
Intercorrelations and Descriptive Statistics of VBSJT Ratings

		M	SD	N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Provocation	3.437	.488	43	1.0														
2	Likeability	2.649	.773	43	-.66	1.0													
3	Rude	.221	.346	43	.56	-.59	1.0												
4	Aggressive	.240	.327	43	.57	-.62	.72	1.0											
5	Suspicious	.379	.396	43	.55	-.53	.10	.11	1.0										
6	Pleasant	.288	.380	43	-.56	.79	-.47	-.42	-.52	1.0									
7	Cooperative	.340	.372	43	-.54	.70	-.44	-.39	-.53	.85	1.0								
8	Sympathy	.207	.305	43	-.56	.67	-.32	-.30	-.51	.52	.56	1.0							
9	Contemptible	.123	.237	43	.19	-.13	-.14	.17	.13	-.08	-.15	.01	1.0						
10	Complaint	.207	.365	43	.21	-.28	.16	.17	.00	-.13	-.02	-.08	.20	1.0					
11	Warrant	.109	.283	43	-.06	.32	-.02	.01	-.18	.37	.31	.40	-.12	-.17	1.0				
12	Broken Law	.423	.395	43	.01	.21	.03	.19	-.33	.20	.17	.19	.23	-.18	.37	1.0			
13	Potential	.556	.330	43	.46	-.30	.43	.33	-.09	-.25	-.24	-.05	.13	.40	.27	.18	1.0		
14	Ethnicity[1]	.390	.494	41	.26	-.04	.14	.03	.15	.00	-.07	-.01	-.05	-.04	.31	.06	.23	1.0	
15	Ethnicity[2]	.342	.481	38	.26	-.08	.12	.03	.17	-.08	-.11	.00	-.01	.01	.29	.00	.23	1.0	1.0
16	Gender	.216	.417	37	-.52	.31	-.11	-.07	-.45	.26	.41	.65	-.04	-.16	.11	.09	-.06	-.40	-.37

Note. Correlations are based on the average of the ten expert raters. Provocation refers to the overall provocation item on the rating sheet. Provocation was coded (Strong Provocation Toward Underexertion of Control = 1; Provocation Toward Underexertion of Control = 2; No Provocation = 3; Provocation Toward Overexertion of Control = 4; Strong Provocation Toward Overexertion of Control = 5). Likeability was coded (Strongly Dislike = 1; Dislike = 2; Neutral = 3; Like = 4; Strongly Like = 5). The variables Rude through Potential correspond to the rating sheet questions about situational provocation (see Appendix B). Average dimension scores were used to compute the correlations. Ethnicity contains the items that had a minority character as the focal character. Ethnicity[1] was dummy coded (Minority = 1; Non-minority = 0). Ethnicity[2] was dummy coded (African American = 1; Caucasian = 0). Gender was dummy coded (Female = 1; Male = 0).

Table 4
Interrater Reliability Estimates of Appropriate Exertion of Control Ratings

Items	Reliability (α)
All Items	.94
Item 1	.96
Item 2	.97
Item 3	.93
Item 4	.96
Item 5	.96
Item 6	.98
Item 7	.95
Item 8	1.00
Item 9	.95
Item 10	.99
Item 11	.95
Item 12	.98
Item 13	.98
Item 14	.99
Item 15	.95
Item 16	.87
Item 17	.98
Item 18	1.00
Item 19	.91
Item 20	.90
Item 21	.96
Item 22	.93
Item 23	.94
Item 24	.96
Item 25	.89
Item 26	.95
Item 27	.93
Item 28	.90
Item 29	.96
Item 30	.98
Item 31	.96
Item 32	.99
Item 33	.97
Item 34	.93
Item 35	.97
Item 36	.82
Item 37	.92
Item 38	.95
Item 39	.99
Item 40	.99
Item 41	.99
Item 42	.90
Item 43	.87

Table 5
Interrater Reliability for Situational Characteristics Items

Provocation Items	Reliability (α)	Interrater Agreement (%)
Overall Provocation	.80	--
Likeability	.93	--
Rude	.95	88
Aggressive	.93	83
Suspicious	.95	81
Pleasant	.94	83
Cooperative	.93	79
Sympathetic	.91	82
Contemptible	.94	89
Complaint	.97	93
Warrant	.98	96
Broken Law	.94	80
Potential	.87	68

Note. α = Cronbach's alpha. Interrater agreement was computed by taking the total number of agreements divided by the total number of agreements plus the total number of disagreements for each pair of raters. The variables Rude through Potential correspond to the situational variable questions on the rating sheet (see Appendix B).

Table 6
Intercorrelations of Ratings on OCS Items

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Provocation															
2	Likeability	-.61														
3	Rude	.49	-.49													
4	Aggressive	.39	-.42	.72												
5	Suspicious	.50	-.39	.02	.12											
6	Pleasant	-.43	.67	-.39	-.34	-.41										
7	Cooperative	-.41	.63	-.44	-.38	-.37	.89									
8	Sympathy	-.37	.49	-.30	-.26	-.27	.35	.55								
9	Contemptible	.18	.05	-.20	.04	.25	-.04	-.08	.04							
10	Complaint	.20	-.22	.24	.02	-.06	-.10	-.03	-.13	.22						
11	Warrant	-.07	.38	-.05	-.01	-.15	.41	.35	.38	-.13	-.21					
12	Broken Law	-.23	.19	.02	.14	-.24	.35	.23	.14	.18	-.17	.24				
13	Potential	.45	-.23	.34	.14	-.03	-.27	-.18	.01	.09	.43	.15	-.07			
14	Ethnicity[1]	.24	-.05	.15	.00	.26	-.02	-.11	-.14	-.09	-.05	.34	.02	.15		
15	Ethnicity[2]	.23	-.12	.16	-.04	.32	-.13	-.21	-.09	-.06	.01	.33	-.02	.16	-.24	
16	Gender	-.41	.36	-.16	-.08	-.34	.24	.45	.74	.01	-.16	.07	.19	-.05	.31	-.42

Note. Correlations are based on the average of the ten expert raters. All correlations are based on the 39 items in the OCS. Provocation refers to the overall provocation item on the rating sheet. Provocation was coded (Strong Provocation Toward Underexertion of Control = 1; Provocation Toward Underexertion of Control = 2; No Provocation = 3; Provocation Toward Overexertion of Control = 4; Strong Provocation Toward Overexertion of Control = 5). Likeability was coded (Strongly Dislike = 1; Dislike = 2; Neutral = 3; Like = 4; Strongly Like = 5). The variables Rude through Potential correspond to the situational variable questions on the rating sheet (see Appendix B). Average dimension scores were used to compute the correlations. Ethnicity contains the items that had a minority character as the focal character. Ethnicity[1] was dummy coded (Minority = 1; Non-minority = 0). Ethnicity[2] was dummy coded (African American = 1; Caucasian = 0). Gender was dummy coded (Female = 1; Male = 0).

Table 7
Intercorrelations of Ratings on UCS Items

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Provocation															
2	Likeability	-.40														
3	Rude	.	.													
4	Aggressive	.22	-.22	.												
5	Suspicious	.55	-.55	.	.22											
6	Pleasant	-.20	.66	.	-.19	-.66										
7	Cooperative	-.20	.66	.	-.19	-.66	1.00									
8	Sympathy	-.33	.60	.	-.13	-.60	.67	.67								
9	Contemptible	.40	-.20	.	.54	.40	-.05	-.05	-.24							
10	Complaint	.32	-.08	.	-.09	.32	.08	.08	-.19	.79						
11	Warrant	-.04	.40	.	-.09	-.40	.44	.44	.24	-.16	-.13					
12	Broken Law	-.22	.15	.	.24	-.67	.34	.34	.27	.15	.00	.00				
13	Potential	.58	-.05	.	-.22	.13	-.03	-.03	-.21	.20	.40	.04	.00			
14	Ethnicity[1]	.46	.07	.	-.21	-.03	.13	.13	.10	-.07	.07	.44	-.07	.41		
15	Ethnicity[2]	.42	.04	.	-.20	.07	.04	.04	.15	-.04	.10	.33	-.16	.36	-.54	
16	Gender	-.40	.16	.	-.09	-.40	.44	.44	.66	-.16	-.13	-.13	.35	-.32	.29	-.31

Note. Correlations are based on the average of the ten expert raters. All correlations are based on the 39 items in the OCS. Provocation refers to the overall provocation item on the rating sheet. Provocation was coded (Strong Provocation Toward Underexertion of Control = 1; Provocation Toward Underexertion of Control = 2; No Provocation = 3; Provocation Toward Overexertion of Control = 4; Strong Provocation Toward Overexertion of Control = 5). Likeability was coded (Strongly Dislike = 1; Dislike = 2; Neutral = 3; Like = 4; Strongly Like = 5). The variables Rude through Potential correspond to the rating sheet questions about situational provocation (see Appendix B). Average dimension scores were used to compute the correlations. Ethnicity contains the items that had a minority character as the focal character. Ethnicity[1] was dummy coded (Minority = 1; Non-minority = 0). Ethnicity[2] was dummy coded (African American = 1; Caucasian = 0). Gender was dummy coded (Female = 1; Male = 0).

Table 8
Validity Coefficients and Intercorrelations for the VBSJT, OCS, and UCS

Performance Dimensions	VBSJT	OCS	UCS
VBSJT	1.000	-.667**	.141*
OCS	-.667**	1.000	.030
UCS	.141*	.030	1.00
1) Public Contact Situations – Communication Style	.239**	-.207**	.026
2) Maturity When Intervening in Stressful Situations	.256**	-.170**	.050
3) Interrogation/Investigation	.224**	-.145**	-.012
4) Initiative in Handling and Resolving Situations	.279**	-.198**	-.004
5) Officer Safety Orientation	.203**	-.204**	-.017
6) Sensitivity to Diverse Groups	.154**	-.075	.063
7) Situational Assessment and Analysis	.246**	-.178**	.057
8) Relations with Supervisors and Management	.190**	-.177**	.067
9) Relations with Co-workers	.160**	-.138*	.053
10) Work Habits	.196**	-.111*	.060
11) Professional Behavior and Bearing	.169**	-.153**	-.036
12) Paperwork	.154**	-.131**	.097
13) Persistence in Learning and Keeping Up-to-Date	.214**	-.151**	.041
14) Physical Skill	.019	-.070	-.013
15) Driving Skill	.194**	-.202**	-.042
16) Weapons Skill	.226**	-.171**	-.031
Overall Evaluation	.332**	-.234**	.054

Note. ** = Correlation is significant at the .01 level (2-tailed). * = Correlation is significant at the .05 level (2-tailed). N = 334. OCS = Overexertion of control scale. UCS = Underexertion of control scale. VBSJT = Frontline™ video-based situational judgment test.

Table 9.
Hierarchical Regression of OCS and VBSJT on Overall Performance.

Predictor Variables	β	R^2	df	ΔR^2	Δdf
Step 1					
VBSJT	.332*	.110	1		
Step 2					
OCS	-.023	.111	3	.001	1

Note. * = Correlation is significant at the .05 level (2-tailed). OCS = Overexertion of control scale. UCS = Underexertion of control scale. VBSJT = Frontline™ video-based situational judgment test.

Table 10
Descriptive Statistics for Provocation Item Subsets

Scale	Subscale	Items	M	N	SD
OCS	Provocation	16	.109	5426	.093
	No Provocation	22	.077	5426	.063
UCS	Provocation	10	-.146	5426	.127
	No Provocation	7	-.063	5426	.084

Note. Provocation refers to provocation toward overexertion of control. OCS = Overexertion of control scale. UCS = Underexertion of control scale.

Table 11
Dependent Samples t-test Between Provocation Item Subsets

	Subscale	M_D	SD	t	df	p
OCS	No Provocation - Provocation	.031	.090	25.46	5425	.000
UCS	No Provocation - Provocation	-.083	.139	-44.36	5425	.000

Note. Provocation refers to provocation toward overexertion of control. OCS = Overexertion of control scale. UCS = Underexertion of control scale.

Table 12
Specific Provocation Subgroups on OCS

Subscale	Items	Min	Max	M	SD
Rude	10	.00	.75	.118	.110
Not Rude	29	.00	.43	.078	.062
Aggressive	9	.00	.66	.130	.124
Not Aggressive	31	.00	.43	.078	.061
Suspicious	15	.00	.68	.072	.082
Not Suspicious	24	.00	.50	.099	.069
Pleasant	12	.00	.51	.067	.067
Not Pleasant	27	.00	.57	.098	.077
Cooperative	14	.00	.44	.060	.059
Not Cooperative	25	.00	.62	.104	.082
Sympathetic	8	.00	.57	.039	.074
Not Sympathetic	31	.00	.52	.101	.071
Contemptible	4	.00	.85	.049	.112
Not Contemptible	35	.00	.44	.093	.065
Complaint	9	.00	.68	.068	.094
No Complaint	30	.00	.42	.095	.067
Warrant	5	.00	1.12	.060	.128
No Warrant	34	.00	.47	.093	.066
Broken Law	19	.00	.44	.063	.065
No Broken Law	20	.00	.61	.112	.085
Potential	24	.00	.55	.091	.074
No Potential	15	.00	.53	.084	.071

Note. N = 5426. OCS = Overexertion of control scale. UCS = Underexertion of control scale.

Table 13
Paired Samples t-tests on Specific Provocation Subgroups for OCS

	Comparison	M_D	SD	t	df	p
Pair 1	Rude – Not Rude	.040	.106	28.11	5425	.000
Pair 2	Aggressive – Not Aggressive	.052	.120	31.91	5425	.000
Pair 3	Suspicious – Not Suspicious	-.026	.085	-22.81	5425	.000
Pair 4	Pleasant – Not Pleasant	-.031	.089	-26.10	5425	.000
Pair 5	Cooperative – Not Cooperative	-.044	.089	-36.72	5425	.000
Pair 6	Sympathetic – Not Sympathetic	-.062	.091	-50.03	5425	.000
Pair 7	Contemptible – Not Contemptible	-.044	.117	-27.62	5425	.000
Pair 8	Complaint – No Complaint	-.027	.097	-20.30	5425	.000
Pair 9	Warrant – No Warrant	-.033	.136	-17.86	5425	.000
Pair 10	Broken Law – No Broken Law	-.049	.088	-41.14	5425	.000
Pair 11	Potential – No Potential	.007	.081	6.22	5425	.000

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. M_D = Mean Difference.

Table 14
Specific Provocation Subgroups on UCS

Subscale	Items	Min	Max	M	SD
Suspicious	10	-.91	.00	-0.092	0.115
Not Suspicious	8	-.59	.00	-0.124	0.104
Pleasant	7	-.74	.00	-0.122	0.112
Not Pleasant	11	-.89	.00	-0.095	0.110
Cooperative	7	-.74	.00	-0.122	0.112
Not Cooperative	11	-.89	.00	-0.095	0.110
Sympathetic	4	-.70	.00	-0.073	0.121
Not Sympathetic	14	-.74	.00	-0.115	0.100
Contemptible	3	-1.20	.00	-0.099	0.188
Not Contemptible	15	-.69	.00	-0.107	0.087
Complaint	2	-1.35	.00	-0.147	0.278
No Complaint	16	-.66	.00	-0.101	0.082
Warrant	2	-.95	.00	-0.172	0.242
No Warrant	16	-.68	.00	-0.098	0.087
Broken Law	9	-.57	.00	-0.134	0.111
No Broken Law	9	-.99	.00	-0.078	0.109
Potential	9	-.84	.00	-0.139	0.141
No Potential	9	-.58	.00	-0.079	0.078

Note. N = 5426. OCS = Overexertion of control scale. UCS = Underexertion of control scale.

Table 15
Paired Samples t-tests on Specific Provocation Subgroups for UCS

	Comparison	M_D	SD	t	df	p
Pair 1	Suspicious – Not Suspicious	0.032	0.147	16.08	5425	.000
Pair 2	Pleasant – Not Pleasant	-0.027	0.151	-13.26	5425	.000
Pair 3	Cooperative – Not Cooperative	-0.027	0.151	-13.26	5425	.000
Pair 4	Sympathetic – Not Sympathetic	0.043	0.155	20.34	5425	.000
Pair 5	Contemptible – Not Contemptible	0.008	0.195	3.03	5425	.002
Pair 6	Complaint – No Complaint	-0.046	0.277	-12.23	5425	.000
Pair 7	Warrant – No Warrant	-0.074	0.254	-21.45	5425	.000
Pair 8	Broken Law – No Broken Law	-0.055	0.143	-28.44	5425	.000
Pair 9	Potential – No Potential	-0.060	0.147	-30.12	5425	.000

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. M_D = Mean Difference.

Table 16
Descriptive Statistics for Ethnic Item Subsets

Scale	Subscale	Items	N	Min	Max	M	SD
OCS	Minority	13	5426	.00	.65	.079	.087
	African American	11	5426	.00	.69	.063	.089
	Non-minority	21	5426	.00	.48	.097	.072
UCS	Minority	7	5426	-1.00	.00	-.199	.178
	African American	5	5426	-1.16	.00	-.173	.189
	Non-minority	10	5426	-.68	.00	-.053	.076

Note. N = 5426. OCS = Overexertion of control scale. UCS = Underexertion of control scale.

Table 17
Ethnic Dependent Samples t-tests on OCS and UCS

Scale	Comparison	M_D	SD	t	df	p
OCS	Minority - Non-minority	-.018	.092	-14.59	5425	.000
	African American - Non-minority	-.034	.095	-26.08	5425	.000
UCS	Minority - Non-minority	-.145	.178	-60.11	5425	.000
	African American - Non-minority	-.119	.187	-46.85	5425	.000

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. M_D = Mean Difference.

Table 18
Descriptive Statistics for Ethnic Item Subsets by Ethnic Group of Test Taker

Scale	Ethnicity	Subscale	M	N	SD
OCS	African American	African American	.075	371	.095
		Non-minority	.133	371	.088
	Caucasian	African American	.061	3826	.087
		Non-minority	.091	3826	.067
UCS	African American	African American	-.208	371	.217
		Non-minority	-.065	371	.083
	Caucasian	African American	-.168	3826	.185
		Non-minority	-.052	3826	.076

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale.

Table 19
Ethnic Dependent Samples t-tests on OCS and UCS by Ethnic Group of Test Taker

Scale	Ethnicity	Comparison	M_D	SD	t	df	p
OCS	African American	Minority – Non-minority	-.058	.101	-10.94	370	.000
	Caucasian	Minority – Non-minority	-.030	.094	-19.88	3825	.000
UCS	African American	African American – Non-minority	-.143	.207	-13.26	370	.000
	Caucasian	African American – Non-minority	-.116	.185	-38.89	3825	.000

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. M_D = Mean Difference.

Table 20
Descriptive Statistics for Gender Item Subsets

Scale	Subscale	Items	Min	Max	M	SD
OCS	Male	25	.00	.58	.084	.074
	Female	8	.00	.55	.087	.079
UCS	Male	16	-.68	.00	-.119	.093
	Female	2	-1.10	.00	-.002	.033

Note. N = 5426. OCS = Overexertion of control scale. UCS = Underexertion of control scale.

Table 21
Gender Dependent Samples t-tests on OCS and UCS

Scale	Comparison	M_D	SD	t	df	p
OCS	Male - Female	-.004	.097	-2.86	5425	.004
UCS	Male - Female	-.117	.099	-87.3	5425	.000

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. M_D = Mean Difference. M_D = Mean Difference.

Table 22
Descriptive Statistics for Gender Item Subsets by Gender of Test Taker

Scale	Gender	Subscale	M	N	SD
OCS	Male	Male	.083	4375	.073
		Female	.089	4375	.080
	Female	Male	.090	808	.080
		Female	.085	808	.077
UCS	Male	Male	-.119	4375	.093
		Female	-.002	4375	.035
	Female	Male	-.127	808	.099
		Female	-.001	808	.028

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale.

Table 23
Gender Dependent Samples t-tests on OCS and UCS by Gender of Test Taker

Scale	Gender	Comparison	M_D	SD	t	df	p
OCS	Male	Male - Female	-.006	.097	-3.97	4374	.000
	Female	Male - Female	.005	.100	1.52	807	.128
UCS	Male	Male - Female	-.117	.099	-78.27	4374	.000
	Female	Male - Female	-.126	.103	-34.62	807	.000

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. M_D = Mean Difference.

Table 24
Descriptive Statistics for Likeable Item Subsets

Scale	Subscale	Items	Min	Max	M	SD
OCS	Like	6	.00	.67	.092	.087
	Neutral	12	.00	.56	.086	.093
	Dislike	21	.00	.56	.089	.075
UCS	Like	3	-.93	.00	-.097	.161
	Neutral	6	-.77	.00	-.106	.115
	Dislike	9	-.89	.00	-.109	.121

Note. N = 5426. OCS = Overexertion of control scale. UCS = Underexertion of control scale. VBSJT = Frontline™ video-based situational judgment test.

Table 25
Likeable Item Subset Dependent Samples t-test on OCS and UCS

Scale	Subgroup	M_D	SD	t	df	p
OCS	Like - Neutral	.005	.117	3.43	5425	.001
	Like - Dislike	.003	.105	1.98	5425	.048
	Neutral - Dislike	-.003	.095	-2.02	5425	.043
UCS	Like - Neutral	.009	.195	3.48	5425	.001
	Like - Dislike	.013	.200	4.62	5425	.000
	Neutral - Dislike	.003	.151	1.63	5425	.102

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. M_D = Mean Difference.

Table 26
Independent Samples t-test Between Applicant and Incumbents on OCS, UCS and VBSJT

Scale	M_D	df	t	p
OCS	.618	5758	-4.62	.000
UCS	.641	5758	7.7	.000
VBSJT	18.88	5758	14.76	.700

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. VBSJT = Frontline™ video-based situational judgment test. M_D = Mean Difference.

Table 27
Descriptive Statistics for Tenure

Scale	Tenure	N	M	SD
OCS	<2 years	261	2.90	1.92
	>2 years	73	2.60	1.86
UCS	<2 years	261	-1.19	1.14
	>2 years	73	-1.55	1.16
VBSJT	<2 years	261	242.27	17.33
	>2 years	73	243.03	18.59

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. VBSJT = Frontline™ video-based situational judgment test.

Table 28
Tenure Independent Samples t-tests on OCS, UCS and VBSJT

Scale	M_D	df	t	p
OCS	.295	332	1.17	.243
UCS	.364	332	2.40	.017
VBSJT	-.759	332	-.326	.745

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. VBSJT = Frontline™ video-based situational judgment test. M_D = Mean Difference.

Table 29
Descriptive Statistics by Location

Scale	Location	N	M	SD
OCS	Suburban	196	3.12	2.01
	Urban	138	2.43	1.67
UCS	Suburban	196	-1.24	1.16
	Urban	138	-1.30	1.15
VBSJT	Suburban	196	241.99	18.35
	Urban	138	243.06	16.48

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. VBSJT = Frontline™ video-based situational judgment test.

Table 30
Independent Samples t-test between location means on OCS, UCS, and VBSJT

	M_D	df	t	p
OCS	.690	332	3.31	.001
UCS	.062	332	.485	.628
VBSJT	-1.06	332	-.543	.587

Note. OCS = Overexertion of control scale. UCS = Underexertion of control scale. VBSJT = Frontline™ video-based situational judgment test. M_D = Mean Difference.

Table 31
Validity Coefficients for VBSJT, OCS, and UCS by location

Location	N		VBSJT	OCS	UCS
Suburban	196	Overall	.261**	-.202**	.077
Urban	138	Overall	.443**	-.305**	.022

Note. ** = Correlation is significant at the .01 level (2-tailed).

Table 32

Summary of Hierarchical Regression of Job Performance on Overexertion of Control and Location Interaction

Predictor Variables	β	R^2	df	ΔR^2	Δdf
Step 1					
OCS	-.243*				
Urban	-.050	.057*	2		
Step 2					
OCS/Urban Interaction	.339	.064*	3	.007	1

Note. * = Significance at the .05 level. OCS = Overexertion of control scale.

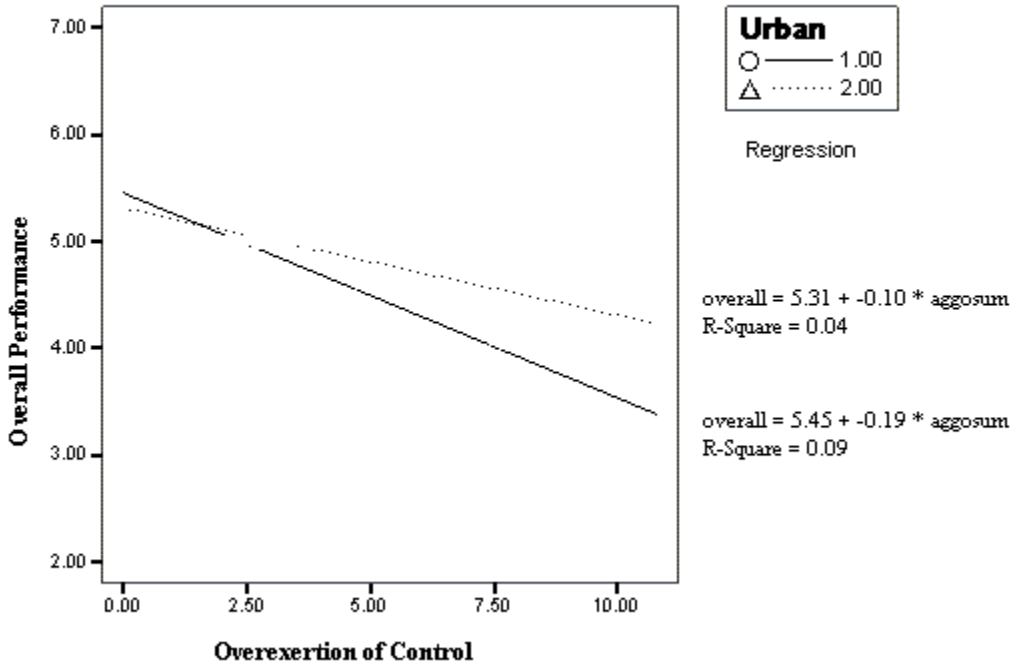

Figure 1. Graph of the location of law enforcement organization by overexertion of control interaction. Urban was dummy coded (Suburban = 1; Urban = 2).

Appendix A

Introduction for Expert Evaluators

Thank you for agreeing to participate in this research project. The purpose of the project is to investigate whether or not certain individual factors, in this case exertion of control, can be identified by looking at response patterns in tests. This type of research has not been done before and may provide information that is useful in understanding and addressing behavioral differences in work settings. For instance, analyzing circumstances of inappropriate exertion of control through use of a test is less costly than discovering the same thing through actual events in the field.

The test you will be looking at is a video based entry level test for law enforcement. There are 78 multiple choice questions in the test, but only 48 of them have at least one answer that relates to exertion of control. As a law enforcement professional, we are asking for your expert opinion in evaluating these 48 test questions.

One hypothesis that we wish to look at is whether or not there are certain triggers within a situation that are more likely to provoke inappropriate exertion of control. Provocation to over-exertion of control includes such things as the *focal character* to which the officer is responding is being rude, aggressive, or suspicious or the crime in question being particularly offensive. It also includes circumstances where an officer may assume resistance or escalation will occur. Provocation to under-exertion of control include such things as the *focal character* evoking sympathy. To that end, the first thing you will be asked to evaluate as you look at the questions is whether or not you find particular triggers to be present. Then you will be asked your opinion regarding the degree of overall provocation represented in the situation. Remember we are not asking how you would respond to the situation or whether you would be affected by the provocation. We are simply asking whether you view the situation as containing provocation. Provocation is here defined as behaviors or circumstances that may cause greater temptation to over- or under-exertion of control. Furthermore, some of the situations may contain multiple focal characters. If there are multiple focal characters, then mark the provocation exhibited by any of the characters. The character or characters in question are listed at the top of each item rating sheet. Be sure to identify the focal character before watching the item.

The third factor we ask you to rate is likeability of the focal character(s). Although most of the questions contain several characters, the answers usually revolve around one or perhaps a few. You will be asked how likeable you find the characters about whom the questions are asked. The questions are short and consequently you will have little exposure upon which to base your opinion, however, we are asking for your initial impression based only on the information you are given. Mark “no opinion” only if you have no positive or negative impressions on which to base an opinion. The character or characters in question are described at the top of each rating sheet.

The last thing you will be asked is your opinion on the appropriateness of the control that each answer represents. Certainly police work involves situations in which an officer must use police authority to exert control in order to accomplish the job. The amount of control necessary is

dependent upon the situation. Over-exertion of control exceeds the control necessary in the situation. Besides exerting more control than is necessary in a situation that actually requires some control, over-exertion of control also includes interventions into non-police matters and harsh or rude behavior that is not necessary and would not be tolerated but for the officer's position. Under-exertion of control is also relative. For instance, exerting less control than is necessary to properly manage a situation. Under-exertion of control can also be reluctance to enforce the law or investigate suspicious situations.

This is the scale you will be using to rate the exertion of control:

Extreme Under-Exertion of Control	Under-Exertion of Control	Appropriate Exertion of Control	Over Exertion of Control	Extreme Over Exertion of Control	Does Not Apply

Remember that exertion of control is relative to the situation. Extreme Under- or Over-Exertion of Control would be a response that in your opinion oversteps or does not meet the role and expectations of a police officer. Less extreme Under- and Over-Exertion of Control indicate that the officer may be technically within his or her discretion, however, in your opinion the officer's discretionary judgment could have been better.

Many of the answers are not related to control, even though they may be good or poor answers for other reasons. This research is concerned only with exertion of control. If an answer is not related to control, mark "Does Not Apply" regardless of whether or not it is a good answer.

In order to have enough time to appropriately fill out these forms, you will need to stop the video after each question. Feel free to rewind and review the questions as many times as you need to.

Thank you again for your assistance with the project. If you have any questions, call Carl Swander at 425-774-5700.

Ergometrics Viewing Agreement

During this presentation you will be viewing actual confidential test materials. To maintain our copyright and confidentiality assets we must require you to sign the following agreement. We at Ergometrics are strong proponents of improvements in the protection of test materials from damaging exposure and copyright infringement. Thank you for your cooperation.

Name *(Please Print)*_____

Confidentiality and Copyright Agreement

I understand the necessary confidentiality of the materials I will see in this session. I agree to refrain from any action that would jeopardize the security of these materials. I agree to maintain confidentiality by refraining from talking about the test with others who have not signed a confidentiality agreement. I also agree to respect the copyright of these materials by refraining from taking notes on test content or attempting to reproduce in any format the materials that will be shown to me.

Signed_____

Rater Information

Job Title: _____

Age: _____ years

Ethnicity (circle one):

African American
Hispanic
Asian or Pacific Islander
American Indian
Caucasian
Other: _____

Gender (circle one):

Male
Female

How long have you worked in law enforcement?

_____ Years _____ Months

How long have you worked in a command position (Rank of Sergeant or higher)?

_____ Years _____ Months

Question 2

Character(s) in question: The men in the park.

Circumstances that may provoke under or over-exertion of control	Yes	No	No Opinion or DNA
Was the character rude?	①	②	③
Was the character aggressive?	①	②	③
Was the character suspicious?	①	②	③
Was the character pleasant?	①	②	③
Was the character cooperative?	①	②	③
Did the character evoke sympathy?	①	②	③
Was the crime or alleged crime contemptible (socially repulsive)?	①	②	③
Were there complaints about the character from bystanders?	①	②	③
Was a warrant being enforced?	①	②	③
Has a law ever been broken?	①	②	③
Was there potential for others to become involved?	①	②	③

	Provocation for Under-Exertion of Control			Provocation for Over-Exertion of Control	
Degree of Provocation	**High**	**Moderate**	**No Provocation**	**Moderate**	**High**
Based on your overall impression, to what degree does this situation include provocation, or temptation, for over- or under-exertion of control?	①	②	③	④	⑤

Likeability	Strongly Dislike	Dislike	No Opinion	Like	Strongly Like
How do you think most people would feel about the character in question as he/she is presented in this situation?	①	②	③	④	⑤

POLICE HAVE UNIQUE, DISCRETIONARY AUTHORITY TO EXERT CONTROL OVER OTHERS. IN YOUR OPINION, TO WHAT DEGREE DOES THIS ANSWER REPRESENT APPROPRIATE EXERTION OF CONTROL IN THE SITUATION PRESENTED?

	Extreme Under-Exertion of Control	Under-Exertion of Control	Appropriate Exertion of Control	Over Exertion of Control	Extreme Over Exertion of Control	Does Not Apply
A. Greet them in a friendly way and ask how they are.	①	②	③	④	⑤	⑥
B. Kid them about yesterday.	①	②	③	④	⑤	⑥
C. In a serious way, ask them what they're doing.	①	②	③	④	⑤	⑥
D. Just ride by and make sure they have no visible alcohol.	①	②	③	④	⑤	⑥

Appendix B

Table B1. Demographics (Law Enforcement Experts)

Variable	Category	Frequency	Percent	Cumulative Percent
Age	32	1	10	11.1
	34	1	10	22.2
	38	1	10	33.3
	39	1	10	44.4
	45	3	30	77.8
	53	1	10	88.9
	56	1	10	100.0
	Missing	1	10	
	Total	10	100	
Gender	Male	7	70	70
	Female	3	30	100
	Total	10	100	
Ethnicity	African American	2	20	20
	American Indian	1	10	30
	Caucasian	7	70	100
	Total	10	100	
Tenure (Law Enforcement)	5-10 years	2	20	20
	11-20 years	3	30	50
	21-30 years	5	50	100
	Total	10	100	
Tenure (Command Position)	None	1	10	10
	1-5 years	3	30	40
	6-10 years	2	20	60
	11-15 years	1	10	70
	16-20 years	3	30	100
	Total	10	100	
Job	Commander	4	40	40
	Inspector	1	10	50
	Officer	1	10	60
	Sergeant	4	40	100
	Total		10	100

Table B2. Demographics (Validation Sample)

Variable	Category	Frequency	Percent		Cumulative Percent
Ethnicity	American Indian	3	.8	.9	.9
	African American	25	7.3	8.2	9.1
	Caucasian	256	68.9	78.2	87.4
	Hispanic	15	4.1	4.7	92.1
	Asian	24	6.2	7.1	99.1
	Other	3	.8	.9	100.0
	Missing	8	11.9		
	Total	334	100.0		
Gender	Male	277	84.7	84.7	84.7
	Female	57	15.3	15.3	100.0
	Total	334	100.0	100.0	

Table B3. Demographics (Validation Sample)

Variable	Category	Frequency	Percent	Cumulative Percent
Education	High School Graduate or GED	24	.4	1.3
	College, No Degree	427	7.7	24.7
	College with emphasis in law enforcement, No Degree	328	5.9	42.7
	2 year Degree	258	4.7	56.9
	2 year Degree in law enforcement	94	1.7	62.0
	More than 2 years College, No Degree	176	3.2	71.7
	4 year College Degree	238	4.3	84.7
	4 year College Degree in law enforcement	179	3.2	94.5
	Advanced Degree	100	1.8	100.0
	Total	1824	32.9	
	Missing	3602	67.1	
	Total	5426	100.0	
Ethnicity	American Indian	104	1.9	2.1
	African American	371	6.9	9.7
	Caucasian	3826	70.3	86.8
	Hispanic	292	5.4	92.8
	Asian	246	4.5	97.7
	Other	111	2.0	100.0
	Total	1	91.2	
	Missing	4951	8.8	
	Total	5426	100.0	
Gender	Male	4375	80.6	84.4
	Female	808	14.9	100.0
	Total	5183	95.5	
	Missing	243	4.5	
	Total	5426	100.0	

Carl Swander

104 Camelot Court
Blacksburg, VA 24060
(540) 961-3818
cswander@vt.edu

Education

2001 Ph.D. **Industrial/Organizational Psychology,** Virginia Polytechnic Institute and State University, Blacksburg, VA.
Dissertation: *Video-Based Situational Judgment Test Characteristics: Multidimensionality at the Item Level and Impact of Situational Variables*

1999 M.S. **Industrial/Organizational Psychology,** Virginia Polytechnic Institute and State University, Blacksburg, VA.
Thesis: *Assessing the Differential Functioning of Items and Tests of a Polytomous Employee Attitude Survey*

1997 B.S. **Psychology, minor in General Business,** Washington State University, Pullman, WA.

Related Experience

Consulting

Industrial/Organizational Psychologist
Ergometrics, Seattle, WA, 5/95-present.
Responsible for statistical analysis and customer relations. Helped Industrial/Organizational Psychologists with projects involving job analysis and selection procedures. Analyzed and organized data pertaining to employee satisfaction, validation, and 360 performance appraisals.

Teaching

Research Methods Instructor
Virginia Tech, Blacksburg, VA 8/99 – 12/00.
Taught multiple sections of research methods. Responsible for organizing lectures, tests and grading.

Undergraduate Advisor
Virginia Tech, Blacksburg, VA 8/98 – 5/99.
Advised approximately 400 undergraduate students for a major in Interdisciplinary Studies. Designed, implemented, and analyzed surveys to assess student satisfaction.

Graduate Teaching Assistant
Virginia Tech, Blacksburg, VA 8/97 – 5/98.
Taught 3 sections with approximately 30 students in each class. Responsible for organizing lectures, quizzes and grading

Research

Dissertation Research
Robert J. Harvey (Chair). Identified a unidimensional construct within a multidimensional VBSJT and explored the situational characteristics that contributed to the measurement of this construct.

Preliminary Examination Research
Two alternate formats (video and written) of a situational judgment test were compared in relation to the constructs measurement and validity.

Thesis Research
Robert J. Harvey (Chair). Analyzed and evaluated attitude survey data using item response theory and differential item functioning methods for polytomous data.

Independent Research
Robert J. Harvey (Chair). Designed and implemented a computerized adaptive version of a well known critical thinking appraisal.

Research Assistant
Washington State University, 1/95-5/95, 8/96-12/96
Conducted group decision-making experiments with introductory psychology students.

Conference Presentations

Swander, C. J. (2001). Video-based and written situational judgment tests: Method effects and construct validity. Paper presented at the 2001 annual conference of the Society for Industrial/Organizational Psychology, San Diego, CA, April 2001.

Swander, C. J. (2001). Exploring the Criterion Validity of Two Alternate Forms of a Situational Judgment Test. Paper presented at the 2001 annual conference of the Society for Industrial/Organizational Psychology, San Diego, CA, April 2001.

Donovan, J. J. & **Swander**, C. J. (2001). The impact of self-efficacy, goal commitment, and conscientiousness on goal revision. Paper presented at the 2001 annual conference of the Society for Industrial/Organizational Psychology, San Diego, CA, April 2001.

Swander, C. J., Spurlin, O. & Swander, C. (2000). Using video to enhance training and selection. Presented at the 2000 IPMAAC conference, Washington D.C.

Swander, C. J. & Harvey, R. J. (2000). Assessing the Differential Functioning of Items and Test of a Polytomous Employee Attitude Survey. Paper presented at the 2000 annual conference of the Society for Industrial/Organizational Psychology, New Orleans, LA, April 2000.

Spurlin, O. & **Swander** C. J. (2000). Ergonomic Principles and the Development of Physical Ability Standards. Paper presented at the 2000 annual conference of the Society for Industrial/Organizational Psychology, New Orleans, LA, April 2000.