5-2009

The Validation of a Situational Judgment Test to Measure Leadership Behavior

Kaci Lyn Grant
Western K, kaci.grant3785@gmail.com

THE VALIDATION OF A SITUATIONAL JUDGMENT TEST TO MEASURE LEADERSHIP BEHAVIOR

A Thesis
Presented to
The Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Arts, Industrial/Organizational Psychology

By
Kaci Lyn Grant

May 2009

THE VALIDATION OF A SITUATIONAL JUDGMENT TEST TO MEASURE LEADERSHIP BEHAVIOR

Date Recommended__May 6, 2009_____

___Betsy Shoenfelt_____
Director of Thesis

___Reagan Brown_____

___Anthony Paquin_____

_____
Dean, Graduate Studies and Research Date

TABLE OF CONTENTS

THE VALIDATION OF A SITUATIONAL JUDGMENT TEST TO MEASURE LEADERSHIP BEHAVIOR

Kaci Lyn Grant May 2009 76 Pages

Directed by: Drs. Elizabeth Shoenfelt, Reagan Brown, and Anthony Paquin

Department of Psychology Western Kentucky University

Assessment centers, although useful for assessing behaviors and competencies associated with a targeted construct, can be low in economic utility. The current study sought to validate a situational judgment test (SJT) that was developed as an alternate form of assessment for a leadership development program. The first study examined the content validity of the SJT by performing retranslation on item stems and calibration of the item responses. The second study examined alternate forms reliability between the two forms of the leadership SJT that were developed. The third and final study evaluated the relationship between assessment center performance scores and SJT scores by demonstrating their convergent validities. Results from Study 1 demonstrated that the SALSA© test was a content valid measure of leadership ability. Results from Study 2 demonstrated that all available items from SALSA© could be used to make two forms of the test that demonstrate good alternate forms reliability. Finally, Study 3 suggests a moderate correlation between the assessment center and situational judgment test. Future research should focus on the underlying issues pertaining to significant group differences between English as primary language and English as second language students. Alternate developmental procedures, especially with alternate form assignment, should also be considered.

The Validation of a Situational Judgment Test to Measure Leadership Behavior

Western Kentucky University provides a unique opportunity for their students in offering a Certificate in Leadership Studies through the Center for Leadership Excellence (CLE). This program combines education in ethics, social responsibility, and core leadership theory to enable students to become knowledgeable about the field of leadership and its practice today. Students enrolled in the program are given the opportunity to participate in a leadership skills assessment center.

The assessment center was developed in 2006 and has become very popular over the past few years. Despite the success of the assessment center, the CLE would like an alternate form of assessment to be available for their students. The assessment center is a valuable tool for leadership development. However, the growing number of students enrolling in the certificate program is making it more difficult to provide the services of the assessment center to each of these students because of reasons specific to the university such as time, cost, and lack of resources. The university has increased their emphasis on distance learning, which likewise makes it difficult for off-campus students to participate in the assessment center.

In addition, assessment centers have come under much scrutiny (for a review, see Lance, 2008). Problems include exercise effects, rater biases, scoring methods and realism among tasks. Research has also suggested that assessment centers cannot measure complex constructs such as leadership (Lowry, 1995). However, a study by Waldman and Korbar (2004) proves that student assessment can be beneficial and can also predict future success. The authors found that scores on an academic-based assessment center for undergraduate business students were able to predict both intrinsic and extrinsic aspects

of career success. They also found that the assessment center was a better predictor of early career success than student GPA. Research of this kind proves that student assessment is an important and worthwhile investment.

An alternate mode of measuring the dimensions of performance assessed in the assessment center would be beneficial so that Leadership Studies could offer some sort of appraisal and feedback opportunity to all of their students. It is for this reason that the CLE has enlisted the help of the director of the university's Industrial/Organizational Psychology Masters program to develop a situational judgment test (SJT). The SJT, which will be a paper-and-pencil/computer-based format, is easy to administer to students on and off campus, is cost efficient, and will be developed to measure the specific dimensions in the assessment center.

This paper will review the current assessment center model being used by the university, along with a brief overview of its development. Issues with assessment centers will also be discussed. The paper will then review the available literature on situational judgment tests. History, validity, development, and special issues such as response instructions and scoring will be covered. A review of the development of the leadership SJT will then be addressed.

The current studies seek to validate the SJT to ensure it is a psychometrically sound measure of leadership ability. The first study will evaluate the content validity of the SJT through a retranslation of the items. This study also addresses calibration of response options. The second study involves alternate forms reliability. Because students participate in an entry and exit assessment center while in the certificate program, it is important that two forms of the test be available. Last, the third study will look at

convergent validities between assessment center dimension scores and SJT dimension scores.

Overview of Assessment Center

A steering committee for the CLE was formed to develop the assessment center (Ashburn & Love, 2006). The committee first identified nine core competencies of effective leaders: Problem Solving and Innovation, Influencing Others, Verbal/Non-Verbal Communication, Team Skills, Visioning and Planning, Results Orientation, Knowledge of Leadership Theories, Written Communication, and Self-Analysis and Improvement. Seven of these competencies were identified previously in a meta-analysis of assessment center dimensions (Arthur, Day, McNelly, & Edens, 2003). Behavioral checklists were then developed for each dimension to provide assessors with key behaviors to represent effective, average, and ineffective leadership behavior in each exercise. After the competencies were identified and defined, the committee developed exercises that were specific to the targeted behaviors of the competencies. The checklists and targeted behaviors were then used to create behaviorally anchored rating scales (BARS) to be used for assessment center ratings for the participants. Last, assessors were required to take part in frame of reference training. This type of training was chosen because the assessors needed a common understanding of the standards used for rating the participants, and because leadership is inconsistently defined in the literature. Along with the frame of reference training, assessors also receive behavior observation training, by watching assessment center exercises on tape, and performance dimension training, by reviewing the definitions of the competencies (Woehr & Huffcutt, 1994).

Before participating in the live assessment center, students complete a leadership theories knowledge test and write a problem-solving essay via an electronic campus communication tool. These exercises are used to provide scores in Knowledge of Leadership Theories, Written Communication, and Visioning and Planning. Because of the method used to administer these exercises, some students do not complete the preliminary steps to the assessment center. In the live assessment center, students first complete an individual oral presentation and then participate in a leaderless group discussion. The last two exercises in the assessment center are team based. Students are rated by two assessors on each exercise and special care is taken so that each student is rated by as many different assessors as possible throughout the assessment center. This is consistent with Guion's (1998) suggestion of a 2:1 assessor to participant ratio. It also helps cut down on rater biases. To complete the assessment center, students fill out a self-rating form so they can compare their opinions of their performance with the scores given by the assessors. If a student completes the leadership certificate program, he or she will typically participate in an "entry" assessment center at the beginning of the program and an "exit" assessment center at the end of the program. This is done not only to enable them to see their own personal growth, but also as a form of feedback for the certificate program.

Based on the response of students eligible to participate, and teachers from leadership classes, it is obvious that opinions of the CLE assessment center have been favorable in the past and many students find great value in the feedback given to them. However, it is no longer feasible to assess all of the interested undergraduate and graduate students due to an increase in the number of Leadership Certificate students and

decreases in time, money, and other resources. There also is no other option for those students who cannot attend the assessment center. These disadvantages are consistent with those mentioned in the literature (Joiner, 2002).

There are other disadvantages to the assessment center model. For example, there may be a lack of realism in the exercises (Howard, 2008). In the CLE assessment center, it is not likely that students will encounter the team exercises again, and it may not be as helpful to learn how they performed in that particular exercise. In other words, the student may be particularly good at the puzzle or problem presented in the exercise, but it may not be something used on a daily basis in leadership. Therefore, the more comfortable the student is with the exercise, the more likely they will perform well. One of the exercises is the "Blind Puzzle" where the students are blindfolded and are asked to work together to complete a puzzle. Students receive scores in competencies such as Team Skills and Verbal/Non-Verbal Communication, but the context in which they receive these scores may not be applicable to other situations. On a similar note, the exercises may lack face validity (Moses, 2008). If the participants do not see the value of the exercise, they may not perform at maximal levels, which may influence the accuracy of the ratings they receive.

The accuracy of the ratings lies in the hands of the assessor, participant, and the design of the exercises. Assessment center ratings are subject to rater biases, even if raters are trained to avoid such errors (Moses, 2008; Lance, 2008). Rating errors include halo error, leniency error, and severity error, among others. Ratings also may suffer from lack of interrater unreliabity (Connelley, Ones, Ramesh, & Goff, 2008). Interrater reliability usually improves with experience and refresher training. Because graduate

students in the I/O Psychology program serve as the primary source for assessors, there is a high turnover rate (50% annually) precluding raters with more than two years of experience. Some may argue that using graduate students as assessors is not as effective as using trained psychologists in attaining accurate and reliable ratings (Lowry, 1995). On the other hand, Borman (1978), who developed a performance appraisal model that attempts to explain the cognitive processes involved in establishing performance ratings, argues that graduate students can provide ratings as accurately as practicing I/O Psychologists if the right training is used.

Other problems include the scoring and the exercises themselves. If development procedures are followed correctly, an assessment center can successfully measure the intended skills and constructs. However, accurate measurement tends to be more difficult to achieve in assessment centers used for developmental purposes. Participants may act differently in assessment centers used for development and assessment centers used for hiring or promotion. Participants likewise may demonstrate inconsistent behavior across the assessment center exercises because of exercise effects (Lance, 2008; Brannick, 2008). In other words, participants may be able to perform well as a function of the exercise or tasks they have to complete rather than as a function of KSAs. The resulting rating may not be a true measure of their typical or maximal ability. Another source of error may lie in the fact that exercises can cause assessors to measure things they are not trying to measure (Arthur, Day, & Woehr, 2008). Lievens (2008) distinguishes between "incidentals" and "radicals." Incidentals are those characteristics of an exercise that do not determine actual performance and are simply surface characteristics. Radicals, on the other hand, determine performance and are the structural characteristics of the exercises.

Slight variations in both the incidentals and the radicals of the exercises can affect performance.

In conclusion, when used for developmental purposes, assessment centers can be very valuable tools but have potential flaws. To fix these problems there are two options: redesign the assessment center to fix the mentioned problems or choose a different method of measuring the leadership dimensions. Because the fix has to be economically feasible for the organization (Jones & Klimoski, 2008), a different method of measurement, a SJT, was developed. Specifically, a SJT was developed to measure the exact dimensions being measured by the assessment center. SJTs have a long history and have also been used as a type of exercise in assessment centers (Lowry, 1995). It was the opinion of the development team and CLE that the SJT will address many of the problems, both those specific to the university and those characteristics of ACs in general identified empirically in the reviewed literature.

Overview of Situational Judgment Tests

SJTs present hypothetical but realistic situations and are intended to measure typical or maximal performance of a certain construct. Test takers indicate how they would respond to the given situation (Lievens, Peeters, & Schollaert, 2008; McDaniel & Nguyen, 2001; Weekley & Ployhart, 2006). Mostly used for hiring and promotion, SJTs are used to predict how applicants will respond to job-related situations. Because they can be developed to measure a variety of constructs, SJTs can be used to predict how someone would act in virtually any situation, including leadership situations. Examples of leadership SJTs include the Leadership Evaluation and Development Scale (Mowry, 1957) and the Leadership Skills Assessment (Bergman, Drasgow, Donovan, Henning, &

Juraska, 2006). Regardless of the targeted construct, most SJTs measure interpersonally oriented skills and tacit knowledge. Tacit knowledge is the ability to solve problems faced in the real world and is gained through experience (Weekley & Jones, 1999).

The use of SJTs dates back to the 1920s and, as with assessment centers, were popularized by the military and civil services (O'Connell, Hartman, McDaniel, Grubb, & Lawrence, 2007; Lievens, et al., 2008). These early SJTs were used to predict the reactions of military personnel to problematic situations and to provide a realistic preview to those interested in civil service. Use of the tests lowered attrition rates among new officers. In the 1940s, SJTs were developed to measure potential in supervisors and, in the 1950s and 1960s, managerial success (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Since then, SJTs have been used to predict job performance for a variety of positions such as labor supervisors and entry-level managers, and to help identify training needs.

A renewed interest in SJTs was prompted by Motowidlo, Dunnette, and Carter (1990) when they developed what they called a low-fidelity simulation. Fidelity, more specifically physical fidelity, refers to how a test or simulation represents a stimulus event to elicit a response. Physical fidelity increases when the situation uses very realistic materials, equipment, and environments and when applicants can respond exactly as they would respond to the situation in the transfer setting. The assessment center is an example of a high-fidelity simulation as it involves an environment where the participants role play hypothetical leadership situations. On the other hand, with a low-fidelity simulation, such as the SJT, the situation is a written description rather than experiential and, as such, does not allow applicants actually to demonstrate how they would respond.

However, SJTs are considered to have high psychological fidelity. Psychological fidelity refers to a stimulus that represents the same psychological demands as the transfer setting. Test takers have to have the experience, knowledge, or skills needed to know how to respond to the hypothetical situation. The SJT scenarios can be highly specific to the job or position.

There are three consistent features of SJTs that make them low physical fidelity simulations (Lievens et al., 2008). First, the tests present realistic situations unique to the construct being measured. This is typically done by a written description; the physical fidelity can increase slightly with the use of video-based presentation. The scenarios presented represent real situations that the test taker may experience in the job or position. Second, the responses are presented in a multiple-choice format, which can also be written or video-based. Participants are usually given four to five options to choose from, but the way in which they are supposed to respond can vary. For example, they may be asked what they "would do" or what they "should do." They could also be asked to rank the effectiveness of the responses or choose both the most effective and least effective responses. Last, because of the format of the simulation, assessors are not needed. There are no behaviors to evaluate. The scoring key is developed a priori, either empirically or by subject matter experts (SMEs). This may seem like a disadvantage, but it has been proven that low-fidelity simulations can predict performance just as well as high-fidelity simulations (Motowidlo et al., 1990). Psychometric properties of SJTs will be discussed next.

Validity, Reliability and Utility of SJTs

Because SJTs are mostly used in job settings, many studies have been completed to establish their criterion-related validity. SJT scores have been demonstrated to correlate with job performance, cognitive ability, and the Big Five factors of personality, among other things. For example, Motowidlo, et al. (1990) found that scores on a SJT for entry-level managers correlated from .28 to .37 with supervisory ratings of performance. McDaniel and Nguyen (2001) found significant correlations between a situational judgment test and emotional stability (r = .31), agreeableness (r = .25), and conscientiousness (r = .26). SJTs have been proven to have useful levels of validity when predicting job performance (p = .34) and a strong relationship with cognitive ability (p = .46), which lends support for the continued use of SJTs (McDaniel et al., 2001). Most relevant to the current research, significant correlations have been found between supervisors' SJT scores and their assessment center performance ratings (Wagner, 1987 as cited in Weekley & Jones, 1999). O'Connell et al. (2007) demonstrated that SJTs can add incremental validity to a test battery. They found that the SJT added a .03 validity increment to a cognitive test and a .04 validity increment to a battery of five personality predictors.

Along with the wealth of data about validity, researchers have presented potential antecedents to performance on SJTs and relationships with different abilities and constructs. For example, Weekley and Ployhart (2005, 2006) suggested that cognitive ability may be related to job performance and scores on SJTs. In other words, more intelligent people may perform better on SJTs because they are able to deduce the appropriate responses. The authors also suggested that personality and experience are

potential antecedents to performance on SJTs. McDaniel, Hartman, Whetzel, and Grubb (2007) demonstrated in their meta-analysis that the type of response instructions influenced how the test measured the constructs. Tests with knowledge instructions (i.e., should do) correlated stronger with cognitive ability while tests with behavioral tendency instructions (i.e., would do) had stronger correlations with personality constructs. Yet another moderator discussed in the literature is whether a job analysis was used to develop the test. McDaniel et al. (2001) found that tests based on a job analysis demonstrated higher validity. In summary, it may be more difficult than expected to get a clear picture of the criterion-related validity of SJTs, and it may depend on things such as response options and the use of a job analysis. However, there is empirical support in favor of their use and SJT validity is comparable to other assessment methods.

Regarding SJT reliability, the meta-analysis by McDaniel et al. (2001) reported internal consistency coefficients that ranged from .43 to .94. These coefficients were moderated by both length of the SJT (with longer SJTs being more reliable) and type of response instructions. However, to assess reliability, it is suggested that test-retest or alternate forms be used, especially if the test is multidimensional (O'Connell et al., 2007). In doing so, test-retest reliabilities in one of their studies ranged from .77 to .89. In other words, much of the research reports internal consistency coefficients, but test-retest or alternate forms reliability may be more appropriate estimates of reliability. In conclusion, O'Connell et al. (2007) found that test-retest results are satisfactory, especially for longer tests.

Last, utility is an important consideration when deciding when the use of a SJT is appropriate. It has been demonstrated that SJTs have satisfactory criterion-related validity

and incremental validity. However, potentially more important to organizations using the SJT for developmental purposes is the economic utility. Research testing the economic utility of SJTs is nonexistent, but the use of a SJT has clear monetary advantages (Lievens et al., 2008). First, the SJT can be given to large groups of participants in a paper-and-pencil or computer-based format. Second, because there are no behaviors to observe, assessors and assessor training are not needed. Third, when compared to assessments such as job samples or assessment centers, the low physical fidelity of SJTs does not require equipment or realistic settings. Last, the costs of developing an SJT are comparable to, but usually lower than, the costs of developing high-fidelity simulations. The development of SJTs will be covered next.

Development of SJTs

The development of SJTs relies heavily on the critical incidents technique (Lawshe, 1975), and most research follows the approach from Motowidlo et al. (1990). The authors, who developed a low-fidelity simulation for general management performance, started by reviewing several job analyses for the position. They then met with and interviewed incumbents and supervisors as SMEs in order to collect effective and ineffective examples of managerial performance (i.e., critical incidents). They did not specify competencies to be used for these examples. The authors used the critical incidents to write brief scenarios. The scenarios were then presented to a new group of incumbents who wrote short descriptions on how they would respond to the situation. The authors then wrote alternate responses for each situation. Last, they used an experienced group of executives to rate the effectiveness of the responses.

Although most researchers use that particular approach, some variations may be considered. For example, some SMEs are directed to write items for specific competencies (McDaniel & Nguyen, 2001). The origin of the stem may be determined by a job analysis or simply experiences on the job. Another issue is the complexity of the item stem (Weekley & Ployhart, 2006). Some research on this issue suggests that the more complex the item stem, the more valid it is because it is a more realistic sample of the job or position. However, there are mixed results. McDaniel et al. (2001) found that more detailed SJTs demonstrated lower criterion validity than more general SJTs. The complexity of item stems may have implications when reading level is important. In addition, complexity may impact the psychological fidelity of the item and whether successful performance on the test actually requires the KSAOs needed for the job or position (Weekley & Ployhart, 2006). This is because the more detail that is included in the situation, the more complex and similar to the real job the situation will be. If the situation depicted is much like the real situation, test takers will be required to demonstrate greater strength in the skills needed for the job.

Finally, there are many options when trying to calibrate response options. SMEs can rate the effectiveness of the responses, determine the most and least effective responses, or options may be empirically correlated with a criterion (Ployhart & Ehrhart, 2003). Although the literature provides a variety of methods, research fails to indicate the most effective development procedures. Once the test is developed, there are other issues that also need attention, such as response instructions and scoring.

Special Issues

After the final version of the test is complete, the developer must decide which instructions should be used and which scoring method would be best. Both can have great effects on validity and reliability.

SJT instructions can elicit different types of responses (Ployhart & Ehrhart, 2003). Some SJTs may ask the participant to select the most effective response, or both the most effective and least effective responses. Other tests have instructed participants to rate the effectiveness of the responses.

Furthermore, the wording of the instructions has important implications. Ployhart and Ehrhart (2003) grouped the different types of response instructions into two categories: "would do" instructions and "should do" instructions. "Would do" instructions include asking the participant to indicate what they would do or what they have done, or what they are most and least likely to do. "Should do" instructions include asking the participant to indicate what they should do, the most effective response, or the best response. They may also ask respondents to identify the best and worst responses, the best and second best responses, or to rate the effectiveness of each response. "Would do" instructions are considered behavioral tendency instructions and "should do" instructions are known as knowledge instructions (McDaniel et al., 2007). SJTs with knowledge instructions measure maximal performance. They assess how the participant performs at optimal levels and give a measure of ability. Other examples of assessments that measure maximal performance are cognitive ability tests, job knowledge tests, or work sample tests. On the other hand, SJTs with behavioral tendency instructions measure typical behavior. These measures have larger non-cognitive correlates and are

similar to personality tests. The authors note that behavioral instructions are more susceptible to self-deception and impression management.

McDaniel et al. (2007) found stronger correlations between knowledge instruction SJTs and cognitive ability than behavioral tendency SJTs and cognitive ability. However, correlations between behavioral tendency SJTs and personality factors were higher. There were no differences in criterion-related validity between the two types of instructions.

Another issue to consider is the scoring of the test. Bergman et al. (2006) identified six different scoring methods. The first, empirical scoring, derives scores from the relationship between the item options and a criterion measure. They have been found to have high validity coefficients, but the outcome depends on the quality of the criterion. Becker (2005) used this method for his employee integrity SJT by dummy-coding participants' responses and correlating them with integrity ratings. The second method, theoretical scoring, reflects theory that is related to the construct being measured and helps determine which answers are the most and least effective (Bergman et al., 2006). This type of scoring depends on the underlying fundamental components of the theory, which may make this method more susceptible to faking. The next method, hybridized scoring, combines two independently generated keys to potentially increase predictive power. Fourth, expert-based scoring is where the scoring key is based on responses from SMEs or from the comparison of responses between novices and experts. This method requires that a decision rule be implemented beforehand. When the correct answer is identified, it is scored as 1 point, while choosing any other choice results in 0 points. Fifth, factorial scoring is based on factor analysis and item correlations. This method is

generally used when a test is not set to measure a certain construct, yet the test can eventually produce meaningful constructs. Last, subgrouping identifies patterns or groupings according to responses on biodata items, and is infrequently used.

Of the six scoring methods, expert-based scoring and empirical scoring are used most frequently (Lievens et al., 2008). Once the scoring method is chosen, researchers must also decide how to assign scores. For example, some may give 1 point for a correct answer and 0 points for all other answers. Other SJTs assign a -1 point value if the participant chooses the least effective answer as the most effective. It also depends on the response instructions. For those SJT items that ask participants to rank the effectiveness of the responses, a special scoring key must be determined (for a review of options that have been used in the past, see Weekley & Ployhart, 2006).

Summary

The review of the literature indicates that situational judgment tests can address many problems that are associated with an assessment center. As with assessment centers, SJTs have been in use since the 1920s and both can be used to assess a number of constructs. Even though SJTs are considered low-fidelity simulations and assessment centers are considered high-fidelity simulations, there appears to be little difference in their ability to predict performance. SJTs have an advantage over assessment centers in that scoring is determined a priori and they do not require the use of assessors to rate behaviors. Therefore, rater errors and rater agreement are not concerns when using SJTs. SJT validity and reliability have proven to be satisfactory, and the utility of SJTs is superior in most situations. Economic utility was of particular interest to the CLE, and factored greatly into the decision to develop and use the SJT with the Leadership

Certificate students. The development of the SJT, which emphasized the dimensions used in the assessment center, will enable the CLE to administer the test to all of their students, rather than only a subset. The SJT will save money and also will reduce the amount of time needed to provide students with feedback on their leadership skills.

Leadership Situational Judgment Test Development

The SJT was developed to assess the seven assessment center dimensions identified by Arthur et al. (2003). In addition, an eighth dimension, Integrity/Ethics, was targeted by the SJT. Because the SJT is to be used as an alternate form of assessment for leadership development, it is important to note that six of the eight SJT dimensions correspond to six of the nine dimensions used in the CLE assessment center: Problem Solving and Innovation, Influencing Others, Verbal/Non-Verbal Communication, Team Skills, Visioning and Planning, and Results Orientation. The CLE assessment center dimension of Knowledge of Leadership Theories was not measured because it is in the form of a paper-and-pencil test. The CLE assessment center dimension of Written Communication was not measured because of the nature of the SJT. The CLE assessment center dimension of Self-Analysis and Improvement was not used because it serves as a way for the students to compare their thoughts on their performances in the simulations to those of the raters in the assessment center and, as such, was not amenable to the SJT format. The SJT targeted two dimensions not included in the CLE assessment center: Tolerance for Stress and Integrity/Ethics. In Arthur et al.'s 2003 meta-analysis identifying the dimensions most frequently observed in leadership assessment centers, Stress Tolerance was the only dimension not assessed in the CLE assessment center.

Thus, the SJT will measure the seven dimensions identified in the Arthur, et al. meta-analysis and an eighth dimension of Integrity/Ethics.

The first step in developing the SJT was to generate critical incidents (Lawshe, 1975) from SMEs. This method is consistent with the SJT development approach described by Motowidlo et al. (1990). SMEs were provided definition of these dimensions (see Appendix A) and were asked to write short descriptions of good, bad, and average leadership performance (i.e., critical incidents) for one of the eight dimensions of leadership. SMEs also wrote three to four responses to each situation. Three critical incident workshops were facilitated by students enrolled in the WKU Industrial/Organizational (I/O) Psychology Masters program. These graduate students received training prior to the workshops. SMEs utilized in the workshops included 28 Cadets from WKU's ROTC program, 11 advanced students from the Dynamic Leadership Institute (DLI) program, and 14 students in an honors section of Effective Leadership Studies. These students qualified as SMEs because of their knowledge of and familiarity with leadership concepts and theory. Students were used as SMEs because the target audience for the SJT is students enrolled in the Leadership Certificate Program. It was expected that the use of student SMEs would help ensure the situations would be appropriate for students. SMEs generated the critical incidents and responses (see Appendix B). This differs from the Motowidlo et al. (1990) approach in that the same SMEs were used to generate both the scenarios and the responses. However, a similar approach was used by Weekley and Ployhart (2006).

During each workshop, SMEs were divided into eight teams; one dimension was assigned to each team. Facilitators ensured that the definitions of the dimensions were

clearly communicated. After generation of the critical incidents and responses, the facilitators were responsible for collecting, editing, and organizing the critical incidents into a spreadsheet. An I/O Psychologist performed a final edit of each of the 300 incidents to ensure each incident met the specifications needed for the SJT and to ensure that each situation was written in the same format. ROTC Cadets generated 126 critical incidents, DLI students generated 108 critical incidents, and honors students generated 55 critical incidents.

In total, across the three workshops, over 50 undergraduate students participated as SMEs and a total of 289 critical incidents were generated, with at least four response choices for each. Additional critical incidents were developed by I/O graduate students to bring the total number of critical incidents to approximately 300.

The second and third steps in the development of the SJT are described in more detail in subsequent sections of this paper describing Study 1 and Study 2. The second step of the process, retranslation, ensured the incidents were clear examples of the targeted leadership dimension. The third step of the process was to calibrate each of the response alternatives in terms of leadership effectiveness. Response instructions and the scoring key were then developed. The SJT items were later assigned to one of two forms of the test, as described in the section on Study 2. The test was put into a platform that allowed electronic administration. The SJT was named Situational Assessment of Leadership: Student Assessment © (SALSA©; Shoenfelt, 2009). Informed consent of participants was acquired through a message included on the first page of the SALSA© website informing participants that completing SALSA© implied informed consent. The WKU Human Subjects Review Board approval form may be found in Appendix C.

Completion of all 130 SALSA© items took approximately one hour. Students were instructed to select the option that represented the most effective leadership behavior for the situation identified in each item. The test was scored by awarding one point for a correct answer and zero points for an incorrect response. Dimension scores were obtained by summing the correct responses for a given dimension. A total test score was obtained by summing the total number of correct responses across all dimensions.

The Current Research

The current research is part of three studies evaluating the leadership SJT as an alternate form of assessment for the CLE's leadership development program. Although the second and third studies are the focus of this thesis, the first study will be described as it laid the foundation for the focal studies. The first study examined the content validity of the test through retranslation of the items and calibrated the response options for each item on the SJT. The second study assessed alternate forms reliability of two forms of the leadership SJT. Finally, the third study evaluated the relationship between assessment center performance scores and SJT dimension scores by examining convergent validities.

Study 1 Overview

In Study 1, the critical incidents generated in the SME workshops were retranslated (Smith & Kendall, 1963) by a different group of six SMEs to determine if they were measuring their intended dimension of leadership. Items were combined across dimensions and listed in random order. After reading the definitions of the eight dimensions, the SMEs assigned each critical incident to the dimension it best represented. Items were retained only if agreement was demonstrated across SMEs in terms of the dimension the item represents. The retranslation process ensured that each retained item represented a given dimension of leadership.

A different group of six SMEs completed the calibration step. For calibration, those items surviving retranslation were grouped by dimension in a database. The SMEs read each situation along with four response options and rated each response option on a 5-point scale. The mean rating reflected the level of effectiveness of a response option. The calibration process ensured that the response used as the correct answer on the test is consistent with the option experts rated as the most effective response.

Study 1 Method

Retranslation

Participants

Six faculty in the disciplines of Industrial/Organizational Psychology, Business, Leadership Studies, and Military Science who are knowledgeable about the field of leadership served as SMEs. Three were female and three were male, with an average age of 52 years (*SD* = 5.33). Four of the SMEs reported receiving graduate training in leadership. The six SMEs reported an average of 16 years experience in teaching leadership (*SD* = 12.68). Three SMEs held Masters degrees; the other three held Ph.D.s.

Procedure

The 300 critical incidents (situations) were combined across all dimensions and listed in random order in an Excel worksheet. The file was then sent by e-mail to the SMEs, along with instructions for completing the retranslation. The SMEs were provided with definitions of the eight dimensions to assist them in assigning each critical incident to a dimension. SMEs were instructed first to read the definition of each dimension and then to read each situation and assign each critical incident to the dimension the incident best represented. An inclusion criterion of 66.7% SME agreement on the dimension for an item resulted in 106 items surviving retranslation. To retain additional items, the criterion was lowered to 50% agreement, resulting in 213 items retained. Table 1 contains the number of items retained for each dimension.

Table 1
Number of Items Retained in Each Dimension After Retranslation

Dimension	Number of Items Retained
Organizing/Planning/Visioning	31
Consideration/Team Skills	32
Problem Solving/Innovation	36
Influencing Others	18
Communication	21
Drive/Results-Orientation	35
Tolerance for Stress	18
Integrity/Ethics	22
Total	213

Calibration

Participants

Six faculty in the disciplines of Industrial/Organizational Psychology, Business, Leadership Studies, and Military Science who are knowledgeable about the field of leadership served as SMEs. There were two female and four male SMEs. Their average age was 49.33 years (*SD* = 5.47). All six SMEs reported receiving graduate training in leadership and averaged 17.67 years experience in teaching leadership (*SD* = 7.31). Two SMEs held Masters degrees, and the other four held a Ph.D.

Procedure

Items that survived retranslation were grouped by dimension and put into an Excel database containing a separate worksheet for each dimension. The definition of the dimension appeared at the top of each worksheet; the relevant items (situations) and four response options (i.e., descriptions of behavioral responses to the situation) per item appeared below the dimension definition. SMEs were directed first to read the definition of the dimension then to read each situation and the four response options. SMEs then rated each response option on a 5-point scale of Leadership Effectiveness (*1* = Extremely Ineffective Leadership Behavior, *2* = Ineffective Leadership Behavior, *3* = Somewhat Effective Leadership Behavior, *4* = Effective Leadership Behavior, *5* = Extremely Effective Leadership Behavior).

The mean of the SME ratings was used to indicate the effectiveness of the behavior described by the response option. Only items with at least one response rated as "Effective" or better were retained, ensuring that there would be a correct response to each item. In addition, only items where the best answer was at least .5 better than the next best answer were retained, ensuring that there would be only one best answer. An exception to this rule was made for seven items included for which the best answer was only .33 better than the next best answer; these exceptions helped ensure an adequate number of items for the dimensions of Organizing/Planning/Visioning, Consideration/Team Skills, Influencing Others, Drive/Results-Orientation, Tolerance for Stress, and Integrity/Ethics. These decision rules eliminated 83 items either because there was no effective response or because there was more than one equally effective best answer. Some 130 items were retained after the calibration process. Table 2 indicates the

number of items retained for each dimension following the calibration process. The calibration process ensured that the response keyed as the correct answer on the SJT is consistent with the opinion of the leadership experts as the most effective response.

Table 2
Number of Items Retained for Each Dimension After Calibration

Dimension	**Number of Items Retained**
Organizing/Planning/Visioning	18
Consideration/Team Skills	21
Problem Solving/Innovation	19
Influencing Others	11
Communication	12
Drive/Results-Orientation	25
Tolerance for Stress	11
Integrity/Ethics	13
Total	130

Study 2 Overview

Study 2 examined the alternate forms reliability of two forms of the leadership SJT. Two forms of the SJT were developed because participants in the assessment center usually complete the assessment center twice, once at the beginning of the Leadership Certificate Program and again after they fulfill the requirements for the program. Assessment center pre and post feedback given to the students enables them to determine if their leadership skills have changed during the course of the program. These data also help the CLE determine the strengths and weaknesses of the Leadership Certificate Program. Thus, it was desirable to have two alternative forms of the SJT. Equivalent forms of the SJT used pre and post to evaluate participation in the Leadership Certificate Program would enable students to determine if their leadership skills have changed over the course of the program. Suggestions by O'Connell et al. (2007) were followed for alternate forms reliability.

Two forms of the SJT were created by splitting the test items to include an equal number of items of each difficulty level in each dimension on each form. The coefficient of equivalence was computed and the following hypothesis was tested:

H1: There will be a positive correlation between the scores on the two forms of the leadership SJT (overall and for each dimension).

Study 2 Method

Participants

A total of 61 students (56 graduate students, 4 undergraduate seniors, and 1 doctoral student) participated in this study. Thirty-three participants were female and 28 were male. The mean age of the participants was 29.51 years (*SD* = 9.39). Of this sample, 42 were Caucasian, 14 were Asian, 2 were other, 1 was African American, 1 was Hispanic, and 1 was non-resident alien. Forty-three students reported English as their primary language and 18 reported English as their second language. Fifty-nine percent of the participants had completed or were currently enrolled in LEAD 200, 400, 500 or 600 (i.e., leadership (LEAD) courses in the CLE Leadership Certificate Program); 20 percent were enrolled in the Leadership Certificate Program. Those who had not completed a LEAD class had completed a graduate level Organizational Psychology class. Thus, all participants had completed some formal coursework on leadership.

Procedure

The 61 participants completed all 130 items on the SJT. Response data and SME data from the response calibration process were used to create two equivalent forms of the SJT. Data from the calibration step were used to calculate the difference in mean ratings for the best and next best response option for each item. Items where the mean difference was .5 or less were considered difficult items; items where the mean difference was between .5 and 1.0 were considered of moderate difficulty; and items where the mean difference was greater than 1.0 were considered easy items. Based on participant responses to the SJT items, p-values (percent of participants answering an item correctly) were calculated as a second method to determine the level

of difficulty for each item. Items with p-values above .75 were categorized as easy, items with p-values between .51 and .74 were categorized as moderately difficult, and items with p-values of .50 and below were categorized as difficult items.

The items were then grouped by dimension and paired by difficulty level. One item from each pair was assigned to either Form A or Form B of the test. If a dimension contained an odd number of items, the "odd item out" was assigned to both forms of the test. Thus, each form of the SJT had an equal number of items for each dimension and the items were of equivalent difficulty. Composites were calculated for each dimension on each form of the test. Coefficients of equivalence were calculated for each dimension and for overall SJT scores.

Study 2 Results

Analyses of the SME data from the calibration step resulted in 53 items categorized as easy, 49 items categorized as moderate, and 28 items categorized as difficult. The results for each of the eight dimensions are presented in Table 3.

Table 3
Number of Items by Dimension and Difficulty Category Based on SME Ratings

Dimension	Easy	Moderate	Difficult	TOTAL
Organizing/Planning/Visioning	8	7	3	18
Consideration/Team Skills	10	6	5	21
Problem Solving/Innovation	8	8	3	19
Influencing Others	3	5	3	11
Communication	6	4	2	12
Drive/Results-Orientation	9	10	6	25
Tolerance for Stress	2	5	4	11
Integrity/Ethics	7	4	2	13
TOTAL	53	49	28	130

Next, p-values (i.e., percent correct) were calculated for each item after the participants completed SALSA©. After this step, 39 items were categorized as easy, 57 as moderate, and 34 as difficult. The results for each of the eight dimensions are presented in Table 4.

Table 4
Number of Items by Dimension and Difficulty Category Based on P-Values

Dimension	Easy	Moderate	Difficult	TOTAL
Organizing/Planning/Visioning	5	10	3	18
Consideration/Team Skills	7	7	7	21
Problem Solving/Innovation	4	11	4	19
Influencing Others	3	5	3	11
Communication	5	3	4	12
Drive/Results-Orientation	7	13	5	25
Tolerance for Stress	4	5	2	11
Integrity/Ethics	4	3	6	13
TOTAL	39	57	34	130

These two analyses were then compared to reach a final difficulty categorization for each item. The results from the first difficulty analysis and the second difficulty analysis were significantly correlated ($r = .63$, $p < .01$) and 65.4% of the items were categorized into the same difficulty level by both methods. For those items where the different methods resulted in different difficulty categorization, a rational decision process was used to categorize the item. Generally, p-values were used to make this decision, but if the difference between means was close to being classified as a different category, that was factored into the decision. For example, if the difference between the average rating for the most effective response and the average rating for the second most effective response was .67 (i.e., moderate), but 83.6% of the participants answered the item correctly (i.e., easy), the item was categorized as easy. The final difficulty

categorization yielded 45 easy items, 53 moderately difficult items, and 32 difficult items.

Table 5
Final Difficulty Categorization of Items by Dimension

Dimension	Easy	Moderate	Difficult	TOTAL
Organizing/Planning/Visioning	7	7	4	18
Consideration/Team Skills	9	6	6	21
Problem Solving/Innovation	5	11	3	19
Influencing Others	2	5	4	11
Communication	5	4	3	12
Drive/Results-Orientation	7	13	5	25
Tolerance for Stress	4	4	3	11
Integrity/Ethics	6	3	4	13
TOTAL	45	53	32	130

To assign the items to the two different forms, each item within a dimension for each category (i.e., easy, moderate, and difficult) was paired with another item of equivalent difficulty. One item from each pair was randomly assigned to either SALSA© - Form A or SALSA© - Form B. If there were an odd number of items, the final item was assigned to both forms to keep the forms equivalent in terms of both difficulty and number of items. This occurred for a total of fourteen items. After completing this step, each form contained 72 total items. The distribution of items by difficulty is described in Table 6. A test map indicating which item numbers are contained on each of the SJT forms may be found in Appendix D.

Table 6
Format of Alternate Test Forms

Dimension	Easy	Moderate	Difficult	TOTAL
Organizing/Planning/Visioning	4	4	2	10
Consideration/Team Skills	5	3	3	11
Problem Solving/Innovation	3	6	2	11
Influencing Others	1	3	2	6
Communication	3	2	2	7
Drive/Results-Orientation	4	7	3	14
Tolerance for Stress	2	2	2	6
Integrity/Ethics	3	2	2	7
TOTAL	25	29	18	72

Coefficient alpha was calculated as an estimate of internal consistency for SALSA© overall and for each form of SALSA©. Internal consistency for SALSA© (i.e., all 130 items) was $\alpha = .91$, SALSA© - Form A was $\alpha = .82$ and SALSA© - Form B was $\alpha = .87$. Coefficient alpha was computed for each dimension overall and for each dimension on Forms A and B. These coefficients are reported in Table 7.

Table 7
Coefficient Alpha by Overall and Dimension

Dimension	Overall	Form A	Form B
Overall (130 items)	.91	.82	.87
Organizing/Planning/Visioning	.49	.26	.45
Consideration/Team Skills	.64	.47	.47
Problem Solving/Innovation	.55	.45	.37
Influencing Others	.56	.28	.54
Communication	.44	.12	.42
Drive/Results-Orientation	.74	.68	.50
Tolerance for Stress	.45	.07	.46
Integrity/Ethics	.41	-.02	.33

Hypothesis 1 predicted that there would be a significant, positive correlation between the scores on the two forms of the leadership SJT (overall and for each dimension). Accordingly, the two forms were compared for equivalence. The performance on the two forms were significantly correlated ($r = .91$, $p < .01$); the Spearman Brown coefficient was .95. Items that were included on both forms of the test were removed and the coefficient of equivalence was recalculated. After removing the redundant items, the two forms were still significantly correlated ($r = .85$, $p < .01$); the Spearman Brown coefficient was .92. Correlation coefficients were calculated between dimension scores from Form A and Form B for each of the eight dimensions. Correlations between Form A and Form B were Organizing/Planning/Visioning ($r = .47$), Consideration/Team Skills ($r = .55$), Problem Solving/Innovation ($r = .52$), Influencing

Others (r = .51), Communication (r = .53), Drive/Results-Orientation (r = .80), Tolerance for Stress (r = .51), and Integrity/Ethics (r = .61); all correlations were significant at $p <$ .01. Thus, Hypothesis 1 was supported.

Means and standard deviations were computed for each dimension on both forms and for the total on each form. These findings are reported in Appendix E.

Additional Analysis

Although no hypotheses were offered concerning participants with English as a second language, gender, LEAD students, and Leadership Certificate Students, it was of interest to determine if SALSA© scores were moderated by any of these variables. A 2 (gender) x 2 (ESL: yes or no) x 3 (Program: Certificate, Industrial/Organizational Psychology (I/O), LEAD class only) factorial ANOVA was conducted on the total SALSA© score to determine if any of these factors moderated performance on SALSA©. Data for one participant was not used for this analysis, as it appeared to be an outlier. Significant main effects were found for gender (F (1,59) = 10.770, p = .002, η^2 = .180), ESL (F (1,59) = 41.309, p = .000, η^2 = .457), and Program (F (1,59) = 3.97, p = .025, η^2 = .140). Females (n = 33) scored an average of 87.52 (SD = 15.01) while males (n = 27) scored an average of 79.96 (SD = 15.89) on SALSA©. Students who reported that English was their primary language (n = 42) had an average score of 92.02 (SD = 10.58) and students who reported English as their second language (n = 18) scored an average of 65.67 (SD = 10.58). A Tukey B post hoc analysis indicated that I/O students (n = 20) scored significantly higher (M = 96.55, SD = 8.22) than Leadership Certificate students (M = 78.71, SD = 15.09) and LEAD class only students (M = 77.30, SD = 15.13). None of the interactions were significant.

Study 3 Overview

The final study addressed the relationship between assessment center performance scores and the SJT. Significant convergent validities were expected even though different methods (i.e., SJT and assessment center) were used to measure the dimensions; the same construct (i.e., leadership) is being measured by the methods. Current university students in the leadership program who participated in the assessment center during the 2008-2009 school year also completed the leadership SJT. Students receive assessment center dimension scores across different exercises such that a composite score for each dimension could be computed for each student. These assessment center dimension scores were correlated with the dimension scores from the SJT. The following hypothesis was tested:

H2: There will be positive correlations between the assessment center composite score and scores on the leadership SJT (overall and for each dimension).

Frequently when evaluating the relationship between two tests purporting to measure multiple dimensions of the same construct, one evaluates discriminant validity as well as convergent validity. That is, one would expect different measures of the same construct to correlate more highly with each other than with measures of other constructs with the same or different instruments. In the current situation, however, there is a priori evidence that the assessment center dimensions and the SJT dimensions are not independent. Previous studies of data from the assessment center (as well as data in the current study) indicated the nine dimensions are highly intercorrelated. The retranslation step of the SJT development indicated that, while each item represents the dimension to which it was assigned, close to half of the items were sorted into dimensions other than

the assigned dimensions by half of the SMEs. Thus, each dimension may represent a different aspect of leadership, but the data indicate that neither the assessment center dimensions nor the SJT dimensions are independent of each other. The lack of independence would diminish the magnitude of any discriminant validity coefficients and make them difficult to interpret.

Study 3 Method

Participants

In the past, the CLE assessment center has assessed both undergraduate and graduate students enrolled in the leadership program. However, because of lack of funds and resources, CLE recently has been assessing only graduate students in the entry assessment center. During the fall 2008 assessment center, a short information session was given on the SJT and student participants were asked to volunteer to complete the SJT. However, of the 26 students who volunteered to complete the test at this time, only eight actually completed SALSA© (30.8%). There was a time interval of approximately five months from the time the students completed the assessment center and when they completed the SJT. An additional 32 students completed the assessment center in spring 2009 and all of these students completed SALSA© approximately two months afterward. Their SALSA© scores and assessment center scores were used for Study 3.

Forty students participated in the study; 37 were graduate students and 3 were undergraduate seniors participating in an exit assessment center. Participants in this study were a subsample of the participants in Study 2. The sample was made up of 23 females and 17 males. Twenty participants were Caucasian, 15 were Asian, 2 were other, 1 was African America, 1 was Hispanic, and 1 responded as a non-resident alien. Twenty-two of the participants listed English as their primary language while the other 18 listed English as their second language. All participants in the sample have completed or are currently enrolled in LEAD200/500 (Leadership Theory) or LEAD400/600 (Leadership Practicum). Of the 40 participants, 16 are enrolled in WKU's Leadership Certificate

Program. Because the Leadership Certificate Program and LEAD classes are open to all majors and programs, the participants represented many different disciplines.

Procedure

Participants began by completing the leadership skills assessment center. Some of these students completed the "entry" assessment center at the start of the program (n = 34), while others completed the "exit" assessment center at the end of the program (n = 6). The only difference between the "entry" and "exit" assessment centers is that the participants are presented with different problems in the exercises. The format and dimensions remain the same. All participants completed an oral presentation, a leaderless group discussion, and two team simulations. Students received scores for each dimension across the different exercises and composite scores for each dimension were computed. Students then completed all 130 items on the leadership SJT through a campus electronic platform. Composite scores were computed for each SJT dimension. The composite scores for both the assessment center and leadership SJT were used to compute a correlation matrix so that convergent validities between the competencies on the two different methods of assessment could be evaluated.

Study 3 Results

Hypothesis 2 predicted that assessment center scores and scores on the leadership SJT would be positively correlated (overall and for each dimension). Six of the eight SJT dimensions (i.e., all except for Tolerance for Stress and Integrity/Ethics) match up with an assessment center dimension. The correlation matrix for all dimension and overall assessment center and SJT scores may be found in Appendix F. All but seven of the correlations were significant, as noted in the matrix. Convergent validities for the matched dimensions ranged from $r = .28$ to $r = .44$, and all were significant. All of the assessment center dimensions were highly intercorrelated except for Visioning and Planning; all correlation coefficients were significant. The SJT dimensions were correlated with each other in the $r = .40$ to .70 range. The composite assessment center score was significantly correlated with the individual assessment center dimensions, ranging from $r = .97$ to $r = .99$, except for Visioning and Planning ($r = .40$). The composite SJT score was significantly correlated with the individual SJT dimensions, ranging from $r = .55$ to $r = .81$. Finally, the composite assessment score was significantly correlated with the composite SJT scores ($r = .55$, $p < .01$). Generally, convergent validities for the matched dimensions were poor, but significant: Problem Solving/Innovation ($r = .40$, $p < .01$), Visioning and Planning ($r = .28$, $p < .05$), Influencing Others ($r = .44$, $p < .01$), Verbal/Non-Verbal Communication ($r = .29$, $p < .05$), Team Skills ($r = .33$, $p < .05$), and Results-Orientation ($r = .37$, $p < .01$). Thus, Hypothesis 2 was partially supported.

Additional Analysis

Means and standard deviations aggregated by gender, ESL, and program were calculated for the assessment center total scores and SJT total scores and are presented in Table 8 and Table 9, respectively.

Table 8
Mean Assessment Center Total Scores by Gender, ESL, and Program[a]

		ACPSI	VP	ACIO	VNV	TS	RO	Composite
Female	*M*	14.65	4.30	13.87	14.78	11.48	10.17	69.26
	SD	5.66	1.40	5.51	5.14	4.61	4.70	25.72
Male	*M*	16.53	4.00	15.94	16.47	13.12	11.82	77.88
	SD	4.40	1.41	4.74	3.89	3.18	3.94	19.96
English- 1st	*M*	18.50	4.59	17.86	18.50	14.45	13.27	87.18
Language	*SD*	3.81	1.30	3.44	2.84	3.05	3.15	15.55
English- 2nd	*M*	11.72	3.67	10.94	11.83	9.39	7.94	55.50
Language	*SD*	4.13	1.37	4.51	3.78	3.48	3.99	19.75
Non-Certificate	*M*	15.00	4.17	14.52	15.13	11.65	10.43	70.91
Students	*SD*	5.45	1.37	5.67	4.89	4.41	4.54	24.82
Certificate Students	*M*	16.06	4.18	15.06	16.00	12.88	11.47	75.65
	SD	4.91	1.47	4.72	4.46	3.66	4.30	22.19
TOTAL	*M*	15.45	4.18	14.75	15.50	12.17	10.88	72.93
	SD	5.19	1.39	5.23	4.67	4.10	4.42	23.56

[a] Note: ACPSI= Assessment Center Problem Solving/Innovation; VP= Visioning & Planning; ACIO= Assessment Center Influencing Others; VNV= Verbal/Non-Verbal Communication; TS= Team Skills; RO= Results-Orientation

Table 9
Mean Assessment Center Total Scores by Gender, ESL, and Program[a]

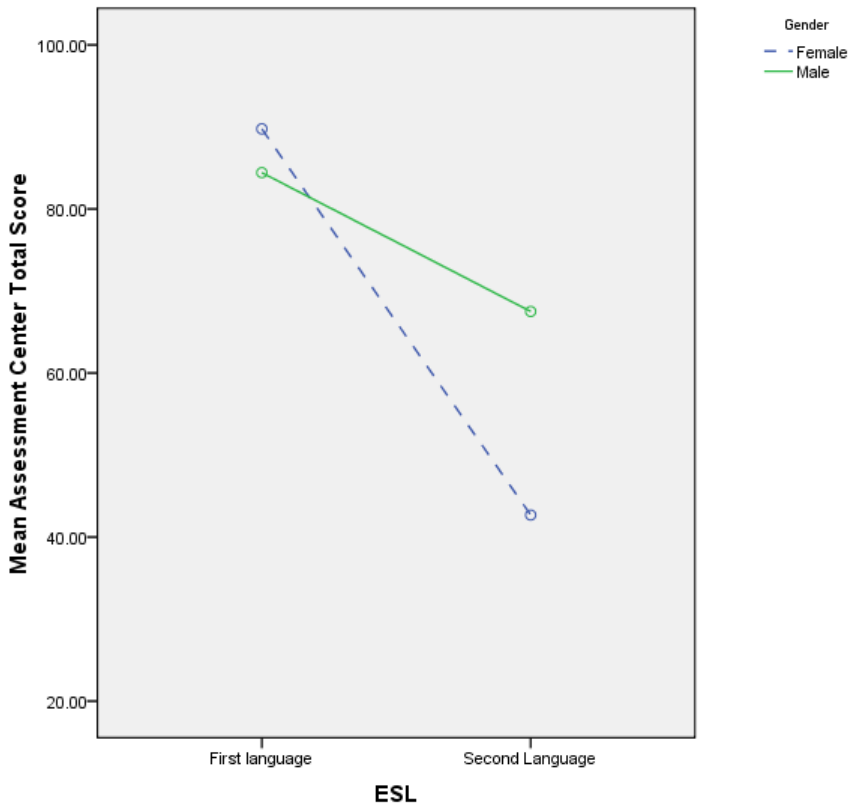
		OPV	CTS	PSI	IO	Comm	DRO	TS	IE	Composite
Female	*M*	11.74	12.78	12.61	6.22	8.04	16.52	6.91	7.48	82.30
	SD	2.34	2.30	2.73	1.86	1.49	3.80	2.15	2.02	14.44
Male	*M*	11.18	10.53	10.47	5.82	6.65	14.53	6.53	6.71	72.41
	SD	2.43	3.26	2.88	2.68	2.42	4.13	2.15	1.40	15.04
English- 1st	*M*	12.77	13.45	13.50	7.00	8.41	17.18	7.82	8.14	88.27
Language	*SD*	1.97	1.65	2.18	2.09	1.92	3.29	1.71	1.36	15.55
English- 2nd	*M*	9.94	9.83	9.50	4.89	6.28	13.83	5.44	5.94	65.67
Language	*SD*	1.83	2.96	2.20	1.81	1.53	4.13	1.89	1.55	19.75
Non-Certificate	*M*	11.04	11.61	11.09	6.00	7.30	14.52	6.65	7.09	75.30
Students	*SD*	2.71	2.31	2.56	2.37	2.23	4.62	2.27	2.04	24.82
Certificate	*M*	12.12	12.12	12.53	6.12	7.65	17.24	6.88	7.24	75.65
Students	*SD*	1.69	3.67	3.32	2.06	1.80	2.33	2.00	1.48	22.19
TOTAL	*M*	11.50	11.82	11.70	6.05	7.45	15.67	6.75	7.15	72.93
	SD	2.36	2.93	2.95	2.22	2.04	4.02	2.13	1.81	23.56

[a] Note: OPV= Organizing/Planning/Visioning; CTS= Consideration/Team Skills; PSI= Problem Solving/Innovation; IO= Influencing Others; Comm= Communication; DRO= Drive/Results-Orientation; TS= Tolerance for Stress; IE= Integrity/Ethics

Although no hypotheses were offered concerning participants with English as a second language, gender, LEAD students, and Leadership Certificate Students, it was of interest to determine if assessment center scores were moderated by any of these variables. A 2 (gender) x 2 (ESL: yes or no) x 2 (Program: Certificate, LEAD class only) factorial ANOVA was conducted on assessment center composite scores. A significant main effect was found for ESL (F (1,32) = 33.06, p = .000, η^2 = .508). Students who reported that English was their primary language (n = 22; M = 87.18, SD = 15.55) scored

significantly higher than did students who reported English as their second language (n = 18; M = 55.50, SD = 19.75). A significant gender by ESL interaction (F (1,32) = 7.34, p = .011, η^2 = .187) also was found. As seen in Figure 1, males (M = 85.00, SD = 17.82) and females (M = 88.69, SD = 14.33) with English as their primary language obtained equivalent assessment center scores and outscored ESL students. ESL males (M = 69.87, SD = 20.21) scored significantly higher than ESL females (M = 44.00, SD = 9.40), who scored the lowest of all groups. No other main effects or interactions were significant.

Figure 1. Interaction Between Gender and ESL for Assessment Center Total Score



Discussion

Retranslation and Calibration

Study 1 sought to ensure that the SJT was psychometrically sound through retranslation of the items and calibration of the response options. Following these steps, 130 of the original 300 critical incidents (43.3%) were successfully retranslated to their intended dimensions and calibrated. The 130 items provided a sufficiently large sample of items to work with to develop alternate forms of the SJT.

Initially in the retranslation task, a criterion of 66.7% agreement was used (i.e., agreement between four of the six SMEs); however, this resulted in only 106 items surviving retranslation. Thus, the agreement criterion was lowered to three of the six SMEs (50%), resulting in 213 items surviving retranslation. This suggests that the eight dimensions on the SJT are not independent. It further suggests that a given item may represent more than one leadership dimension. Psychometrically, it is desirable to have both independent dimensions and items that represent only one dimension. However, in reality, most leadership situations likely involve more than one dimension of leadership. As such, the fact that the SJT by definition involves hypothetical but realistic situations, it should not be that surprising that the test items represent more than one dimension. In fact, the overall internal consistency coefficient alpha of .90 indicates that the test is measuring a unitary underlying construct, presumably leadership.

Initial response options were generated by the same SMEs who generated the critical incidents. This is not typical of the procedure described in most of the literature on the development of situational judgment tests. For example, it is suggested by Motowidlo et al. (1990) that a different group of SMEs be used to generate the item

responses. However, Weekly and Ployhart (2006) used a process similar to that used in the current study. Based on anecdotal reactions from the students participating in the workshops, generating four response options for each critical incident was the most challenging task for them. In fact, the SME response options were substantially edited and additional response options were added during the editing process. It would be interesting to employ other methods to generate response options (e.g., using a different group of SMEs to write critical incidents and response options). The difficulty of generating four viable response options representing a range of leadership effectiveness was also illustrated by the fact that 83 items were lost because they either failed to have a correct answer (i.e., there was not a response rated as at least effective) or because there was more than one "best" answer.

For the calibration step, a different group of SMEs was asked to rank the response options on a 1-5 scale of leadership effectiveness. It has been suggested that the SMEs used to calibrate response options be representative in terms of demographics of the individuals that will complete the test (Shyamsunder, Lima, Burke, Tamanini, Horgen, & Teeter, 2009). This would suggest that individuals similar to students who will typically take SALSA in the future should act as SMEs during the calibration of responses (i.e., students in the Leadership Certificate Program).

In sum, our efforts to develop a leadership SJT appear to have been successful. A 130- item test with a sufficiently large number of items across eight dimensions of leadership was developed. The retranslation process ensured the items are representative of the dimension to which they were assigned. The calibration process ensured the response options reflect a range of leadership effectiveness and that there is a "correct"

answer for each item. The items also appear to represent an appropriate range of difficulty.

Alternate Forms Reliability

Alternate forms reliability was assessed in Study 2. Because the Center for Leadership Excellence uses the assessment center at the beginning of the certificate program and at the end of the certificate program, we were tasked with developing alternate forms of the SJT to be used in lieu of the pre and post assessment center. Two forms of the SJT were developed with items equated on difficulty and dimension representation. The resulting forms, SALSA© - Form A and SALSA© - Form B, contained 72 items each. A coefficient of equivalence showed a strong, positive correlation between the two forms (r = .91), which indicates that the two forms are equivalent measures and can used to similarly measure leadership ability. As such, they should work well as pre and posttests for assessment of students in the CLE Leadership Certificate program. Correlations between the dimensions on the two forms ranged from r = .47 to r = .80. Some of these correlations are lower than one would prefer given that the assessment center and SJT purport to be measuring the same underlying constructs. The small number of SJT items for some dimensions (i.e., 6 to14 items) may have played some role in the small magnitude of the correlations as longer tests likely would be more reliable. The low internal consistency coefficients for the dimensions, overall and particularly for the alternate forms, suggest that dimension scores should be interpreted with caution. Until more data can be collected and the dimension scores evaluated further, it might be prudent to report only SALSA© total scores if alternate forms are used.

Two types of difficulty analyses were utilized to determine item difficulty, the SME calibration ratings, and p-values from student participants. The two analyses resulted in 65.4% agreement on the difficulty categorization of the items. Some 34.6% of the items on the test are considered easy items, 40.7% are moderately difficult, and 24.6% of the items are difficult. For a test that is to be used as a pre and post evaluation for students enrolled in a multi-year leadership training program, this appears to be an appropriate distribution of item difficulty. Ideally, we would have a test that will accurately measure at both ends of the distribution of leadership knowledge. Students taking a pre-test presumably have a relatively low level of knowledge of leadership principles while those taking a post-test should have considerable knowledge of leadership. Because we used two different methods to determine difficulty and had reasonable agreement across the methods, we can be confident that the final difficulty of each item is an accurate indication, at least for the present sample.

Additional Findings

The results of ANOVA on total SALSA© scores indicated that I/O students outperformed other students; that females outperformed males; and that students with English as their primary language outperformed ESL students. One possible explanation for the program effect is that the I/O program students have completed graduate coursework that focuses on a broad array of organizational effectiveness factors in addition to leadership. This training may have provided an edge in understanding how to deal with the organizational situations contained on SALSA©. It is not clear why females outperformed males on SALSA©. English as primary language students scored more than 20 points better than ESL students on SALSA©. Language accounted for nearly

46% of the variance in SALSA© scores. This finding suggests that SALSA© may not measure leadership ability equally for all students, especially for those students who do not speak English as their primary language.

Convergent Validity

The third study compared scores in the assessment center and scores on the SJT. A moderate correlation was found between assessment center scores and SJT scores. Correlations between assessment center dimensions were found to be very strong. This finding suggests a lack of independence between the assessment center dimensions, the occurrence of halo error in the ratings of assessment center performance, or both. Correlations between the dimensions on the SJT were significant but not as strong, suggesting that the SJT dimensions may be somewhat more independent than the assessment center dimensions. Convergent validities between the matched assessment center dimensions and SJT dimensions were poor, although significant. This likely is a result of the high intercorrelations between the assessment center dimensions. Because the assessment center dimensions apparently are measuring a common underlying construct, it is difficult for the SJT dimensions to correlate differentially with them. The high assessment center intercorrelations also suggest a lack of construct validity for the dimensions in the assessment center, although the overall assessment center score may reflect a measure of leadership. The correlations between the assessment center dimensions and the SJT composite were moderate; the correlations between the SJT dimensions and the assessment center composite were slightly lower. These findings suggest that the two forms of assessment do not measure leadership in the same way.

The significant correlation between the SJT composite and the assessment center composite suggests that the SJT may be used as a substitute form of measurement for the assessment center. This is an important finding for the CLE and for the university because of the decrease in funding for administering the assessment center and the increased emphasis on distance learning (i.e., students are not physically on campus to participate in the assessment center). As of now, the Center for Leadership Excellence only allows students enrolled in LEAD classes to participate in the assessment center, therefore, assessment center data for non-LEAD students are not available. However, with the SJT it would be cost effective to collect data for comparison purposes. Theses results should be regarded with caution and additional data should be collected to continue the evaluation of SALSA©.

Additional Findings

The results of an ANOVA on total assessment center scores indicated that students with English as their primary language outperformed ESL students in general, and that male ESL students outperformed female ESL students, who performed least well of all groups. Language accounted for almost 51% of the variance in the assessment center performance. Both the ANOVA on assessment center scores and the ANOVA on SALSA© scores show a significant difference between ESL groups and language accounted for significant variance in performance on both evaluations. This finding suggests that future research should address other potential explanations for the ESL differences. For example, females from other cultures may be taught to defer to males in leadership situations. This cultural norm would likely impact their performance in assessment center exercises and, perhaps, on SALSA© as well. Female ESL students

may perform better in the assessment center in all female groups as compared to mixed gender groups.

Limitations

There were several limitations to the current study. A potential limitation of the retranslation and calibration steps may be the small number of SMEs used to retranslate the items and calibrate the responses. It is possible that if a larger sample of SMEs was used, a higher threshold of agreement could have been achieved. By using additional SMEs, the effects of outliers would be minimized. In other words, if two or more of the SMEs chose an answer that differed from the general consensus, the item did not meet the requirements. If more SMEs were used, a small number of the SMEs would not affect the results in this manner. Likewise, if more SMEs were used in the calibration step, the effects of extreme high and extreme low ratings would be minimized.

Another issue may be the quality of the critical incidents. Undergraduate students were given a brief overview of the development of the critical incidents during the workshops and were instructed on the definition of the dimensions to aid in writing the critical incidents. Still, it is possible that the students who generated the critical incidents lacked the experience necessary to tap into the critical nature of leadership knowledge, skills, and abilities. However, this concern should have been addressed by the fact that graduate students in industrial/organizational (I/O) psychology provided an initial edit of the items following the SME workshops and that an I/O psychologist subsequently provided substantial additional editing to the items.

Difficulty for alternate forms was based on one sample of SMEs (i.e., mean ratings) and one sample of students (i.e., p-values). The same scores were used to

calculate p-values and the convergent validity coefficients. As the test has not been cross-validated, it is impossible to determine if these findings will be consistent with other administrations of the test. It is recommended that a cross-validation study be conducted with a new sample of participants. Given the low internal consistency coefficients for the alternate forms dimensions, it is recommended that alternate forms be developed that address this concern. Items within a dimension could be matched on difficulty and item-total correlation before randomly assigning an item from each match to one of the forms. Collecting additional data on SALSA© may also increase the reliability of the dimensions as the development sample of 61 was relatively small.

Some of the dimensions on SALSA© - Form A and SALSA© - Form B contain a small number of items. This is likely to limit the reliability of the test. This aspect of SALSA© should continue to be monitored as more data is collected. If the alternative forms lack sufficient reliability, SALSA© may need to be administered as a single 130-item test.

Research on the reliability of the assessment center scores is lacking and we are unsure how accurate the exercises are in determining a student's leadership ability. Variability in the ratings suggests that the scores distinguish between effective and ineffective student leaders using the assessment center. However, empirical studies have not been conducted to validate the assessment center scores with a leadership criterion measure.

Directions for Future Research

We found that students for whom English is a Second Language (ESL) scored significantly lower on both the Assessment Center and on the SJT. This finding warrants

further research to determine whether this is a reflection of lower leadership knowledge or if the assessment center and SJT are biased against ESL students. It is possible that the tests measure Western ideals of leadership. These differences in test scores likely reflect that the SJT and assessment center have a strong verbal component; both tests are in English. It would be interesting to develop response options provided by non-English speaking students to examine how they might compare to the current response options. Moreover, future research should focus on underlying constructs, such as cognitive factors, that may lead to better scores on the SJT.

Different developmental procedures and different ways of matching up items for alternate forms could also be studied. As can be seen by the alphas for each dimension on the different forms, there is quite a bit of variability within a dimension, suggesting some groupings used for the current alternate forms are not equivalent across forms. The current alternate forms weighted p-values more heavily than difficulty based on SME ratings. Relying on SME ratings to determine difficulty may result in better alphas on the alternate forms. As mentioned above, the alternate forms of SALSA© should be cross-validated on a new sample of participants.

Future research might examine the correlation between grades in LEAD classes and scores on the SJT. Numeric grades, rather than letter grades, might provide a criterion measure with sufficient variability to determine if SALSA© is related to performance in leadership classes. SALSA© could also be correlated with overall GPA to determine how scores on the test are related to how well the students do in their other academic classes. It is important to determine if the test is actually measuring leadership ability or some other construct, such as general mental ability.

Conclusions

In sum, a leadership SJT, SALSA©, was developed measuring eight dimensions of leadership. The retranslation process ensured the items are representative of the dimension to which they were assigned. The calibration process ensured the response options reflect a range of leadership and that there is a "correct" answer for each item. The items also appear to represent an appropriate range of difficulty. Equivalent alternate forms were developed and are likely suitable for measuring change in leadership ability. These two forms are likely appropriate for use as pre and post assessment of students enrolled in the CLE Leadership Certificate Program. The high coefficient of equivalence suggests the amount of error in measurement is low so that we can be confident that any differences in pre and posttest scores are due to changes in ability or knowledge. Finally, SJT scores were significantly correlated with assessment center scores, contributing to the modest literature comparing these two different methods of assessment.

References

Arthur, W., Day, E.A., McNelly, T.L., & Edens, P.S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125-154.

Arthur, W., Day, E.A., & Woehr, D.J. (2008). Mend it, don't end it: An alternate view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology, 1,* 105-111.

Ashburn, C., & Love, R. (2006). The Development of a Leadership Assessment Center. Western Kentucky University, Bowling Green, KY.

Becker, T.E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment, 13,* 225-232.

Bergman, M.E., Drasgow, F., Donovan, M.A., Henning, J.B., & Juraska, S.E. (2006). Scoring situational judgment tests: Once you get the data, your trouble begins. *International Journal of Selection and Assessment, 14,* 223-235.

Borman, W.C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63,* 135-144.

Brannick, M.T. (2008). Back to basics of test construction and scoring. *Industrial and Organizational Psychology, 1,* 131-133.

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

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology, 1,* 98-104.

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

Jones, R.G., & Klimoski, R.J. (2008). Narrow standards for efficacy and the research playground: Why either-or conclusions do not help. *Industrial and Organizational Psychology, 1,* 137-139.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

Lance, C.E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology, 1,* 84-95.

Lievens, F. (2008). What does exercise-based assessment really mean? *Industrial and Organizational Psychology, 1,* 112- 115.

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

Lowry, P.E. (1995). The assessment center process: Assessing leadership in the public sector. *Public Personnel Management, 24,* 443-450.

McDaniel, M.A., Hartman, N.S., Whetzel, D.L., & Grubb, W.L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63-90.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A., & Braverman, E.P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730-739.

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

Moses, J. (2008). Assessment centers work, but for different reasons. *Industrial and Organizational Psychology, 1,* 134-136.

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

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

O'Connell, M.S., Hartman, N.S., McDaniel, M.A., Grubb, W.L., & Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment, 15,* 19-29.

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

Shoenfelt, E.L. (2009). Situational Assessment of Leadership – Student Assessment (SALSA©): The development and validation of a situational judgment test to assess leadership effectiveness. Unpublished manuscript, Western Kentucky University.

Smith, P.C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149-155.

Shyamsunder, A., Lima, L., Burke, E., Tamanini, K.B., Horgen, K., & Teeter, L. (2009, April). *Practical issues in developing construct-based situational judgment tests.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Waldman, D.A., & Korbar, T. (2004). Student assessment center performance in the prediction of early career success. *Academy of Management Learning and Education, 3,* 151-167.

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

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

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

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

Appendix A

SJT Dimension Definitions

ORGANIZING / PLANNING / VISIONING

The extent to which the individual systematically arranges his/her own work and resources, as well as that of others, for efficient task accomplishment. The extent to which an individual anticipates and prepares for the future. The extent to which the individual effectively creates an image of the future for the organization and develops the necessary means to achieve that image.

CONSIDERATION / TEAM SKILLS

The extent to which the individual's actions reflect a consideration for the feelings and needs of others as well as an awareness of the impact and implications of decisions relevant to others inside and outside the organization. The extent to which the individual engages and works in collaboration with other members of the group so that others are involved in the process and the outcome.

PROBLEM SOLVING / INNOVATION

The extent to which an individual gathers information; understands relevant technical and professional information; effectively analyzes data and information; generates viable options, ideas, and solutions; selects supportable courses of action for problems and situations; uses available resources in new ways; and generates and recognizes creative solutions.

INFLUENCING OTHERS

The extent to which the individual persuades others to do something or adopt a point of view in order to produce desired results (without creating hostility) and takes action in which the dominate influence is one's own convictions rather than the influence of others' opinions.

COMMUNICATION

The extent to which the individual effectively conveys both oral and written information. The extent to which the individual effectively responds to questions and challenges.

DRIVE / RESULTS-ORIENTATION

The extent to which the individual originates and maintains a high activity level, sets high performance standards and persists in achievement, and expresses the desire to advance to higher job levels. The extent to which the individual establishes clear direction, pushes self and others for high quality and results, monitors progress and results, and demonstrates a bias for action.

TOLERANCE FOR STRESS

The extent to which the individual maintains effectiveness in diverse situations under varying degrees of pressure, opposition, and disappointment.

INTEGRITY / ETHICS

The extent to which the individual demonstrates consistency between word and deed across situations and circumstances. The extent to which the individual does "the right thing" across situations and circumstances, especially in difficult and challenging situations.

Appendix B

Critical Incidents Student SME Worksheet

Student Leader Workshop:
Examples of Leadership Situations and Behavioral Responses

DIMENSION OF PERFORMANCE:_____
Please do NOT use specific names or entities. Your examples should be written in generic terms.

1. Think of a time when you were or another student leader was particularly **effective** at this dimension of performance. Below describe the situation, the behavior that was effective, and why the behavior was so effective:

Antecedent / Situation:
_____
(2 to 4 sentences)

_____

_____

_____

Behavior:
_____

_____

Consequence (Why behavior was effective):
_____

What are other responses that would have been less effective?

a. _____

b. _____

c. _____

2. Now think of a time when you were or another student leader was particularly **ineffective** at this dimension of performance. Below describe the situation, the behavior that was not effective, and why the behavior was so ineffective:

Antecedent / Situation:
_____
(2 to 4 sentences)

_____

_____

_____

Behavior:
_____

_____

Consequence (Why behavior was in effective):
_____

What are other responses that would have been more effective?

a. _____

b. _____

c. _____

3. Now think of a time when you were or another student leader was **moderately effective** at this dimension of performance. Below describe the situation, the behavior that was of only average effectiveness, and why the behavior was only moderately effective:

Antecedent / Situation:
_____
(2 to 4 sentences)

_____

_____

_____

Behavior:
_____

_____

Consequence (Why behavior was only moderately effective):
_____

What are other responses that would have been more or less effective?

a. _____

b. _____

c. _____

4. Now think of any other behaviors that are **particularly relevant** to this dimension of performance. Below list the actual behavior, the situation in which it occurred, and why the behavior was **either effective or ineffective**:

Antecedent / Situation:

_____

(2 to 4 sentences)

_____

_____

_____

Behavior:

_____

_____

Consequence (Was behavior effective?)

_____

What are other responses that would have been more or less effective?

a. _____

b. _____

c. _____

Thank you for participating in this important workshop!

Appendix C

WKU HSRB Approval Form

WKU

A LEADING AMERICAN UNIVERSITY WITH INTERNATIONAL REACH

HUMAN SUBJECTS REVIEW BOARD

In future correspondence, please refer to HS09-186, March 30, 2009

Kaci Grant
c/o Dr. Shoenfelt
Psychology
WKU

Kaci Grant:

Your research project, *The Validation of a Situational Judgment Test to Measure Leadership Behavior*, was reviewed by the HSRB and it has been determined that risks to subjects are: (1) minimized and reasonable; and that (2) research procedures are consistent with a sound research design and do not expose the subjects to unnecessary risk. Reviewers determined that: (1) benefits to subjects are considered along with the importance of the topic and that outcomes are reasonable; (2) selection of subjects is equitable; and (3) the purposes of the research and the research setting is amenable to subjects' welfare and producing desired outcomes; that indications of coercion or prejudice are absent, and that participation is clearly voluntary.

1. In addition, the IRB found that you need to orient participants as follows: (1) signed informed consent is required; (2) Provision is made for collecting, using and storing data in a manner that protects the safety and privacy of the subjects and the confidentiality of the data. (3) Appropriate safeguards are included to protect the rights and welfare of the subjects.

   This project is therefore approved at the Expedited Review Level until August 1, 2009.

2. Please note that the institution is not responsible for any actions regarding this protocol before approval. If you expand the project at a later date to use other instruments please re-apply. Copies of your request for human subjects review, your application, and this approval, are maintained in the Office of Sponsored Programs at the above address. Please report any changes to this approved protocol to this office. A Continuing Review protocol will be sent to you in the future to determine the status of the project. Also, please use the stamped form that accompanies this letter.

Sincerely,

Paul J. Mooney, M.S.T.M.
Compliance Manager
Office of Sponsored Programs
Western Kentucky University

HUMAN SUBJECTS REVIEW BOARD WKU APPROVED

cc: HS file number Grant HS09-186

HSRB APPLICATION # 09-186
APPROVED 3/30/09 to 8/1/09
EXEMPT EXPEDITED FULL BOARD
DATE APPROVED 3/30/09

Appendix D

Test Map for Alternate Forms

Item	Rating Difference	Difficulty 1	P-value	Difficulty 2	Final Difficulty	Form
Org01	1.17	Easy	.902	Easy	Easy	A
Org02	1	Moderate	.623	Moderate	Moderate	A
Org03	0.67	Moderate	.607	Moderate	Moderate	B
Org04	1.16	Easy	.803	Easy	Easy	B
Org05	0.83	Moderate	.738	Moderate	Moderate	A
Org06	0.83	Moderate	.705	Moderate	Moderate	B
Org07	1.33	Easy	.721	Moderate	Easy	A
Org08	1	Moderate	.623	Moderate	Moderate	A
Org09	2.33	Easy	.869	Easy	Easy	B
Org10	1.16	Easy	.705	Moderate	Moderate	B
Org11	0.5	Difficult	.459	Difficult	Difficult	A
Org12	0.33	Difficult	.525	Moderate	Difficult	B
Org13	0.5	Difficult	.443	Difficult	Difficult	A
Org14	1.66	Easy	.902	Easy	Easy	A
Org15	0.66	Moderate	.328	Difficult	Difficult	B
Org16	0.83	Moderate	.541	Moderate	Moderate	A,B
Org17	1.33	Easy	.820	Easy	Easy	B
Org18	1.34	Easy	.738	Moderate	Easy	A,B
Con01	0.80	Moderate	.574	Moderate	Moderate	A
Con02	0.67	Moderate	.836	Easy	Easy	A
Con03	0.84	Moderate	.705	Moderate	Moderate	B

Con04	1.17	Easy	.754	Easy	Easy	B
Con05	0.5	Difficult	.246	Difficult	Difficult	A
Con06	0.5	Difficult	.492	Difficult	Difficult	B
Con07	1.17	Easy	.902	Easy	Easy	A
Con08	1.17	Easy	.787	Easy	Easy	B
Con09	1.83	Easy	.902	Easy	Easy	A
Con10	0.33	Difficult	.311	Difficult	Difficult	A
Con11	1.83	Easy	.836	Easy	Easy	B
Con12	1	Moderate	.197	Difficult	Difficult	B
Con13	1.17	Easy	.656	Moderate	Moderate	A
Con14	1.34	Easy	.738	Moderate	Easy	A
Con15	0.66	Moderate	.590	Moderate	Moderate	B
Con16	0.5	Difficult	.459	Difficult	Difficult	B
Con17	1	Moderate	.410	Difficult	Moderate	A
Con18	1.17	Easy	.705	Moderate	Moderate	B
Con19	0.33	Difficult	.262	Difficult	Difficult	A
Con20	1.83	Easy	.705	Moderate	Easy	B
Con21	1.66	Easy	.852	Easy	Easy	A,B
Prob01	1.5	Easy	.656	Moderate	Easy	A
Prob02	1.33	Easy	.754	Easy	Easy	B
Prob03	1.17	Easy	.951	Easy	Easy	A
Prob04	0.5	Difficult	.115	Difficult	Difficult	A
Prob05	0.66	Moderate	.656	Moderate	Moderate	A

Prob06	1.17	Easy	.623	Moderate	Moderate	B
Prob07	1.16	Easy	.721	Moderate	Moderate	A
Prob08	0.84	Moderate	.410	Difficult	Moderate	B
Prob09	1.5	Easy	.951	Easy	Easy	B
Prob10	0.5	Difficult	.590	Moderate	Moderate	A
Prob11	0.66	Moderate	.656	Moderate	Moderate	B
Prob12	0.5	Difficult	.475	Difficult	Difficult	B
Prob13	0.84	Moderate	.295	Difficult	Difficult	A,B
Prob14	1.16	Easy	.639	Moderate	Moderate	A
Prob15	1.17	Easy	.918	Easy	Easy	A,B
Prob16	0.67	Moderate	.508	Moderate	Moderate	B
Prob17	1	Moderate	.721	Moderate	Moderate	A
Prob18	0.84	Moderate	.902	Easy	Moderate	B
Prob19	0.67	Moderate	.623	Moderate	Moderate	A,B
Influ01	1	Moderate	.508	Moderate	Moderate	A
Influ02	0.67	Moderate	.459	Difficult	Difficult	A
Influ03	0.83	Moderate	.721	Moderate	Moderate	B
Influ04	0.5	Difficult	.869	Easy	Moderate	A
Influ05	1.34	Easy	.754	Easy	Easy	A
Influ06	1.16	Easy	.639	Moderate	Moderate	B
Influ07	1	Moderate	.672	Moderate	Moderate	A,B
Influ08	0.33	Difficult	.525	Moderate	Difficult	B
Influ09	0.67	Moderate	.344	Difficult	Difficult	A

Influ10	0.17	Difficult	.246	Difficult	Difficult	B
Influ11	1.5	Easy	.803	Easy	Easy	B
Comm01	0.83	Moderate	.672	Moderate	Moderate	A
Comm02	1.84	Easy	.820	Easy	Easy	A
Comm03	1.33	Easy	.475	Difficult	Moderate	B
Comm04	0.83	Moderate	.525	Moderate	Moderate	A
Comm05	0.67	Moderate	.721	Moderate	Moderate	B
Comm06	0.84	Moderate	.377	Difficult	Difficult	A
Comm07	2	Easy	.754	Easy	Easy	B
Comm08	0.5	Difficult	.393	Difficult	Difficult	B
Comm09	1.83	Easy	.934	Easy	Easy	A
Comm10	0.5	Difficult	.393	Difficult	Difficult	A,B
Comm11	1.17	Easy	.836	Easy	Easy	B
Comm12	1.17	Easy	.754	Easy	Easy	A,B
Res01	0.5	Difficult	.721	Moderate	Moderate	A
Res02	1.34	Easy	.639	Moderate	Moderate	B
Res03	1.13	Easy	.721	Moderate	Moderate	A
Res04	2.5	Easy	.918	Easy	Easy	A
Res05	1	Moderate	.361	Difficult	Difficult	A
Res06	0.5	Difficult	.492	Difficult	Difficult	B
Res07	1.5	Easy	.836	Easy	Easy	B
Res08	0.5	Difficult	.443	Difficult	Difficult	A
Res09	0.84	Moderate	.557	Moderate	Moderate	B

Res10	1.5	Easy	.754	Easy	Easy	A
Res11	0.84	Moderate	.721	Moderate	Moderate	A
Res12	0.5	Difficult	.426	Difficult	Difficult	B
Res13	0.83	Moderate	.852	Easy	Easy	B
Res14	1	Moderate	.738	Moderate	Moderate	B
Res15	2.17	Easy	.918	Easy	Easy	A
Res16	0.84	Moderate	.705	Moderate	Moderate	A
Res17	1.17	Easy	.820	Easy	Easy	B
Res18	1	Moderate	.738	Moderate	Moderate	B
Res19	0.84	Moderate	.738	Moderate	Moderate	A
Res20	1.33	Easy	.836	Easy	Easy	A,B
Res21	1.16	Easy	.721	Moderate	Moderate	B
Res22	1	Moderate	.672	Moderate	Moderate	A
Res23	0.33	Difficult	.475	Difficult	Difficult	A,B
Res24	1	Moderate	.574	Moderate	Moderate	B
Res25	0.34	Difficult	.672	Moderate	Moderate	A,B
Tol01	0.33	Difficult	.475	Difficult	Difficult	A
Tol02	0.5	Difficult	.541	Moderate	Difficult	B
Tol03	1	Moderate	.787	Easy	Easy	A
Tol04	0.33	Difficult	.525	Moderate	Moderate	A
Tol05	0.67	Moderate	.803	Easy	Easy	B
Tol06	1.34	Easy	.836	Easy	Easy	A
Tol07	0.66	Moderate	.541	Moderate	Moderate	B

Tol08	0.67	Moderate	.721	Moderate	Moderate	A
Tol09	2.17	Easy	.754	Easy	Easy	B
Tol10	0.5	Difficult	.738	Moderate	Moderate	B
Tol11	0.67	Moderate	.475	Difficult	Difficult	A,B
Int01	0.67	Moderate	.656	Moderate	Moderate	A
Int02	0.84	Moderate	.492	Difficult	Moderate	B
Int03	0.83	Moderate	.393	Difficult	Difficult	A
Int04	0.34	Difficult	.475	Difficult	Difficult	B
Int05	1.67	Easy	.885	Easy	Easy	A
Int06	1.83	Easy	.639	Moderate	Easy	B
Int07	1.34	Easy	.738	Moderate	Easy	A
Int08	2	Easy	.787	Easy	Easy	B
Int09	2.5	Easy	.934	Easy	Easy	A
Int10	1.34	Easy	.492	Difficult	Moderate	A,B
Int11	0.5	Difficult	.443	Difficult	Difficult	A
Int12	1.33	Easy	.787	Easy	Easy	B
Int13	0.67	Moderate	.279	Difficult	Difficult	B

Appendix E

Group Means for Alternate Forms

		FormA	FormB	OrgA	OrgB	ConA	ConB	ProbA	ProbB	InfluA	InfluB	CommA	CommB	ResA	ResB	TolA	TolB	IntA	IntB
Female	*M*	48.33	48.61	7.00	6.94	7.06	7.64	7.27	7.48	3.67	3.73	4.79	4.73	10.15	9.91	3.76	4.03	4.64	4.15
	SD	8.26	8.89	1.60	1.84	1.41	1.75	1.89	1.70	1.24	1.49	1.17	1.15	2.56	2.16	1.17	1.55	1.32	1.40
Male	*M*	44.07	42.89	6.36	6.36	6.25	6.43	6.43	6.68	3.54	3.46	4.11	3.86	9.07	8.75	3.89	3.64	4.43	3.71
	SD	8.54	10.48	1.59	1.83	2.22	2.01	1.73	1.77	1.35	1.56	1.23	1.76	2.78	2.37	1.17	1.34	.92	1.54
English- 1st Language	*M*	50.05	50.67	6.98	7.44	7.19	7.91	7.67	7.51	3.95	4.09	4.67	4.74	10.53	10.05	4.05	4.37	5.00	4.56
	SD	6.76	6.83	1.61	1.52	1.52	1.31	1.38	1.58	1.23	1.31	1.23	1.40	2.31	1.89	1.11	1.05	.93	1.12
English- 2nd Language	*M*	37.61	34.78	6.06	4.83	5.50	5.11	5.00	6.17	2.78	2.44	4.00	3.33	7.56	7.78	3.28	2.61	3.44	2.50
	SD	5.67	6.95	1.47	1.10	2.09	1.84	1.46	1.86	1.00	1.38	1.14	1.33	2.41	2.49	1.13	1.58	.86	1.15
Non-LEAD Students	*M*	48.84	48.72	6.88	7.24	7.48	7.56	7.16	7.04	3.72	3.92	4.80	4.44	10.08	9.88	3.84	4.20	4.88	4.44
	SD	9.04	10.27	1.79	1.92	1.87	2.00	1.77	1.54	1.40	1.15	1.32	1.45	2.96	2.56	1.14	1.38	.97	1.64
LEAD Students	*M*	44.67	44.08	6.58	6.28	6.14	6.75	6.69	7.17	3.53	3.39	4.25	4.25	9.36	9.03	3.81	3.61	4.31	3.61
	SD	7.94	9.46	1.50	1.70	1.66	1.87	1.91	1.92	1.21	1.71	1.13	1.57	2.50	2.09	1.19	1.48	1.22	1.25
Non-Certificate Students	*M*	47.27	46.56	6.73	6.88	7.10	7.15	6.80	7.17	3.83	3.59	4.54	4.41	9.68	9.32	3.90	4.02	4.68	4.02
	SD	8.76	10.09	1.64	1.94	1.69	1.87	1.89	1.52	1.32	1.47	1.31	1.55	2.85	2.53	1.00	1.53	1.25	1.51
Certificate Students	*M*	44.55	44.80	6.65	6.25	5.85	6.95	7.05	7.00	3.15	3.65	4.35	4.15	9.60	9.50	3.65	3.50	4.25	3.80
	SD	8.13	9.92	1.60	1.59	1.95	2.16	1.82	2.22	1.10	1.66	1.10	1.46	2.42	1.82	1.46	1.28	.85	1.40
TOTAL	*M*	46.38	45.98	6.70	6.67	6.69	7.08	6.89	7.11	3.61	3.61	4.48	4.33	9.66	9.38	3.82	3.85	4.54	3.95
	SD	8.59	9.99	1.62	1.84	1.86	1.95	1.85	1.76	1.28	1.52	1.23	1.51	2.70	2.31	1.16	1.50	1.15	1.47

Appendix F

Correlations Between Assessment Center and SJT

	ACPSI	ACVP	ACIO	ACVNV	ACTS	ACRO	PSI	OPV	IO	Comm	CTS	DRO	TolS	IE	AC	SJT
ACPSI	1.00															
ACVP	.375**	1.00														
ACIO	.959**	.414**	1.00													
ACVNV	.942**	.380**	.937**	1.00												
ACTS	.948**	.228	.939**	.914**	1.00											
ACRO	.939**	.254	.950**	.909**	.961**	1.00										
PSI	**.400****	.306*	.362*	.422**	.313*	.314*	1.00									
OPV	.493**	**.284***	.460**	.490**	.422**	.399**	.492**	1.00								
IO	.501**	.428**	**.443****	.502**	.380**	.378**	.425**	.538**	1.00							
Comm	.257	.432**	.275*	**.293***	.150	.160	.607**	.538**	.420**	1.00						
CTS	.376**	.277*	.411**	.388**	**.333***	.349*	.695**	.468**	.391**	.554**	1.00					
DRO	.383**	.427**	.406**	.430**	.346*	**.372****	.463**	.439**	.509**	.247	.463**	1.00				
TolS	.578**	.601**	.530**	.525**	.389**	.443**	.602**	.534**	.539**	.434**	.497**	.622**	1.00			
IE	.329*	.264*	.341**	.386**	.301*	.266*	.489**	.499**	.619**	.497**	.455**	.392**	.496**	1.00		
AC	.983**	.396**	.985**	.966**	.966**	.968**	.384**	.473**	.471**	.258	.391**	.415**	.522**	.343*	1.00	
SJT	.543**	.502**	.540**	.574**	.444**	.455**	.806**	.736**	.721**	.683**	.772**	.747**	.789**	.704**	.546**	1.00

**. Correlation is significant at the 0.01 level (1-tailed). *. Correlation is significant at the 0.05 level (1-tailed).

NOTE: ACPSI= Assessment Center Problem Solving & Innovation; ACVP= Assessment Center Visioning & Planning; ACIO= Assessment Center Influencing Others; ACVNV= Assessment Center Verbal/Non-Verbal Communication; ACTS= Assessment Center Team Skills; ACRO= Assessment Center Results-Orientation; PSI= SJT Problem Solving & Innovation; OPV= SJT Organizing/Planning/Visioning; IO= SJT Influencing Others; Comm= SJT Communication; CTS= SJT Consideration/Team Skills; DRO= SJT Drive/Results-Orientation; TolS= SJT Tolerance for Stress; IE= SJT Integrity/Ethics; AC= Composite Assessment Center score; SJT= Composite Assessment Center score.

NOTE: Convergent validities between corresponding assessment center dimensions and SJT dimensions are in bold.