5-2014

The Adaptation of a Situational Judgement Test to Measure Leadership Knowledge in the Workplace

Ebo K. A Osam
Western Kentucky University, ebo.osam840@topper.wku.edu

THE ADAPTATION OF A SITUATIONAL JUDGMENT TEST TO MEASURE LEADERSHIP KNOWLEDGE IN THE WORKPLACE

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

By
Ebo K.A Osam

May 2014

THE ADAPTATION OF A SITUATIONAL JUDGMENT TEST TO MEASURE LEADERSHIP BEHAVIOR IN THE WORKPLACE

Date Recommended April 25, 2014

Elizabeth L. Shoenfelt

Elizabeth L. Shoenfelt, Director of Thesis

Reagan Brown

Cindy J. Ehresman

Cindy Ehresman

Dean, Graduate Studies and Research 5-6-14 Date

ACKNOWLEDGMENTS

I would like to express my sincere gratitude and appreciation to Dr. Elizabeth Shoenfelt for her assistance and support as I completed my thesis. She provided me with valuable insights pertaining to this project that extends to other aspects of my life. I would also like to thank Dr. Reagan Brown and Dr. Cindy Ehresman for serving on my committee, and for giving me their support and encouragement throughout the research process. Finally, I would like to thank Dr. Cecile Garmon for her support, and all the accommodations she provided to ensure that this thesis was completed on time.

TABLE OF CONTENTS

THE ADAPTATION OF A SITUATIONAL JUDGMENT TEST TO MEASURE LEADERSHIP KNOWLEDGE IN THE WORKPLACE

Ebo K.A Osam May 2014 42 Pages

Directed by: Dr. Elizabeth Shoenfelt, Dr. Reagan Brown, and Dr. Cindy Ehresman

Department of Psychology Western Kentucky University

In recent times, situational judgment tests (SJTs) have emerged as an instrument of choice in organizations. This emergence is partly due to the high costs associated with developing and conducting high fidelity simulations such as assessment centers, coupled with the recent economic downturn affecting many organizations. The current study sought to validate an SJT as a low cost, alternate form of assessing leadership within an organizational context. A content validation study was carried out by retranslating items into eight dimensions and calibrating item responses. This study resulted in a content valid measure of leadership knowledge. Future studies should focus on further evaluating the psychometric properties of this new leadership assessment. Alternate forms reliability, convergent validity, and divergent validity studies, in particular, should be conducted to evaluate the new test.

Introduction

Management in organizations today is preparing for the impending wave of baby boomer retirements by taking steps to identify and nurture leadership talent (Gowing, Morris, Adler & Gold, 2008). This identification and nurturing of leadership talent, however, is not simply a reaction to the impending baby boomer retirements, but has been a key part of organizations for many years. Gowing et al. indicated the use of assessment centers has been at the forefront of the identification of leadership talent, and has been seen by some as the most valid method for assessing leadership potential over the past two decades. However in more recent times, situational judgment tests (SJTs) have emerged as the tool of choice for use in organizations (Patterson et. al 2012). This emergence is partly due to the high costs associated with developing and conducting assessment centers, coupled with the recent economic downturn affecting many organizations. In addition to this, Lance (2008) indicated that the use of assessment centers has come under scrutiny due to problems with rater bias, scoring methods, and realism among tasks. It is clear, therefore, that an alternate mode of measuring dimensions of leadership assessed in assessment centers would be of interest to organizations today, especially given the costs associated with conducting assessment centers.

At Western Kentucky University, a situational judgment test called Situational Assessment of Leadership: Student Assessment (SALSA©) is used by the Leadership Studies Department and the Doctoral Program in Educational Leadership to assess eight specific leadership dimensions of students studying leadership. The introduction of this test has been beneficial because it enables appraising and providing feedback to students

concerning various dimensions of leadership. Feedback further provides the opportunity for students to improve aspects of leadership on which they fall short. Given the successful implementation of SALSA©, this study seeks to adapt SALSA© into a similar situational judgment test that could be used in an organizational setting. Such a test will enable employers to provide feedback on eight specific leadership dimensions for their employees, design training programs to improve aspects of leadership, and potentially even use the tool as a part of a selection system.

In the next section, the paper will review the available literature on SJTs. Areas that will be covered specifically will include history, validity, reliability as well as a number of other strengths, and potential limitations of situational judgment tests. The paper will then review the SALSA© that is currently used at Western Kentucky University along with a brief overview of its development. The procedure to be followed to develop the new leadership SJT for organizational use will also be discussed. The current study will seek to evaluate the content validity of the test through the retranslation and calibration of SALSA© test items.

Overview of Situational Judgment Tests

A situational judgment test (SJT) is a type of assessment that consists of hypothetical scenario-based questions requiring test takers to use careful judgment to pick answers from a list of plausible courses of action (O'Connell, Hartman, McDaniel, Grubb, & Lawrence, 2007). These questions are related to situations that occur within the work place, and seek to measure the typical or maximal performance of given constructs (McDaniel &Nguyen, 2001). Although SJTs tend to be associated with hiring and promotion at work, they can be developed to assess many other types of constructs such

as leadership (Weekley & Ployhart, 2006). An illustration of an SJT item from SALSA© is presented below.

1) During a team exercise, you notice one team member is falling behind and cannot keep up with the group. What is the most effective leadership action you would take in this situation?

    a. Modify the exercise so the individual is able to participate.
    b. Continue with the exercise and leave the team member behind.
    c. Speak to the individual and tell him/her they need to keep up.
    d. Point out the individual to the rest of the team as an example of failure.

SJTs have a long history within the field of psychology, stretching as far back as the early 20th century (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). McDaniel et al., indicated that a subtest of the George Washington Social Intelligence Test called the Judgment in Social Situations was one of the first widely used SJTs that required a keen sense of acumen to be successful on the test. SJTs have since been used in military settings, resulting in lower attrition rates among new officers, as well as in organizations in the 1950s and 1960s to predict managerial success (Whetzel & McDaniel, 2009). Despite the fact that SJTs have been around for quite a while, the past decade has seen a marked increase in the use of SJTs in several occupational groups to predict performance for different positions as well as to identify training needs (Grant, 2009; Patterson et al., 2012). Patterson et al. further indicated that SJTs are not one special kind of test, but rather a type of measurement method that can be presented in different forms such as written and video formats.

Strengths of Situational Judgment Tests

According to Patterson et al. (2012), SJTs have emerged as an established tool for selection, and offer significant advantages over more traditional methods such as interviews and personality tests. The strengths of SJTs are demonstrated through research findings on their reliability, validity, and adverse impact. These strengths are discussed below.

Reliability

Reliability as a strength of SJTs has been questioned by some researchers. Patterson et al. (2012) indicated that one of the reasons for this assertion is that SJTs typically measure multiple constructs, and therefore present a challenge of accurately estimating their reliability. It is for this reason that Patterson et al. stated that SJTs are often referred to as construct heterogeneous, since one item on a test could encompass multiple performance dimensions. However, in general, most research has typically reported good levels of reliability using internal consistency. For example, a meta-analysis conducted by McDaniel et al. (2001) focusing on SJT reliability reported internal consistency coefficients ranging from .43 to .94. McDaniel et al., however, stated that the length of SJTs moderates its reliability, and therefore SJTs that contain more items tend to be more reliable. The type of response instructions on SJTs also has a moderating effect on reliability. Ployhart and Ehrhart (2003), for instance, found that rating the effectiveness of each response on an SJT leads to an internal consistency of .73, as opposed to an internal consistency of .24 when the instructions on the SJT were to choose the most effective response.

As mentioned earlier, using reliability as a strength of SJTs has been called into question due to the multifaceted nature of SJTs. Using internal consistency is therefore not the ideal way to measure reliability, as it is more suited for unidimensional tests which measure one construct (Patterson et al., 2012). The most accurate means to measure reliability therefore, is to use test-retest or parallel forms approach (O'Connell et al., 2007). Ployhart and Ehrhart (2003) found test retest reliability as high as .92, moderated by the type of response instructions used. Overall, the research has shown moderate to good levels of reliability for the use of SJTs regardless of whether an internal consistency, test-retest, or parallel forms approach is used.

Validity

The term validity has different interpretations; however, the Society for Industrial and Organizational Psychology (SIOP) has offered some professional guidance with respect to the concept of validity. According to the Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology, Inc., 2003), validity is the extent to which gathered data support the interpretation of a test score. In other words, a valid test shows the extent to which the test is useful based on the inferences that a test developer makes about the test. Whetzel and McDaniel (2009) indicated that criterion-related validity and construct-related validity are the two main sources of evidence to demonstrate the validity of SJTs. In addition to criterion validity and construct validity, face validity and incremental validity are strengths of SJTs that are discussed below.

Construct Related Validity. Bagozzi, Yi, and Phillips (1991) indicated that construct validity is the extent to which an instrument actually measures the construct

that it was designed to measure. Bagozzi et al. also indicated that the most common way to establish construct validity is by making comparisons between a measure and other well established measures of the construct. Research on construct validity provides evidence that suggests that SJTs are related to multiple constructs. Patterson et al. (2012), for example, indicated that SJTs correlate with cognitive ability (r = .46), personality (r = .32 for conscientiousness), and job knowledge (r = .23).

The relationship between SJTs and cognitive ability has been demonstrated by several studies. For example, McDaniel, Hartman, Whetzel, and Grubb (2007) found a moderate correlation between SJTs and cognitive ability of r = .32. They, however, indicated that the magnitude of the correlations between SJTs and cognitive ability is moderated by response instructions, with knowledge instructions (r = .35) correlating more highly with cognitive ability than behavioral instructions (r = .19). Another factor that could moderate the relationship between cognitive ability and SJTs is the use of job analysis. Patterson et al. (2012) stated that SJTs based on a job analysis correlate more highly with cognitive ability (r = .56) than do those constructed without a job analysis (r = .38).

As with cognitive ability, there has been a lot of research examining personality and SJTs. Patterson et al. (2012), in their review of literature, indicated that SJTs correlate moderately well with conscientiousness (r = .26 to .32), agreeableness (r = .22 to .26) and emotional stability (r = .22). As with cognitive ability, there are factors that moderate the relationship between personality and SJTs. Whetzel and McDaniel (2009) identified test instructions as one such moderator. They further mentioned that SJTs that

contain behavioral tendency instructions typically correlate higher with personality than do SJTs with knowledge instructions.

Criterion Related Validity. Motowidlo, Dunnette and Carter (1990) indicated that numerous studies have been conducted to establish criterion validity of SJTs due to the fact that they are typically used in job settings. For example, in a meta-analysis by McDaniel et al. (2001), SJTs were shown to have a mean uncorrected validity of .26 with performance. Other studies that have examined SJTs and performance ratings have indicated moderate correlations ranging from r = .30 to .35 (Chan & Schmidt, 2002; Motowidlo, et al., 1990; Motowildlo & Tippins, 1993; Weekly & Jones, 1997).

Some criterion validity research focused outside of the workplace has indicated significant correlations between SJTs and performance. For example, Patterson et al. (2012) indicated that other research using objective criterion measures such as sales performance has shown significant correlations with SJTs (r = .63). Patterson et al. also noted a correlation of r = .52 between applicant scores on an SJT and performance ratings in multiple mini interviews in a study on selection into graduate programs.

In summary, the literature indicates that getting a clear picture of criterion related validity of SJTs is more challenging than straightforward; however, the information provided from the correlations between SJTs and performance cited above suggests that SJTs have useful validity levels when it comes to predicting job performance.

Face Validity. Perhaps one of the more easily identifiable strengths associated with SJTs is that they typically show a high degree of face validity. Face validity, according to Rynes and Connerley (1993), is one of the reasons SJTs are popular. Rynes and Connerley further explained that because most SJTs appear to be highly work-

related, they usually elicit favorable reactions from respondents. Callinan and Robertson (2000) also provided some further explanation on the favorable respondent reaction towards SJTs. Callinan and Robertson mentioned that respondents tend to perceive SJTs as an opportunity to demonstrate their ability to perform well on job related tasks, which then leads to more motivation and increased engagement with the test.

The format of SJTs has an effect on respondent reaction. For example, Patterson et al., (2012) identified that video based SJTs are more likely to receive favorable ratings from respondents as compared to written SJTs. Kanning and Kuhne (2006) also reported more positive ratings from applicants who were administered interactive SJTs that used a video format as compared to written SJTs

Incremental Validity. Incremental validity, another potential strength of SJTs, refers to an increase in predictive validity when a predictor is used with other established selection measures (Patterson et al., 2012). Schmidt and Hunter (1998) indicated that using additional predictors is only of value from a utility standpoint when the additional predictors add to the variance explained in the criterion beyond other less expensive predictors.

There have been a number of studies that have sought to identify whether SJTs provide significant incremental validity, if any at all. Lievens, Buyse, and Sackett (2005) investigated the incremental validity of an SJT measuring non-academic attributes. Lievens et al. found that for cognitively oriented predictors, the SJT showed significant incremental validity. O'Connell et al. (2007) also indicated that SJTs have incremental validity of .03 and .04 over cognitive and personality measures, respectively, when it came to predicting task performance. For contextual performance, O'Connell et al. found

that SJTs had incremental validity of .04 over cognitive ability. However, the authors found that SJTs did not add incrementally to personality measures for contextual performance. These findings are similar to those found by McDaniel et al. (2007). Results from the McDaniel et al. meta-analysis indicated that SJTs provide incremental validity over cognitive ability and personality estimated at between .03 to .05, and .06 to .07, respectively.

Results from these studies indicate that SJTs can provide a measure of utility when used as an additional predictor. McDaniel et al. (2007), however, warned that one can create situations whereby SJTs appear to have substantial incremental validity or have next to no incremental validity at all.

Adverse Impact. Lievens, Peeters, and Schollaert (2008) indicated that the issue of adverse impact, when it comes to SJTs, deals with whether EEO protected groups systematically receive lower scores than other groups. These groups are based on categories identified by various anti-discrimination laws such as race, sex and gender. This section will examine two of these categories, namely race and gender.

Research that has been conducted on adverse impact points to differences in SJT scores within the race category; however, Patterson et al. (2012) indicated that, despite this difference, SJTs have far less adverse impact compared to cognitive tests. For example, Whetzel, McDaniel, and Nguyen (2008) found differences in mean SJT scores (*SD* = .38) in favor of Whites over Blacks. The authors also found that White test-takers generally perform better on SJTs than do Hispanics (*d* = .24) and Asians (*d* = .38). Lievens et al. (2008) indicated that adverse impact of SJTs can be determined by correlating SJT scores with cognitive ability. The lower the correlations are between the

SJT score and cognitive ability, the lower the adverse impact is likely to be. Therefore, Lievens et al. argued that the differences identified at the race level can be reduced if SJTs measure non-cognitive aspects of a job. This is consistent with the findings of Whetzel et al. (2008); they noticed that differences at the race level were reduced substantially if the SJT focused on the non-academic aspects of job performance.

Findings from research also suggest that the format of SJTs could play a part in reducing adverse impact. Chan and Schmitt (1997), for instance, have found that video-based SJTs appear to result in less adverse impact than do written SJTs. The reason Chan and Schmitt provided to explain this finding is that video-based SJTs are less cognitively loaded than written SJTs. Finally, Lievens et al. (2008) identified that SJTs with behavioral tendency instructions are likely to have less adverse impact than knowledge based instructions on SJTs.

With regard to gender, research has shown that females tend to score higher on SJTs than do males, with difference in mean scores ranging from d =.1 to d = .27 (O'Connell et al., 2007; Patterson et al., 2012). Lievens et al. (2008) stated that a possible reason for the difference in mean scores could be due to gender differences that exist in terms of the personality traits triggered by the SJT questions.

Weaknesses of Situational Judgment Tests

Despite the apparent strengths of SJTs that have been discussed, there are some limitations regarding their use as well. Faking, practice effects, and coaching are discussed below as potential weaknesses of SJTs.

Faking

According to Whetzel and McDaniel (2009), faking on a selection measure refers to deliberate distortion of responses by individuals in order to score favorably. There has been debate as to whether people can fake on selection measures, and what impact it has. Whetzel and McDaniel indicated that faking does not negatively affect the validity of a test. Schmit and Ryan (1992), however, found that faking does occur in selection settings, and that it attenuates the criterion related validity of personality tests.

Nguyen, Biderman, and McDaniel (2005) examined the extent to which SJTs can be faked. The authors' findings suggest that the response instructions provided affects the extent to which SJTs can be faked. Nguyen et al. found that SJTs presented under a behavioral tendency response format could be faked, with effect sizes ranging between .15 and .34. Nguyen et al. also found that SJTs presented under a knowledge response format are relatively immune from faking. Peeters and Lievens (2005) conducted a study using college students to determine the fakabilty of SJTs and found that SJTs with behavioral tendency instructions have limited validity, which is consistent with the findings by Nguyen et al. (2005).

Whetzel and McDaniel (2009) stated that the research on faking clearly shows that people can fake on SJTs. Faking therefore can be considered to be one of the potential weaknesses of using SJTs to predict performance. However, as seen from the study by Nguyen et al. (2005), faking can be reduced by using knowledge-based instructions as opposed to behavioral based instructions.

Practice Effects and Coaching

Lievens, Peeters and Schollaert (2007) suggested that once a selection procedure gains popularity, people will seek out test coaching programs and adopt strategies to increase their chances of improving their test scores, as well as their chances of being selected. This assumption begs the question as to whether performance on SJTs can be boosted through coaching and/or practice.

There have been very few studies that have examined the effects of coaching and practice effects and, as a result, the information concerning this is rather limited. One such study by Cullen, Sackett and Lievens (2006) examined the coachability of SJTs for consideration as selection instruments in high-stakes testing. Cullen et al. concluded that performance on some SJTs could be enhanced by coaching. In terms of practice effects, Cullen et al. indicated that the retest effects of SJTs are not larger than effects for traditional tests such as cognitive ability tests.

High stakes testing may provide the opportunity for coaching to influence performance, thus making coaching a potential weakness of using SJTs. Although there is some evidence to suggest that practice effects and coaching constitutes a weakness of SJTs, research in this area is quite scarce; therefore, further studies need to be conducted in this area.

SALSA© Test Development

SALSA© was developed by Shoenfelt in 2009 in response to the growing time and costs associated with the use of an assessment center in providing students at Western Kentucky University with feedback on their leadership skills. SALSA© was developed to reflect the six leadership dimensions identified in a meta analysis of

leadership assessment centers (Arthur, Day, McNelly & Edens, 2003), and two additional dimensions, Tolerance for Stress and Ethics. The six dimensions identified by Arthur et al. are Organizing/Planning/Vision, Consideration/Team Skills, Problem Solving/Innovation, Influencing Others, Communication, and Drive/Results Orientation.

The first step in developing SALSA© was the generation of critical incidents and responses to those situations from subject matter experts (SMEs). The SMEs that were involved in this process were students from WKUs ROTC program, the Dynamic Leadership Institute program, I-O Psychology graduate students, and the honors section of Effective Leadership Studies. Students were used as SMEs because SALSA© targeted university students, and it was expected that these student SMEs would help ensure that the scenarios created would be appropriate for university students. To develop the critical incidents, the SMEs wrote descriptions of good, bad and average leadership performance to represent each of the eight dimensions. The critical incidents and responses generated by the SMEs were reviewed and edited by an I-O Psychologist to ensure that they met the specifications needed for an SJT, and that each situation was written in the same format. In all, a total of approximately 300 incidents were generated during this process.

The second step in developing SALSA© was retranslation of the critical incidents to ensure that they were clear, reliable examples of the targeted leadership dimensions. Critical incidents were retained only if SMEs agreed on the dimension an item represented. The SMEs involved in this phase of the development of SALSA© were faculty in the disciplines of Industrial/Organizational Psychology, Business, Leadership Studies, and Military Science who were knowledgeable about the field of leadership. During this stage 106 items initially survived the retranslation at an inclusion criterion of

66.7% SME agreement. However, the criterion was lowered to a greater than 50% agreement, resulting in a total of 213 items being retained.

The third step of the process was the calibration of each response option for each item with respect to the level of leadership effectiveness represented by the response option. As with the retranslation process, SMEs from Industrial/Organizational Psychology, Business, Leadership Studies, and Military Science were used. The SMEs 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). Mean SME ratings were used to indicate the effectiveness of behavior described by each response option. Only items that had one correct answer (i.e., mean of 4 and separation of at least .5 from the next best answer) were retained, resulting in a total of 130 items being retained.

After the completion of these steps, response instructions and a scoring key were developed. On average it takes approximately one hour for students to complete all of the 130 items on SALSA©. The test is scored by awarding one point for each correct answer and zero points for each incorrect answer. Dimension and total test scores are obtained by summing the correct responses for a given dimensions and across all dimensions, respectively.

Current Study

The current study focused on laying the foundation for the development of a leadership assessment tool that can be employed in a workplace setting. The items developed for SALSA© consist of hypothetical scenarios designed for a university

setting, but may be applicable to the workplace. SALSA© was used as the foundation for the current organizational leadership assessment tool. Each SALSA© item was reviewed to determine whether it was specific to a university setting or if it would generalize to the workplace. Six items that contained references to university specific events were reworded to reflect generic organizational events. Essentially, this study was a content validation of a potentially new assessment tool. The revised SALSA© items and other SALSA© items were subjected to a retranslation process and the calibration of response options for each item.

Study 1: Retranslation

The methodology employed in this study was guided by information outlined by Smith and Kendall (1963) for the retranslation of behavioral expectations.

Hypotheses

Based on the review of literature, the following hypotheses were identified for this study.

Hypothesis1: The items on SALSA© will be retranslated beyond a random level; that is, greater than 12.5% of the items will be retranslated into the same dimensions for the new leadership assessment tool.

Hypothesis 2: 75% percent of the items on SALSA© will be successfully retranslated and retained for the calibration process for the new leadership assessment tool.

Method

Participants

For Study 1, SMEs knowledgeable about the field of leadership from Western Kentucky University were presented with the items in SALSA© and asked to sort them

into one of eight dimensions of leadership. There were four SMEs that participated in the retranslation process: three females and one male. All of the SMEs held doctoral degrees, had received graduate training in leadership, and had taught university courses in leadership.

Procedure

Six items contained student-based scenarios that had to be modified to reflect scenarios that could occur within an organizational context. These six items included three items from the Results Orientation dimension, two items from the Problem Solving dimension, and one item from the Organizing dimension. In addition to these six items, other items that had weak response options (that is, easily identified as an incorrect option) were rewritten to increase item difficulty. Four industrial-organizational psychology graduate students and one industrial-organizational psychologist carried out the modification and rewriting process.

After the modification and rewriting process, all the items from each dimension in SALSA© were combined in random order in an electronic sheet. The file was then sent by email to the SMEs with instructions for completing the retranslation (see Appendix A for a copy of the retranslation instructions). As part of the instructions, SMEs were asked to read the given definitions of the eight dimensions, and then to assign each item into the dimension that it best represented. The final dimension each item was assigned to was based on a criterion of 75% rater agreement. In situations where there was no clear majority, two SMEs reviewed the items together in order to reach a rational consensus. The SMEs each spent an average of an hour to complete the retranslation exercise. After the SMEs completed the retranslation, their responses were compiled into another

electronic file with separate worksheets for each dimension. This new electronic file indicated both the original dimension prior to the retranslation, as well as the new dimensions to which each item belonged.. In all, a total of 106 out of the 130 items were retranslated back into the original dimensions. Table 1 contains the number of items retained for each dimension.

Table 1

Number of Items in Each Dimension after Retranslation

Dimension	Original Number of Items	Number of Items Retained in Same Dimension	Number of Items Reallocated from Different Dimensions	Total Number of Items in Each Dimension
Organizing/Planning/Visioning	18	17	6	23
Consideration/Team Skills	21	18	6	24
Problem Solving/Innovation	19	12	8	20
Influencing Others	11	8	2	10
Communication	12	9	0	9
Drive/Results-Orientation	25	22	1	23
Tolerance for Stress	11	8	0	8
Integrity/Ethics	13	12	1	13
Total	130	106	23	130

Results

In order to test the first hypothesis, a one sample z-test for proportion was used to compare the percentage of items that was predicted to retranslate into the same dimension to the actual percentage of items that retranslated in to the same dimension. The z-test for proportions indicates whether the proportions observed are significantly different from what would occur by chance. The formula for the z-test is as follows:

$$Z_{obs} = \frac{\text{Observed\%-Hypothesized\%-.5/n}}{\sqrt{(\text{Hypothesized\%})x(1-\text{Hypothesized\%})/n}}$$

The random level, .125, was set based on the number of dimensions, that is, eight. The z-test was significant ($Z_{obs} = 23.9 > 1.65$, $p < .05$), thus providing support for Hypothesis 1.

The aim of Hypothesis 2 was to establish how many of the items on SALSA© would need to be retranslated and retained for the calibration study for the retranslation to be considered successful. An item was considered to have retranslated successfully if it was retranslated back into the same dimension it was on SALSA©. As indicated in Table 1, 106 items on SALSA© were retained in the same dimension following the retranslation process. This represents 82% of the items on SALSA©, and therefore provides support for Hypothesis 2.

Study 2: Calibration

The purpose of Study 2 was to calibrate the response options for each successfully retranslated item from Study 1 to identify the correct answer for each item. The following hypothesis was, therefore, identified for Study 2.

Hypothesis 3: Seventy-five percent of the items will successfully calibrate. That is, seventy-five percent of the items will have a correct answer and only one correct answer

(i.e. a mean of 4 or more) separated by a mean rating of at least .3 from the next best option.

Method

Participants

For this study, ten individuals knowledgeable about the field of leadership from Claremont McKenna College and Western Kentucky University served as SMEs. There were nine female and one male SMEs. Their average age was 40.6 years (SD=14.4). All six SMEs reported receiving graduate training in leadership, averaging 10.8 years (SD= 11.8). There was an average of 7.1 (SD = 10.3) years experience in teaching leadership.

Procedure

All 130 items used during the retranslation study were grouped according to dimension and placed into an electronic sheet; that is, the items for each dimension were on a separate electronic sheet. The 23 items that did not retranslate back into their original dimension were included in the dimension identified by the retranslation process. Twenty-six additional items were included in this study in order to increase the size of the item pool. These additional items were modified to reflect workplace scenarios where necessary, and some response options were also rewritten in order to increase item difficulty. The modification of the items, as well as the rewriting of the response options, was done by four industrial-organizational psychology graduate students and an industrial-organizational psychologist.

After the modification and rewriting process, the items were compiled into a separate electronic sheet by dimension. Each dimension was divided into three item clusters. Each SME was assigned a unique electronic sheet containing one or two item

clusters from each dimension such that each SME rated responses for approximately 60 items and all items were rated by three SMEs. One dimension, Organizing, had four SMEs rate each response item. These electronic sheets were sent via email with instructions on how to complete the calibration process (see Appendix B for a copy of the calibration instructions). As part of the instructions SMEs were asked to read the given definitions of the eight dimensions, as well as 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). On average it took SMEs about an hour to rate the response options provided in each electronic file.

The ratings provided by SMEs were combined by dimension into a new electronic file, and means and standard deviations were computed for each response option. A response option was considered to be the correct answer if it had a mean of 4 or more and was at least .3 (or .25 where there were four raters for an item) better than the next best answer. This criterion ensured that there was a correct response to each item, and that there would be only one correct answer to each item. These decision rules eliminated 55 items because there was either no correct response or because there was more than one correct answer. In all, a total of 101of the 156 items were retained after the calibration process. Table 2 indicates the final number of items retained for each dimension following the calibration process.

Table 2

Number of Items in Each Dimension after Calibration

Dimension	Number of Items Pre-Calibration	Number of Items with One Correct Answer	Number of Items with More than One Correct Answer	Number of Items with no Correct Answer	Final Number of Items in Each Dimension
Organizing	23	14	1	8	14
Consideration	24	18	1	5	18
Problem Solving	20	12	4	4	12
Influencing Others	17	9	3	5	9
Communication	16	12	0	4	12
Results-Orientation	23	17	3	3	17
Tolerance for Stress	13	8	0	5	8
Integrity/Ethics	20	11	2	7	11
Total	156	101	14	41	101

Results

In order to test Hypothesis 3, a one-sample z-test for proportion was used to compare the percentage of items that was predicted to have only one correct answer separated by a mean of .3 (or .25 where there were four raters for an item) from the next best answer, to the actual percentage of items that had only one correct answer. The test value, 75%, was set based on the expectation that nearly all the items would have a

correct answer, and also because a number of items were modified to increase item difficulty.

As mentioned in Study 1, the z-test for proportions indicates whether the proportions discovered is significantly different from what would occur by chance. The formula for the z-test is as follows:

$$Z_{obs} = \frac{\text{Observed\%-Hypothesized\%-.5/n}}{\sqrt{(\text{Hypothesized\%})x(1-\text{Hypothesized\%})/n}}$$

The z-test was non significant ($Z_{obs} = -3.53 < 1.65$, $p>.05$), thus it did not provide support for Hypothesis 3. As indicated in Table 2 only 101items (64%) were retained after the calibration process.

Discussion

The purpose of this study was to ensure that the new leadership assessment tool was psychometrically sound through the retranslation and calibration of items and response options, respectively. The outcome of this process was that 106 out of the 130 items (82%) were successfully retranslated back into the same dimension. The percentage of items (18%) that were retranslated into a different dimension suggests that the eight dimensions on the leadership assessment tool are not independent. Thus, some of the items may represent more than one dimension of leadership. This finding confirms the assertion made by Patterson et al. (2012) that SJTs are construct heterogeneous, because one item on a test could encompass multiple performance dimensions. Although it is psychometrically desirable to have independent dimensions, as well as items representing only one dimension, the nature of most leadership situations is such that they are likely to involve more than one dimension of leadership.

Generally, longer tests tend to be more reliable; therefore SJTs that contain more items tend to be more reliable (McDaniel et. al., 2001). During the calibration study, a total of 55 items were removed from the test, thus reducing the number of items from 156 to 101. Although the removal of these items may have reduced the reliability of the test, several measures were put in place during the study to ensure that the reliability would not be greatly impacted by the loss of items. One of the measures taken was to improve the quality of the items before the retranslation and recalibration processes. In all, 82 of the 156 items were rewritten in order to increase difficulty, and to ensure that they would improve the overall quality of the assessment. An additional 26 items from the original SALSA© item pool that had two correct answers were modified for this study and added to the SALSA© items In addition to this, during the calibration process, a stringent criterion was put in place so that only items with at least one response rated with a mean of 4 or better were retained. Finally, only items where the best answer was at least .33 better (or .25 better in cases where four SMEs rated an item) than the next best answer were retained. Therefore, even though some items failed the retention criteria and were removed from the test, thereby shortening the test, the items that remained should be of high quality due to the steps taken prior to the studies to ensure item quality. In addition to this, it is better to have a shorter test with high quality items than to have a longer test with items of lower quality.

Shyamsunder et al. (2009) suggested that SMEs used to calibrate response options should be similar in terms of demographics to those who will take the test. During the development of SALSA©, this was not the case as the SMEs who calibrated the response options were not students. However, the current study considered the suggestion by

Shyamsunder et al. (2009) and used SMEs that were similar to those who will take the test, that is, employees in workplace settings (albeit university work settings). This will help ensure that the correct answer identified for each item actually reflects the most effective leadership response in each scenario.

In conclusion, the studies conducted appear to have resulted in the successful creation of a leadership assessment suited for a workplace environment. This assessment contains 101 items with an average of 13 items across eight dimensions. The retranslation study ensured that all the items in the assessment represent the dimensions to which they have been assigned. Finally, the calibration study ensured that each item contained four response options that reflect a range of leadership effectiveness, and that there is at least one correct answer for each item.

Limitations

There are a several limitations that affected this study. One of the potential limitations of the retranslation and calibration studies is the small number of SMEs that were used. With a much larger sample of SMEs, a higher threshold of agreement could have been useful, especially for the retranslation study.

Another potential limitation of the study involves the quality of the scenarios and response options that were modified. Although the graduate students who were involved in the modification and rewriting process had had some training in test development, it is possible that they may not have had enough experience and/or knowledge of leadership knowledge, skills, and abilities to generate high quality scenarios and response options. However, this concern was addressed by the fact that an I/O psychologist provided close

supervision, and substantially edited the modifications made by the graduate students to ensure that they were of high quality.

Another potential limitation of these studies is the level at which the criterion was set for to measure the hypotheses in study one. The criterion level, set at 12.5%, could be considered a low and/or lenient measure of success. However, the results showed an 82% success rate for items retranslating into the same dimensions. This success rate should allay any concerns about the level at which the criterion was set.

Finally, one could argue that it is difficult to capture complex leadership situations in two to three sentence scenarios. Thus, the level of leadership knowledge assessed by the new test may be limited.

Future Directions

The leadership assessment tool that has been developed is still new; therefore, more studies need to be conducted in order to evaluate it. For example, a convergent validity study using SALSA© could be conducted to ensure that this new leadership assessment is actually measuring the leadership construct it was developed to measure. Likewise, divergent validity studies should also be conducted to ensure that this new leadership assessment is not measuring some construct other than leadership. These studies would help improve the overall validity of this new assessment. Also, further studies should be conducted to assess the reliability of this assessment. According to O'Connell et al. (2007), a test-retest or parallel forms approach would be the most accurate means to measure reliability of this leadership assessment due to the fact that it has multiple dimensions. Therefore, future studies should focus on test-retest and/or

alternate forms reliability to further evaluate the reliability of this new leadership assessment.

Future research should be conducted to examine the correlation between scores on this leadership assessment and job performance. Numeric scores on performance appraisals could serve as a potential criterion measure with enough variability to determine if this new leadership assessment is related to on the job performance. This would also serve to determine whether the test is actually measuring leadership or some other construct.

Conclusion

In summary, a new SJT for measuring leadership within an organizational context has been developed. This assessment was developed to measure eight dimensions of leadership. The processes involved in the development of this assessment were the retranslation of items from an existing SJT measuring leadership, and the calibration of response options of the items. The retranslation process was carried out to ensure that each item in the assessment accurately reflects the dimension into which it was assigned. The calibration study ensured that each item contained four response options that reflect a range of leadership effectiveness, and only one correct answer. This new leadership assessment tool has the potential to assess leadership knowledge, and identify employees for employment decisions such as training and promotion. However, in order to effectively achieve this, further studies need to be conducted in order to further evaluate the psychometric properties of the assessment.

APPENDIX A: INSTRUCTIONS FOR RETRANSLATION

Thank you for agreeing to help with the revision of the Situational Assessment of Leadership: Student Assessment (SALSA). SALSA consists of a collection of hypothetical, but realistic leadership scenarios.

We are asking for your assistance in assuring that the leadership scenarios represent specific dimensions of leader behavior. To accomplish this, we have attached an Excel file that contains 130 leadership scenarios. (You will notice that each scenario is followed by four potential leadership responses. These responses will be used in a later stage of this project – they will become the response options on the Situational Judgment Test.)

After you have assigned a dimension to each leadership scenario, please save the Excel file and return it to betsy.shoenfelt@wku.edu. All responses will be anonymously aggregated into a single Excel file to determine the consensus of which dimension is represented by each leadership scenario.

We need you to**:**

1.Please carefully read the definitions of the 8 dimensions of leadership.

(You may want to print the attached document with the dimension definitions and keep it in front of you.)

2. Please read each brief scenario and decide which dimension of leader behavior is best represented by the scenario. Mark the number of the leadership dimension in the column to the left of the scenario (column A).

3. Please save the file and return it to betsy.shoenfelt@wku.edu by close-of-business this Friday, March 28th.

We estimate that it will take about 60 to 90 minutes to complete this task. Please let me know if you have any questions. This is an important project and your time is greatly appreciated. Thank you!

APPENDIX B: INSTRUCTIONS FOR CALIBRATION

We very much appreciate your help in developing the business form of the leadership situational judgment test (SJT).

The attached files are for the calibration step in the test development. That is, Subject Matter Experts (SME) are asked to read test items and to rate the four response options for that item in terms of leadership effectiveness (1 = Extremely Ineffective to 5 = Extremely Effective). These ratings will form the basis for identifying the correct answer to the test item.

Each of the attached files contains a unique set of items. Rather than asking each SME to rate all 150 items, each SME will rate approximately 1/3 of the items (i.e., ~ 50 items). To ensure we have three sets of ratings for each item, we need 8 raters. You indicated that you have some faculty and advanced graduate students who are willing to serve as raters. We very much appreciate your and their help.

You should assign each SME/rater to a different Excel file. That is, each Excel file is unique – so each of the 8 raters should get a specific, unique Excel file.

(We do not necessarily need to know who the raters are, but if you keep track of who is assigned to which file, we can let you know which files are missing, should any rater need a reminder.)

Please note there are 10 tabs/worksheets in the Excel Files.

The first tab contains the Directions for the task.

The second tab contains Demographic items.

The third through ninth tabs contain a subset of test items for each of the 8 dimensions of leadership on the (SJT).

Please ask the SMEs to read the directions and answer the demographic items.

For each tab for the test dimensions:

First, read the definition of that dimension of leadership.

Then, for each test item, read the item and rate EACH response option from 1 (Very Ineffective) to 5 (Very Effective) in terms of leadership behavior.

After the SME/rater has rated each response option for each item in each dimension, s/he should save their completed file and return it to me at: betsy.shoenfelt@wku.edu

It will likely take about an hour to complete these ratings. We are working under a deadline to ensure my graduate student has these data to complete his thesis by May. We would very much appreciate receiving the completed ratings by Tuesday, April 15th.

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.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly*, 421-458.

Callinan, M., & Robertson, I.T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8,* 248-160.

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

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

Cullen, M.J., Sackett, P.R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14*, 142-55.

Gowing, M. K., Morris, D. M., Adler, S., & Gold, M. (2008). The next generation of leadership assessments: Some case studies. *Public Personnel Management, 37*, 435-455.

Grant, K. L. (2009). *The Validation of a Situational Judgment Test to Measure Leadership Behavior* (Unpublished master's thesis). Western Kentucky University, Kentucky.

Kanning, U. P., & Kuhne, S. (2006). Social desirability in a multimodal personnel selection test battery. *European Journal of Work and Organizational Psychology*, *15*, 241-261.

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., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology*, *90*, 442.

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

McDaniel, M.A., & Nguyen, N.T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment,9,* 103-13.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A., & Braverman, E.P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730-40.

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*, 63-91.

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

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

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment*, *13*, 250-260.

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.

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgment tests to assess non-academic attributes in selection. *Medical Education, 46,* 850-86.

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

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.

Rynes, S. L., & Connerley, M. L. (1993). Applicant reactions to alternative selection procedures. *Journal of Business and Psychology*, *7*, 261-277.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH:Author.

Schmit, M. J., & Ryan, A. M. (1992). Test-taking dispositions: A missing link? *Journal of Applied Psychology, 77*, 629−637.

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-74.

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.

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.

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

Weekley, J.A., & Ployhart, R.E. (2006). *Situational Judgment Tests: Theory, Measurement and Application*, Jossey-Bass, San Francisco, CA.

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

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, *19*, 188-202.