

5-2010

A Situational Assessment of Student Leadership: An Evaluation of Alternate Forms Reliability and Convergent Validity

Patricia Slack

Western Kentucky University, patricia.slack100@wku.edu

Follow this and additional works at: <http://digitalcommons.wku.edu/theses>



Part of the [Educational Assessment, Evaluation, and Research Commons](#), and the [Industrial and Organizational Psychology Commons](#)

Recommended Citation

Slack, Patricia, "A Situational Assessment of Student Leadership: An Evaluation of Alternate Forms Reliability and Convergent Validity" (2010). *Masters Theses & Specialist Projects*. Paper 172.
<http://digitalcommons.wku.edu/theses/172>

This Thesis is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Masters Theses & Specialist Projects by an authorized administrator of TopSCHOLAR®. For more information, please contact topscholar@wku.edu.

A SITUATIONAL ASSESSMENT OF STUDENT LEADERSHIP: AN EVALUATION
OF ALTERNATE FORMS RELIABILITY AND CONVERGENT VALIDITY

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
Patricia Slack

May 2010

A SITUATIONAL ASSESSMENT OF STUDENT LEADERSHIP: AN EVALUATION
OF ALTERNATE FORMS RELIABILITY AND CONVERGENT VALIDITY

Date Recommended _April 22, 2010_____

_Elizabeth Shoenfelt_____
Director of Thesis

_Reagan Brown_____

_Jacqueline Pope-Tarrence_____

Dean, Graduate Studies and Research Date

Table of Contents

Abstract	ii
Introduction.....	3
Study 1	23
Method	23
Results	25
Study 2	28
Method	28
Results	29
Discussion	30
References.....	34
Appendix A.....	37
Appendix B.....	39
Appendix C	41
Appendix D.....	48

A SITUATIONAL ASSESSMENT OF STUDENT LEADERSHIP: AN EVALUATION OF ALTERNATE FORMS RELIABILITY AND CONVERGENT VALIDITY

Patricia Slack

May 2010

52

Directed by: Elizabeth Shoenfelt, Reagan Brown, and Jacqueline Pope-Tarrence

Department of Psychology

Western Kentucky University

The Situational Assessment of Leadership: Student Assessment (SALSA©) was developed in the spring of 2009 to be used as a measure of student leadership. Study 1 assessed alternate forms reliability of the SALSA using scores from 178 students. The overall scores on SALSA Form A and SALSA Form B showed a significant correlation ($r_{AB} = .906, p < .01$). Dimension scores on the two forms ranged from $r_{AB} = .475$ to $r_{AB} = .804$. Study 2 evaluated the convergent validity between the SALSA and the Western Kentucky University Center for Leadership Excellence assessment center. SALSA scores as well as assessment scores from 53 students were analyzed. The overall scores on the SALSA and CLE assessment center had a significant yet moderate correlation ($r = .513$). Dimension correlations were significant but low, ranging from $r = .310$ to $r = .392$. The strong correlations in Study 1 indicate the two forms of the SALSA may be used as alternate measures such as in a pre and post-test of leadership. The convergent validities in Study 2 demonstrate that both the SALSA and assessment center may be used to assess leadership. However, the low convergent validities across dimensions indicate overall scores likely should be used rather than dimension scores.

A Situational Assessment of Student Leadership: An Evaluation of Alternate Forms Reliability and Convergent Validity

The Center for Leadership Excellence (CLE) on the campus of Western Kentucky University provides students with a unique experience. This program was designed to enhance available leadership experiences and to provide leadership training and education to students. The current means of leadership evaluation available in this program is an assessment center. Gowing, Morris, Adler, and Gold (2008) described the assessment center process. During the assessment center, participants are evaluated by trained assessors on a number of different competencies. They are observed performing multiple exercises and rated based on the consensus of the assessors.

Although the assessment center process is considered a valid process for measuring leadership, it is costly and time consuming. Therefore, a situational judgment test of leadership, Situational Assessment of Leadership: Student Assessment (SALSA®), was developed in the Spring of 2009 and may be used as an alternate method for evaluating leadership qualities in individuals. This assessment tool was designed to measure eight leadership dimensions: Organizing/Visioning/Planning, Consideration/Team Skills, Problem Solving/Innovation, Influencing Others, Communication, Drive/Results Orientation, and Tolerance for Stress. Descriptions of these dimensions can be found in Appendix B.

Given that leadership is assessed by CLE at the beginning and end of a given semester, two different situational judgment tests are needed for this purpose. Therefore, there is a need to create alternate forms of SALSA. This study will develop alternate

forms as well as assess the convergent validity of the SALSA with CLE assessment center scores.

This paper will discuss the current research on situational judgment tests, including evaluations of validity, response formats, uses, and issues that are addressed. The development of alternate forms of a test will be discussed, as well as the process of evaluating the convergent validity of a measure. Thus, the current studies are intended to develop alternate forms of the SALSA as well as to evaluate the convergent validity of SALSA and the CLE assessment center.

Situational Judgment Tests

Situational judgment tests are becoming more popular as predictors of performance. Typically, situational judgment tests (SJTs) present an applicant with several different situations they might encounter on the job. These situations are usually derived from critical incidents or some other method of job analysis. The respondent is then asked to select the best, or sometimes the worst, response to such a situation from a list of possible responses. Scoring of these tests can be done in a variety of ways and will be discussed later. SJTs are usually paper and pencil format, but some SJTs have been developed into a video-based format (Weekley & Ployhart, 2005). Different aspects of situational judgment tests will be discussed as well as benefits of and issues with this testing format. The development of SJTs as well as types of response formats will be discussed first.

Development of SJTs

The most common type of SJT is a multiple response option format. Motowidlo, Crook, Kell, and Naemi (2009) proposed that situational judgment tests could be

administered more easily with a single response option format. The authors claimed that SJTs with a multiple response option format are time consuming and difficult to create. With a single response format, single-response alternatives are derived directly from critical incidents. Thus, a less expensive alternative is available in comparison to a multiple response format. With a single-response format, applicants are presented with a scenario and are asked to rate the effectiveness of that behavior. The difference between their rating and the SME mean rating provided through the development process is then calculated and squared. The mean squared difference across all items is the respondent's final score.

Weekley stated that multiple response SJTs can have anywhere from 3 to 12 (or even more) response options for each item (as cited in Motowidlo et al., 2009). Flanergan wrote that single response SJT items could be created with much less effort than multiple response SJT items by following a well-known process of using critical incidents to develop performance dimensions and rating scales (as cited in Motowidlo et al.).

Motowidlo et al. (2009) described the process of developing the response items. First, critical incidents were collected from subject matter experts. The experts in this study were 26 supervisors and administrators that worked with different volunteer organizations. They were asked to remember a time when they saw a volunteer working with someone who needed help and their behavior was either very effective or very ineffective. The final set of critical incidents consisted of 100 items. The items were then put into behavioral categories. Three researchers reviewed the items independently then discussed their decisions with each other to reach a consensus. The final three dimensions were work effort, professionalism, and personal skill. Next, thirteen volunteer supervisors

and administrators (some of whom had participated in the generation of the critical incidents) completed a questionnaire that asked them to rate the critical incidents in two ways. The SMEs were asked to decide in which of the three dimensions the critical incident belonged. Next, they were asked to rate the critical incident on effectiveness using a 7-point rating scale, which ranged from 1 being very ineffective to 7 very effective.

Motowidlo et al. (2009) found that it was possible to predict job performance with a single response SJT based on critical incidents. Given that this process can be done with less effort and less cost, it is possibly a useful alternative to the multiple response formats. The authors also found that work effort SJT scores were more highly correlated with performance ratings of work effort ($r = .28$) than with performance ratings of professionalism ($r = .02$) or personal skill ($r = .04$). Another significant correlation was between the SJT measure of procedural knowledge of work effort and the personality trait of conscientiousness ($r = .26$).

Response Instructions of SJTs

Response instructions can vary across SJTs. Nguyen, Biderman, and McDaniel (2005) examined how response instructions affected faking on a SJT. There are two types of response formats for multiple response SJTs. Those with a knowledge format present respondents with a scenario and instruct them to indicate what they should do by choosing the best or worst answer from a list of possible responses. A behavioral tendency format consists of a scenario and a list of behaviors, and direct respondents choose the behavior they would be most or least likely to exhibit.

Nguyen et al. (2005) proposed that behavioral tendency formats would be easier to fake than knowledge formats. They also believed the SJT would be more difficult to fake than the Big Five dimensions of personality. Their last supposition was that the knowledge-based scores would correlate more with cognitive ability than would scores from behavioral tendency SJTs. Student participants completed the Wonderlic Personnel Test, a measure of cognitive ability, and the SJT measures. Respondents were asked to complete these measures twice. Half of the respondents were asked to complete the SJT once honestly and once “faking good.” The other half of respondents were asked to complete the measures in the same way, except they were asked to “fake good” the first time they took the SJT and honestly the second time.

Nguyen et al. (2005) found that the SJT could be faked when a behavioral tendency format was used. The faking effect size for this format was between .15 and .34. The order in which the participants completed the measures affected results. For example, researchers found a positive faking to honesty difference for participants that received instructions to “fake good” first. Those who received the “honest” instructions first showed a negative faking to honest difference. Also, when participants received the honest instructions first, researchers found the knowledge response format was correlated more highly with cognitive ability ($r = .56$) than the behavioral format ($r = .38$).

Researchers also examined the type of instructional format used in SJTs and their effectiveness. McDaniel, Hartman, Whetzel, and Grubb (2007) conducted a meta-analysis to examine the relationship between response instructions and construct and criterion related validity of SJTs. They proposed that response instructions would affect the correlations of constructs with the SJT. To be specific, they hypothesized that

knowledge instructions would correlate higher with cognitive ability than would the behavioral tendency format. They also hypothesized that behavioral tendency instruction formats would correlate higher with personality traits than would knowledge formats.

McDaniel et al. (2007) found, in support of one of their hypotheses, that SJTs with behavioral tendency formats correlated higher with the Big Five measure of personality than did knowledge formats. For this relationship, instructions acted as a moderator for the personality components of agreeableness, conscientiousness, and emotional stability. However, there was a different finding when analyzing criterion related validity. Researchers found that the two different instruction formats did not significantly affect the criterion related validity of the SJT. Therefore, the researchers concluded that instruction formats do not moderate criterion related validity. One possible reason for this finding is that applicants taking a SJT may ignore the response instructions of behavioral tendency tests and respond as though the instructions represent a knowledge format. A second possible reason for this finding is that some aspects of job performance can be predicted by either cognitive ability or personality.

Validity of SJTs

Other studies have been conducted evaluating the validity of SJTs. Weekley and Ployhart (2005) examined a SJT in terms of incremental validity and correlations with multiple predictors. Three groups of individual differences measures were used in this study. Cognitive ability was described as general mental ability or grade point average. Personality consisted of the Big 5 measures, which were conscientiousness, emotional stability, agreeableness, extroversion, and openness to new experiences. Experience was operationalized as general work experience, job tenure, and training experience.

Some 27 employees in loss prevention management in a large retail organization were asked to complete the predictor battery during working hours. The battery consisted of a 40-item measure of cognitive ability, a 58-item test of situational judgment, and a 158-item measure of training experience. The cognitive ability measure consisted of 10 items for each of the following: math problems, analogies, word identification problems, and reasoning problems. Grade point average was measured on a five-point scale with 1 being mostly As and 5 being mostly Fs. While the battery was being completed, supervisors provided performance ratings of the participants (Weekley & Ployhart, 2005).

For each SJT item, participants were to select the option they considered to be the best and the option they considered to be the worst. Respondents received one point for getting the “best” answer correct and one point for getting the “worst” answer correct. If the respondents chose the best answer as the “worst” or the worst answer as the “best” they lost 1 point. Therefore, for each situation the participants’ score could range from -2 to +2 (Weekley & Ployhart, 2005).

To assess the five-factor model of personality, researchers used a 125-item inventory, with 25 items measuring each of the Big Five. Job tenure was measured by asking the number of years the participant had worked in the current position. A five-item measure was used to determine general work experience. Training experience was reported using self-report measures consisting of 158 items. These items were short descriptions of areas of knowledge in safety and security, general management, and policies and procedures specific to the company (Weekley & Ployhart, 2005).

Weekley and Ployhart (2005) found that the SJT measured aspects of ability, experience, and personality that were related to managerial job performance indicated by

cognitive ability ($r = .36$), general work experience ($r = .21$), job tenure ($r = .13$), training experience ($r = .12$), conscientiousness ($r = .13$), emotional stability ($r = .17$), and extroversion ($r = .14$). The SJT was also found to reflect general forms of knowledge. The SJT demonstrated incremental validity over the cognitive ability, personality, and experience ($\Delta R^2 = .18$), suggesting that the SJT captured something unique, such as knowledge (Weekley & Ployhart).

Other Uses of SJTs

Situational Judgment Tests are most commonly used to measure job performance, but can be used to measure other constructs as well. Becker (2005) conducted a study evaluating a SJT meant to measure employee integrity. To do this, Becker created a 20-item SJT involving potential workplace dilemmas. This instrument took a total of 20-30 minutes to complete. The first set of subjects in this study consisted of 307 business students enrolled in upper-level courses. Data gathered from these respondents were used to develop an empirical scoring key. These respondents completed the measure along with self-report ratings of integrity. Ratings of integrity were also collected from persons who knew the respondents well. The scoring key was then created. The scores on the items were correlated with the integrity ratings. If the correlation was positive and significant, the response was given a score of “1.” If the correlation was significant and negative, the response was given a “-1.” If there was no significant correlation, the response was given a score of zero.

The participants for the second part of Becker’s (2005) study were employees from restaurants in a fast-food service chain, workers from a large plant that manufactured plastics, and mechanical engineers from the same manufacturing plant. All

of the participants completed the SJT instrument while their managers completed a 30-item rating measure of the integrity of the respondents. The results indicated scores on the SJT were valid predictors of outcomes related to integrity in real world situations. Also, these scores were found to predict rates of promotion, progress in the given career, and team leader status.

As mentioned earlier, SJTs can be used to examine personality traits. However, SJTs can also be used to evaluate underlying theories as well. Motowidlo, Hooper, and Jackson (2006) used a SJT to look at a connection between personality traits and behavioral effectiveness. These researchers developed a theory, the implicit trait policy (ITP), to show why SJT answers are often correlated with personality traits. They believed that when expert judgments are used as a scoring key, procedural knowledge is being measured. The theory goes on to state that if individual differences in personality are exhibited in the answers on the SJT, they will affect the weight of the scores. For example, if agreeableness is shown in an answer, a person who is agreeable will rate the answer higher than a person who is not.

Motowidlo et al. (2006) developed items for this SJT specifically to demonstrate personality traits. Of the item stems developed for this SJT, five were developed to show extraversion, five were developed to show agreeableness, and six were developed to show conscientiousness. These stems were written to allow respondents to display these traits if they had them. Forty-four undergraduates wrote responses to these stems indicating the behavior they would exhibit if they were in the situation described in the stem. A second sample of 63 undergraduates participated in simulated interviews, were told the situation, and asked how someone should react to it. The responses were then

written to show both high and low levels of this trait using the examples collected from the undergraduates. For each stem, there were 5 to 10 response items. Half of the responses for each item showed low levels of the trait, while the other half showed high levels. In the final set of responses, 39 demonstrated extraversion, 36 demonstrated agreeableness, and 45 demonstrated conscientiousness.

The next step in the process required seven graduate students to rate the SJT response options according to the trait they were meant to express. Respondents used a 7-point scale to do so. For example, to measure agreeableness, the item was rated from 1 to 7, with 1 being very disagreeable and 7 being very agreeable. Finally, 96 undergraduates were asked to rate the SJT. Respondents read every stem and every response option. They were asked to rate each response option on a scale of 1 to 7, with 1 being very ineffective and 7 being very effective. The participants were also asked to complete the NEO-FFI extraversion, agreeableness, and conscientiousness scales. ITP scores for each respondent were calculated by correlating the effectiveness of the response with his or her scores on the personality scales. ITP scores on both extraversion ($r = .39$) and agreeableness ($r = .30$) correlated more with their corresponding personality traits on the NEO-FFI than with scores on other personality traits (Motowidlo et al., 2006).

Motowidlo et al. (2006) conducted a second study using SJT items that were created by a government agency to predict performance. Nineteen of these items allowed for the expression of agreeableness and 19 others allowed for the expression of conscientiousness. Answers were developed by incumbents and were reviewed by 10 industrial/organizational psychologists and 2 research assistants. The reviewers rated responses on how they showed either agreeableness or conscientiousness. One hundred

undergraduates then rated each response on the same 7-point scale used in the first study. They also completed the agreeableness and conscientiousness scales of the NEO-FFI. Results in both these studies showed that ITP measuring agreeableness can predict agreeableness as a behavior. People that showed high levels of agreeableness chose the agreeable responses to be more effective than other responses. The same was not true for extraversion and conscientiousness.

Bledow and Frese (2009) examined the use of a SJT in measuring personal initiative. Researchers proposed that this SJT would positively relate to ratings of personal initiatives given by supervisors. They also believed the SJT scores would be positively related to ratings of overall performance. Participants completed a situational judgment test of personal initiative (SJT-PI) self-ratings of personal initiative, a five-item scale of helping, a generalized self-efficacy measure, and a felt responsibility scale. Supervisors rated the participants on overall performance, personal initiative, helping and conscientiousness, and also completed a demographics section. The SJT-PI was found to be a valid measure of personal initiative ($r = .29$). The SJT also was positively related to the supervisors' ratings of overall performance ($r = .37$). This indicated a relationship between personal initiative and work performance, which may be noted by supervisors (Bledow & Frese). This study is a demonstration of how a SJT can be used to measure variables other than job performance.

Scoring of SJTs

Once a SJT has been developed, it is necessary that there is some method of scoring available. Bergman, Drasgow, Donovan, Henning, and Juraska (2006) conducted a review of literature and a study relating to the different types of scoring keys available

for SJTs. Most of the time, SJTs do not have answers that are objectively correct. The answers are usually scored by which is the best, rather than which answer is “right.”

There are many methods for determining the best answer choice. Hogan stated that in empirical scoring methods, items are scored based on how they are related to a criterion measure (as cited in Bergman et al.). Theoretical scoring keys can reflect a certain theory, or a theory can be used when deciding the best or worst answer for an item. If an answer reflects the theory, it is correct, but if it contradicts the theory, it is deemed incorrect (Bergman et al.). Hough and Paullin found theoretical keys could be more susceptible to faking (as cited in Bergman et al.).

Hybridized scoring uses two different keys developed independently. One is deemed primary and the other secondary. The keys could then be used in different ways. For example, a positive score on one key could counteract a negative score on the other, causing them to cancel out. Another option is the secondary key would come into play when a zero score is received for an item using the primary key. When an empirical key is used with a theoretical key in a hybridized fashion, some issues are resolved. It decreases reliance on empiricism while at the same time recognizing theory (Bergman et al., 2006).

Expert based scoring creates keys based on the judgment of subject matter experts (SMEs). In this situation, SMEs make judgments based on the response’s relevance to the criterion. SMEs examine each item and response and choose the best or worst choices based on their expert opinion. The items they choose are deemed correct or incorrect and the rest of the response options are given a score of zero. Another similar approach compares the judgments of novices to the judgments of experts. Items chosen frequently

by experts are marked as correct, while options chosen frequently by novices (and not experts) are deemed incorrect (Bergman et al., 2006).

Bergman et al. (2006) evaluated a Leadership Skills Assessment (LSA) to find out which type of scoring key would be the most effective for this particular measure. The participants in this study were 181 non-academic supervisors who oversaw departments such as housing, building services, and student affairs. All participants completed the LSA as well as other measures. Their supervisors, however, completed the criterion measures. The LSA used in this study was a computerized multimedia assessment of leadership skills consisting of 21 items. Respondents were shown scenarios either in an office or in a manufacturing environment. Each environment had a unique four person team with a leader. Participants were shown the scenario, given four response options, and asked to choose how they would react if they were in the leadership position shown. Other measures used were the Wonderlic Personnel Test and the Sixteen Personality Factor Questionnaire (16PF). Supervisors rated leadership performance and overall performance for each respondent.

Bergman et al. (2006) created several different scoring keys for the purpose of this study. Researchers found the empirical ($r = .25$), SME ($r = .32$), hybrid initiating structure ($r = .17$), hybrid participation ($r = .22$), and hybrid empowerment ($r = .17$) keys demonstrated significant and moderate correlations with leadership ratings. They also noted the Wonderlic Personnel Test predicted leadership ratings significantly ($r = .32$). There were 2 predictors with gender differences, with females ($M = 5.43$) scoring higher than males ($M = 4.50$) on extraversion and males ($M = 6.30$) scoring higher than females ($M = 4.96$) on tough-mindedness. The empirical ($\Delta R^2 = .14$), SME ($\Delta R^2 = .16$), hybrid

participation keys ($\Delta R^2 = .13$), as well as the hybrid initiating structure key ($\Delta R^2 = .14$) showed significant incremental validity over personality measures and cognitive ability. However, the authors suggested that for developing keys, others should assess incremental validity, adverse impact, and construct validity. Using these analyses, the best keys for other SJTs can be identified (Bergman et al.).

Issues with SJT Use

There is a great deal of concern about subgroup differences that occur with SJT use. A 2007 study by O'Connell, Hartman, McDaniel, Grubb, and Lawrence was conducted for several reasons. First, the authors wanted to respond to a need for more research on racial and gender differences on SJTs. Second, they wanted more research on the incremental validity of SJTs in relation to cognitive and personality predictors. Last, authors wanted to examine the validity of SJT in terms of task and contextual performance.

The participants in this study were from seven different manufacturing companies. The measures completed by respondents were a cognitive test, personality tests, and a SJT consisting of 10 scenarios; ratings of contextual and task performance were completed by supervisors. Contextual performance related to dimensions such as leadership and teamwork, while task performance was based more on technical knowledge (O'Connell et al., 2007).

O'Connell et al. (2007) found racial differences in their data. The results showed the Black-White mean difference on the SJT to be $d = .38$. This is consistent with earlier findings from Nguyen, McDaniel, and Whetzel (as cited in O'Connell et al.). Gender differences, however, seemed to be larger than noted by Nguyen et al. (as cited in

O'Connell et al.). This difference favored females and had a $d = -.27$. Researchers found that for task performance, it would be beneficial to have respondents complete a SJT in conjunction with a cognitive ability test ($r = .18$), as it adds to the validity of such a test. Oddly, it would also be useful to supplement a personality test with a SJT ($r = .22$), but only if it is not being used with a cognitive ability test. The SJT only adds to the validity of these tests if they are being used separately. Researchers also noted that the SJT was much better at predicting contextual performance than task performance. Also, for predicting contextual performance, cognitive ability validity ($r = .07$) increased when used in conjunction with a SJT ($r = .11$).

The research into subgroup differences continued. Whetzel, McDaniel, and Nguyen (2008) conducted an analysis of race and gender differences on SJTs to address subgroup differences. Whetzel et al. investigated response instruction format and personality loading as moderators of these differences. Researchers defined cognitive loading on a SJT as how well the test correlated with cognitive ability. Personality loading was defined in the same manner, that is, how well the SJT correlates with some measure of personality.

Whetzel et al. (2008) conducted two different studies. The first was a meta-analysis. The studies used in this analysis were those using participants that were employees, applicants, or students. The format of the SJT in the study could be either written or video-based. Researchers found that, in general, there were mean race differences across studies. These differences appeared to be greater when the test had higher cognitive loading and low emotional stability loading. The higher the loading of emotional stability, the lower the race differences. These differences also appeared to be

moderated by response instructions. The results are most clear for the Black-White comparison because of the large amount of data available. Mean differences for this comparison were significantly linked to cognitive loading on the SJT. Knowledge instruction SJTs were found to have larger mean race differences than were the behavioral tendency formats, likely because of the fact that knowledge based formats have higher cognitive loading than behavioral tendency formats.

Whetzel et al. (2008) also found gender differences. On average, female test takers performed better on SJTs than male test takers. Cognitive ability was not found to be a moderator of gender differences. Conscientiousness and agreeableness were the only personality factors found to be moderators. The more the tests were loaded on these two constructs, the greater the gender differences. Response formats were found to be moderators, but only slightly, as the gender differences related to formats were small. In the second study, Whetzel et al. (2008) noted that because of mean racial differences found in the SJT results, adverse impact might occur if these tests are used in selection decisions. A simulation was conducted to estimate the adverse impact. Results indicated to avoid adverse impact employers would need to hire 70 of 100 Black applicants, 46 of 100 Hispanic applicants, and 58 of 100 Asian applicants.

Situational judgment tests are a tried and true method of evaluating job performance. They also have been shown to have many other uses. Whether they are measuring personal integrity or predicting behaviors, they have exhibited a great deal of validity. It is apparent from this research that a knowledge format is the most effective type of SJT and is less susceptible to faking than a behavioral tendency format.

SALSA ©

The SJT used in this research is the Situational Assessment of Leadership: Student Assessment (SALSA). The SALSA was developed by Shoenfelt (2009) to assess the seven most common dimensions of leadership identified by Arthur, Day, McNelly, and Edens (2003). These dimensions are Organizing/Visioning/Planning, Consideration/Team Skills, Problem Solving/Innovation, Influencing Others, Communication, Drive/Results Orientation, and Tolerance for Stress. An eighth dimension was included, Integrity/Ethics. SALSA consists of approximately 130 items with 10-20 items per dimension. The development of SALSA was based on the critical incident technique (Flanagan, 1954). SMEs used to generate critical incidents consisted of students in the Industrial/Organizational Psychology Masters Program, Honors Leadership Students, and student members of the Dynamic Leadership Institute. These critical incidents consisted of a situation, as well as three to four response options for each (Grant, 2009). The scoring key was developed using the process suggested by Motowidlo, Dunnette, and Carter (1990) and Lievens, Peeters, and Schollaert (2008). Seven university faculty members with a great deal of experience teaching leadership served as SMEs and rated each response option in terms of effectiveness. Only options that were determined to have one correct response were selected for the final version of SALSA.

The SALSA asks respondents to choose the best response to a leadership situation in terms of leadership effectiveness. This is consistent with a knowledge-based format described by Nguyen et al. (2005) and is therefore a predictor of cognitive ability. Previous research found the internal consistency of the SALSA to be $\alpha = .91$ (Grant,

2009). When assessed in terms of difficulty, it was found that across dimensions there were nearly an equal number of items that were easy, moderate, or difficult. Grant also examined the convergent validity between the CLE assessment center scores and SALSA scores. Convergent validity is how well scores on one measure of a construct correlate with scores on a second measure of the same construct. Across dimensions these validities ranged from $r = .28$ to $r = .44$, which indicated low but significant correlations. A limitation of Grant's study was that there was a small sample size ($N = 40$). The current research will reassess the convergent validity of the SALSA with CLE assessment center scores. In this study, the two measures are the SALSA and CLE assessment center, and the construct is leadership.

Alternate Forms Reliability

The need for two separate versions of the SALSA was stated earlier. Equivalent forms of the same test can be established by using the alternate forms method of estimating reliability (Murphy & Davidshofer, 1988). This method will lead to the development of alternate test forms that are essentially equivalent in terms of content, response process, and statistical characteristics. The process involves administering one form of the test to a group of individuals; at a later time, giving the second version of the test to the same group of individuals; and then correlating the scores on the two versions of the test. Given that the two forms of the test are different, the problems with the test-retest method (giving the same test twice at different times) are less of an issue. For example, carry-over effects occur when a test is given twice and the respondents remember answers from the first test, which affects their scores on the second test. As the test items on alternate forms are different, the long interval of time between tests

necessary for the test-retest method is not necessary. With alternate forms, respondents may remember similar test items, but any effect on scores on the second test would be minimal. Reactivity effects that occur with the test-retest method will be reduced as well (Murphy & Davidshofer).

Murphy and Davidshofer (1988) identified some issues with alternate forms reliability. The need for two separate test administrations results in the alternate forms administration to be as costly as the test-retest method. It is costly to develop several alternate forms of a test. However, in this study, a single version of the test will be administered to participants. Test items will be divided into alternate forms, thereby reducing the cost of multiple administrations. To reduce the error likely from simply splitting the test in half, the items will be first grouped by dimension and matched on difficulty within a dimension. One item from each matched pair will be randomly assigned to one of two forms of the test. The scores on two forms of the test will be correlated to determine equivalency.

The Current Research

This research is divided into two studies. The first study evaluated alternate forms of the SALSA. This process required grouping items within dimensions and then pairing them matched on difficulty. The items were matched in accordance with the groupings determined in Grant (2009). Two methods were used to assess difficulty: using SME ratings obtained through the initial development process and using p-value (the percent of test-takers who got the item correct). Once the items were matched, they were randomly assigned to one of two forms (SALSA Form A or SALSA Form B). The scores on the two forms were then evaluated using a bivariate correlation.

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

The second study conducted in this research evaluated the convergent validity between SALSA scores and the CLE assessment center scores. A bivariate correlation of SALSA scores and assessment center scores was conducted to test the second hypothesis:

H2: There will be a positive relationship between assessment center scores and scores on SALSA (overall and for each dimension).

Study 1

Method

Participants. Data in this study were previously collected from students enrolled in classes in the Center for Leadership Excellence leadership and university ROTC members. A total of 178 participants responded to the SALSA situational judgment test. Of these, 22 participants did not complete all eight dimensions of the SALSA, thus their results were excluded.

Procedure. SALSA items were assigned to alternate forms by Grant (2009). SALSA items were categorized into the eight dimensions assessed by SALSA. Pairs of items within a dimension were matched on difficulty. There were two indices of difficulty used by Grant. The first was determined from SME ratings. When SALSA was developed, SMEs rated each of the four possible response options for each SJT item on a 5-point scale of leadership effectiveness. The discrepancy between ratings of the best and second best choices was used as a measure of difficulty, with a smaller difference indicating a more difficult item. The second difficulty rating used was p-value, which was determined by the percentage of participants who got the item correct. The items were matched on difficulty, then one item from each pair was randomly assigned to Form-A and the other to Form-B. In the current study, the allocation of items to forms was done in accordance with the difficulty determinations and item assignments made in Grant (2009). A listing of the difficulties as well as the item assignment to the two forms may be found in Appendix C. When an item did not match another item in the same dimension (i.e., there were an odd number of items in the dimension), that item was assigned to both forms.

Results. Hypothesis 1 proposed that there would be a positive relationship between scores on the two forms of SALSA©. The bivariate correlation between the total scores on the two forms was significant ($r_{AB} = .906, p < .01$). When the items with no match were removed from both forms of the test, the correlation between total scores was lower, but significant ($r_{AB} = .857, p < .01$). The means and standard deviations overall and for each dimension are reported in Table 1.

The correlation between SALSA Form A and SALSA Form B scores for each dimension were also examined. All eight of these correlations were significant: Problem Solving/Innovation ($r_{AB} = .636, p < .01$), Influencing Others ($r_{AB} = .475, p < .01$), Verbal/Non-Verbal Communication ($r_{AB} = .686, p < .01$), Consideration/Team Skills ($r_{AB} = .589, p < .01$), Organizing/Visioning/Planning ($r_{AB} = .523, p < .01$), Results Orientation ($r_{AB} = .804, p < .01$), Integrity/Ethics ($r_{AB} = .499, p < .01$), and Tolerance for Stress ($r_{AB} = .485, p < .01$). The dimension correlations were also computed with the duplicate items removed. All eight of the correlations were significant as well: Problem Solving/Innovation ($r_{AB} = .421, p < .01$), Influencing Others ($r_{AB} = .351, p < .01$), Verbal/Non-Verbal Communication ($r_{AB} = .483, p < .01$), Consideration/Team Skills ($r_{AB} = .479, p < .01$), Organizing/Visioning/Planning ($r_{AB} = .426, p < .01$), Results Orientation ($r_{AB} = .699, p < .01$), Integrity/Ethics ($r_{AB} = .370, p < .01$), Tolerance for Stress ($r_{AB} = .310, p < .01$). These correlations are reported in Table 2.

The SALSA scores were evaluated for internal consistency reliability using coefficient alpha, both overall and within dimensions. The overall SALSA produced a

reliability of $\alpha = .922$, showing strong internal consistency. SALSA Form A produced a reliability of $\alpha = .854$, while SALSA Form B showed an internal consistency of $\alpha = .870$. Coefficient alpha overall and for each dimension for both forms are reported in Table 2.

Table 1

Means (Standard Deviations) for SALSA Form A and SALSA Form B

Dimension	Form A	Form B
Overall (130 items)	42.13 (10.04)	42.53 (10.73)
Organizing/Planning/Visioning	6.26 (1.70)	6.41 (1.86)
Consideration/Team Skills	5.92 (1.88)	6.49 (1.97)
Problem Solving/Innovation	6.36 (1.99)	6.38 (2.13)
Influencing Others	3.26 (1.40)	3.64 (1.34)
Communication	4.04 (1.61)	4.33 (1.73)
Drive/Results-Oriented	8.41 (2.76)	8.20 (2.85)
Tolerance for Stress	3.69 (1.39)	3.53 (1.50)
Integrity/Ethics	4.19 (1.47)	3.56 (1.45)

Table 2

Correlations Between SALSA Form A and SALSA Form B and Coefficient Alpha

Dimension	Number of Items	Number of Duplicate Items	Correlations		Coefficient Alpha		
			With Duplicate Items r_{AB}	Without Duplicate Items r_{AB}	Overall	Form A	Form B
Overall	130	n/a	.906*	.857*	.92	.85	.87
Organizing/Planning/Visioning	18	2	.523*	.426*	.57	.32	.45
Consideration/Team Skills	21	1	.589*	.479*	.65	.44	.44
Problem Solving/Innovation	19	3	.636*	.421*	.65	.49	.54
Influencing Others	11	1	.475*	.351*	.53	.42	.31
Communication	12	2	.686*	.483*	.68	.50	.58
Drive/Results-Oriented	25	3	.804*	.699*	.78	.64	.65
Tolerance for Stress	11	1	.485*	.310*	.54	.35	.45
Integrity/Ethics	13	1	.499*	.370*	.54	.38	.36

Note. * $p < .01$

Study 2

Method

Participants. Data in this study were previously collected from students enrolled in the CLE Leadership Certificate Program. A total of 53 students completed the SALSA and the CLE Assessment Center.

Procedure. Assessment center scores as well as SALSA scores were reported by CLE. Assessment center scores were either collected at the beginning of a semester as an entrance assessment or at the end of the semester as an exit assessment. The activities during the assessment centers included an oral presentation, a leaderless group discussion, and two team related activities. After completing the assessment center, students then completed the online version of SALSA. Six dimensions of leadership are evaluated during the assessment center that match up with six of the eight dimensions evaluated by SALSA.

Results. Hypothesis 2 predicted there would be a significant relationship between scores from the CLE assessment center and scores on SALSA. Correlation coefficients were calculated between the six dimensions of the CLE Assessment Center and the six corresponding dimensions of SALSA (i.e., all except Tolerance for Stress and Integrity/Ethics), as well as between composite scores. The composite scores on the assessment center and SALSA produced a significant correlation coefficient ($r = .513, p < .01$). The correlations between dimensions ranged from $r = .310$ to $r = .392$ and were significant but weak: Visioning and Planning ($r = .336, p < .05$), Problem Solving ($r = .355, p < .01$), Influencing Others ($r = .392, p < .01$), Communication ($r = .310, p < .05$), Team Skills ($r = .358, p < .01$), and Results Orientation ($r = .338, p < .05$). Assessment

center dimensions showed strong intercorrelations, as did SALSA dimension scores. The correlation matrix for all dimensions for both the assessment center and SALSA are reported in Appendix D.

Discussion

Alternate Forms Reliability

The alternate forms used in Study 1 were developed in Grant (2009). In the initial study, the items were grouped by difficulty within dimensions. Items were matched and were randomly assigned to either SALSA Form A or SALSA Form B. If there was an odd number of items within a dimension (i.e., an item did not have a pairing), the item was placed on both forms. This process resulted in 72 items in both SALSA Form A and SALSA Form B. The duplicate items in both forms were taken into account during the calculations in the present study, as the duplicate items were removed and the alternate forms reliability remained significant. Alternate forms bivariate correlations were strong both overall and across dimensions, with dimension correlations ranging from $r = .475$ to $r = .804$; the overall correlation was $r = .906$. Although the overall correlation ($r = .857$), as well as the dimension correlations ($r = .310$ to $r = .699$) were slightly lowered when duplicate items were removed, these relationships still suggest that the alternate forms of SALSA© could be used as a pre and post-test for leadership assessment. The weak correlations between alternate forms of some of the dimensions could be due to the small number of total items on these dimensions. The internal consistency reliabilities calculated for each form of the test suggest that there is some variability across dimensions, with reliabilities ranging from $\alpha = .310$ to $\alpha = .650$. However, overall SALSA reliability ($\alpha = .922$) as well as overall Form A ($\alpha = .854$) and overall Form B ($\alpha = .870$) suggest that internal consistency for the test as a whole and for each form is acceptable. Considering these findings, it may be beneficial to report the overall scores for leadership assessments rather than dimensions scores.

Convergent Validity

The second study evaluated the convergent validity of the SALSA with the CLE assessment center. By comparing scores on the two measures, an assessment of whether they are measuring the same construct of leadership can be made. Scores were collected from students completing the Center for Leadership Excellence Certificate Program. Intercorrelations among dimensions were significant, yet weak, ranging from $r = .310$ to $r = .392$. Intercorrelations among assessment center dimensions were strong overall. It can be inferred that the CLE assessment center is measuring a single underlying construct. The SALSA© dimension intercorrelations are significant, but more moderate. The assessment center composite score had moderate relationships with the SALSA dimensions. The same was noted for the SALSA composite score in relation to the assessment center dimensions. The assessment center and SALSA composite scores showed a moderate relationship as well ($r = .513$). This suggests that the two assessments are measuring the same construct to an extent. The low correlations between dimensions could mean the assessment center and SALSA are measuring the construct in different ways. Furthermore, these low correlations could mean that the items within dimensions are measuring leadership, but not the specific dimension to which they are assigned.

Limitations

There are several limitations to the current study. The relatively low internal consistencies for some SALSA dimensions suggest these dimensions may not be measuring unique dimensions of leadership. This lack of reliability puts a limit on the convergent validity examined in Study 2. Likewise, the high intercorrelations of the Assessment Center dimension suggest that the assessment center may be unidimensional.

The lack of convergent validity suggests that both measures of leadership may suffer from method bias.

In Study 2, the convergent validity of the SALSA and the CLE assessment center was evaluated. Only 53 individuals had completed both measures. The small sample size in this study makes the results of the study less reliable.

Future Research

Given the small sample size in the convergent validity study, it would be beneficial to repeat the study with a larger sample size (i.e., once additional students have completed both measures). Using the alternate forms of SALSA to evaluate individual strengths and weaknesses on the eight dimensions of leadership may require additional research. The low reliabilities of some dimensions suggest that these dimensions may need modification such as refining, adding, or deleting items. Using SMEs to generate additional items for the SALSA may increase both its validity and reliability. Future research might include a factor analysis, to determine the underlying factor structure of SALSA and further refine items within dimensions.

Conclusion

In the alternate forms reliability study it was hypothesized that alternate forms of the SALSA would be positively correlated. The results indicated a high correlation coefficient between the two forms of the test and moderate to strong relationships between the forms across dimensions. It is concluded that SALSA can be used as a pre-test and post-test measure of leadership. The convergent validity of the SALSA and CLE assessment center show that the SALSA would be a valid replacement for assessing leadership. This is beneficial to organizations such as the Center for Leadership

Excellence as the SALSA would be more cost effective than the assessment center.

SALSA is more accessible as it is available online, which will be beneficial to those who live far from the assessment center or have limited mobility. The dimensions of the SALSA should be evaluated further to determine if item refinement would increase dimension reliability.

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.
- Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection & 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 troubles begin. *International Journal of Selection & Assessment*, 14, 223-235.
- Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology*, 62, 229-258.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, 51, 327-358.
- 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. (2009). The validation of a situational judgment test to measure leadership behavior. Unpublished Thesis, Western Kentucky University. Bowling Green, Ky.
- Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, 37, 426-441.

- McDaniel M. A., Hartman, N. S., Whetzel, D. L., & Grubb III, W. L. (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60*, 63-91.
- Motowidlo, S., Crook, A., Kell, H., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business & Psychology, 24*, 281-288.
- Motowidlo, S. J., Dunnette, M.D., & Carter, G.W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640-647.
- Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91*, 749-761.
- Murphy, K. R., & Davidshofer, C. O. (1988). Psychological testing: Principles and applications. Upper Saddle River, New Jersey: Prentice-Hall, Inc.
- Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment. *International Journal of Selection & 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 & Assessment, 15*, 19-29.
- Shoenfelt, E. L. (2009). The development of the situational assessment of leadership: Student assessment (SALSA©). Unpublished Manuscript, Western Kentucky University, Bowling Green, Ky.

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

Appendix A

WKU HSRB Approval Form



A LEADING AMERICAN UNIVERSITY WITH INTERNATIONAL REACH
HUMAN SUBJECTS REVIEW BOARD

In future correspondence, please refer to HS10-143, January 7, 2010

Patricia Slack
c/o Dr. Shoenfelt
Psychology
WKU

Patricia Slack:

Your research project, *A Situational Assessment of Student Leadership (SALSA®): An Evaluation of Alternate Forms Reliability and Convergent Validity*, 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 not 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 Exempt from Full Board Review Level.

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 approval forms to assure participants of compliance with The Office of Human Research Protections regulations.

Sincerely,

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



cc: HS file number Slack HS10-143

The Spirit Makes the Master

Appendix B

Description of SALSA© Dimensions

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 C

Item Difficulty and Assignment to Forms

Item	Rating	Difficulty	P-	Difficulty	Final	Form
	Difference	1	value	2	Difficulty	
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 D

Correlation Coefficients for Convergent Validity

	ACPSI	ACVP	ACIO	ACVNV	ACTS	ACRO	PSI	OPV	IO	Comm	CTS	DRO	TolS	IE	AC	SJT
ACPSI	1.00															
ACVP	.419**	1.00														
ACIO	.951**	.460**	1.00													
ACVNV	.931**	.445**	.933**	1.00												
ACTS	.939**	.294*	.940**	.891**	1.00											
ACRO	.921**	.257	.926**	.872**	.962**	1.00										
PSI	.355**	.241	.355**	.394**	.297*	.283*	1.00									
OPV	.388**	.336*	.391**	.402**	.356**	.298*	.560**	1.00								
IO	.441**	.308*	.392**	.460**	.319**	.315*	.499**	.534**	1.00							
Comm	.257	.435**	.264	.310*	.171	.144	.511**	.563**	.443**	1.00						
CTS	.372**	.284*	.433**	.432**	.358**	.333*	.697**	.536**	.465**	.593**	1.00					
DRO	.371**	.410**	.384**	.417**	.347*	.338*	.515**	.501**	.490**	.387**	.516**	1.00				
TolS	.505**	.608**	.522**	.525**	.394**	.372**	.567**	.531**	.530**	.479**	.544**	.593**	1.00			
IE	.353*	.241	.412**	.481**	.362**	.332*	.489**	.511**	.584**	.508**	.608**	.456**	.471**	1.00		
AC	.979**	.449**	.984**	.959**	.965**	.952**	.552**	.394**	.412**	.264	.409**	.402**	.511**	.408*	1.00	
SJT	.492**	.461**	.512**	.553**	.426**	.397**	.812**	.766**	.723**	.710**	.819**	.768**	.762**	.745**	.513**	1.00

Note. **. 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-Oriented; 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-Oriented; 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.

