5-2011

An Evaluation of Alternate Forms of Reliability of the Situational Assessment of Leadership: Student Assessment (SALSA©)

Ashley N. Wade
Western Kentucky University, ashley.wade156@topper.wku.edu

AN EVALUATION OF ALTERNATE FORMS RELIABILITY OF THE SITUATIONAL ASSESSMENT OF LEADERSHIP: STUDENT ASSESSMENT (SALSA©).

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
Ashley N. Wade

May 2011

AN EVALUATION OF ALTERNATE FORMS RELIABILITY OF THE SITUATIONAL ASSESSMENT OF LEADERSHIP: STUDENT ASSESSMENT (SALSA©).

Date Recommended May 4, 2011

Dr. Elizabeth Shoenfelt, Director of Thesis

Dr. Reagan Brown

Dr. Andrew Mienaltowski

Richard G. Bowker May 16, 2011

Dean, Graduate Studies and Research Date

TABLE OF CONTENTS

List of Tables

AN EVALUATION ALTERNATE FORMS RELIABILITY
OF THE SITUATIONAL ASSESSMENT OF LEADERSHIP:
STUDENT ASSESSMENT (SALSA©).

Name: Ashley N. Wade Date: May 2011 54 Pages

Directed by: Drs. Elizabeth Shoenfelt, Reagan Brown, and Andrew Mienaltowski

Department of Psychology Western Kentucky University

The primary goal of the current study was to re-evaluate, revise, and abbreviate alternate forms of the Situational Assessment of Leadership: Student Assessment (SALSA©) developed by Grant (2009). Archival response sets collected from individuals with extensive experience in leadership who were administered either the full-length SALSA© or Form A or B in previous studies. A total of 80 individual response sets comprised the final sample. Items were categorized by p-value and Subject Matter Expert ratings gathered from the previous study. Items were then selected based on a combination of difficulty and item-total correlation (ITC) values. Selected items were paired based on ITC, and randomly assigned to either Form A or Form B. The newly created forms yielded acceptable alpha coefficients, indicating satisfactory reliability. The coefficient of equivalence between the two forms was high, indicating that the two tests are acceptable alternate forms of the SALSA©.

AN EVALUATION ALTERNATE FORMS RELIABILITY OF THE SITUATIONAL ASSESSMENT OF LEADERSHIP: STUDENT ASSESSMENT (SALSA©).

The Situational Assessment of Leadership: Student Assessment, or SALSA© is a situational judgment test (SJT) developed by Shoenfelt (2009). The instrument assesses the leadership dimensions identified by Arthur, Day, McNelly, and Edens (2003) as the dimensions most commonly evaluated by leadership assessment centers. Western Kentucky University's Center for Leadership Excellence (CLE) expressed a need for a situational judgment test to be used in lieu of their assessment center to assess the performance of students enrolled in the leadership certificate program. To facilitate the use of SALSA© as both a pre-and post assessment, Grant (2009) developed alternate forms of the assessment in a previous study. The current study re-evaluates, revises, and abbreviates these forms using response sets collected from individuals with extensive experience in leadership.

Background

Western Kentucky University's Center for Leadership Excellence (CLE) offers training and assessment of leadership qualities/skills, as well as leadership certificate programs for undergraduate and graduate students. The certificate programs offer instruction in ethics, social responsibility, and core leadership theory, which promotes the understanding of current leadership practices and applications in students. The CLE gives certificate students the option of participating in a leadership assessment center before beginning the program, and upon program completion. This practice allows the CLE to assess the growth of students' leadership abilities during participation in the leadership certificate program. Data gathered during the assessment centers are also used as a

diagnostic tool to ensure that all facets of leadership assessed in the assessment center are being adequately taught by program instructors.

The assessment center program utilized by the CLE was designed in 2006. It examines seven dimensions of leadership identified by a meta-analysis conducted by Arthur et al. (2003). These dimensions include: problem solving and innovation, influencing others, verbal/non-verbal communication, team skills, visioning and planning, tolerance for stress, and results orientation. The assessment center also measures knowledge of leadership theories, written communication, and self analysis and improvement.

The feedback and experience the leadership assessment center provides is a valuable tool for leadership development. With the continued growth and success of the CLE in recent years, student interest in the leadership certificate programs has steadily increased. Although the university fully supports student interest in the CLE and its programs, it does not possess the necessary resources to allow all leadership program students to participate in the assessment center. Additionally, as WKU's distance learning courses and programs expand, off-campus students interested in participating in the leadership program face time and distance constraints barring participation in the assessment center.

Waldman and Korbar (2004) noted that student evaluation via assessment center methods predicts future success. Analyses conducted by the authors revealed that assessment center scores consistently and strongly predicted several aspects of early career success while GPA predicted only salary in a sample of business students. These findings demonstrate the utility of student assessment centers. Assessment centers have

been the target of considerable investigation in recent years. Of special interest are the problems common among assessment centers, including rater/assessor bias, differential impact of scoring methods, psychological and physical fidelity issues, and exercise effects. Such issues have caused a number of scholars and researchers to advise against the use of assessment centers in favor of other, less problematic forms of evaluation (Moses, 2008; Motowidlo, Dunnette, & Carter, 1990; Sackett, 1987). Construct issues, such as those mentioned by Lowry (1995) also have bolstered the case against the use of assessment centers as selection procedures. In order to address growing student interest in leadership assessments, the CLE expressed interest in the development of a situational judgment test (SJT) to assess leadership in students.

Shoenfelt (2009) developed an SJT, termed the Situational Assessment of Leadership: Student Assessment, or SALSA©. The assessment, a paper and pencil/computer-based format, solves many of the issues associated with the CLE assessment center. The SALSA© is low-cost, and easy to administer to students both on and off-campus. It may be used in place of the CLE assessment center to provide feedback to leadership students and the CLE. The SALSA© was developed to measure many of the same dimensions of leadership examined by the original assessment center, which were based on dimensions identified by Arthur et al. (2003). Table 1 illustrates the dimensions examined by each method of assessment.

Table 1.

SALSA© and CLE Assessment Center Dimensions.

SALSA©	CLE Assessment Center
Problem solving & innovation	
Influencing others	
Verbal/non-verbal communication	
Team skills	
Visioning & planning & results- orientation	
Tolerance for stress	Knowledge of leadership theories
Integrity/ethics	Written communication
	Self-analysis & improvement

As shown in Table 1, the SALSA© measures five out of the eight dimensions originally examined by the CLE assessment center, and two additional dimensions identified by Arthur et al. (2003). The CLE assessment center dimension of "Knowledge of Leadership Theories" is in the form of a paper-and-pencil exam, and was not included in the SALSA©. The assessment center dimension of "written communication" was not included in the SALSA© because the SJT is in multiple-choice format, and therefore could not accurately measure the dimension. The assessment center dimension of "Self-Analysis and Improvement" was not included in the SALSA© because it serves as a way for leadership students to compare their thoughts on their performance in the assessment center simulations to those of the assessment center raters, and therefore was also not amenable to the SJT format.

SALSA© examines two dimensions not assessed in the CLE's assessment center: Tolerance for Stress and Integrity/Ethics. A meta-analysis by Arthur et al. (2003) outlined

the most common dimensions examined by leadership assessment centers. Out of these dimensions, tolerance for Stress was the only dimension that was not assessed by the CLE assessment center. To summarize, the SALSA© measures the seven most common assessment center dimensions identified by Arthur et al., and an eighth dimension of Integrity/Ethics. A detailed explanation of each of the 8 dimensions of the SALSA© may be found in Appendix A.

Initially, SALSA© was a single form assessment composed of 130 items. As such, its use as a pre-and post-assessment is debatable; scores may be inflated due to practice effects or question familiarity. In order for the SALSA© to be used as both a pre- and post-assessment, alternate forms were created by Grant (2009). Alternate forms of an instrument consist of different items, with approximately equivalent internal consistency (Cronbach's alpha), difficulty level, and scores. Alternate forms of the SALSA© enabled users to more accurately and quickly measure relevant dimensions of leadership without the influence of practice effects or other issues.

To further increase the utility of the instrument for use by the CLE, Grant (2009) constructed alternate forms of the SALSA©. The study produced two alternate forms of the SALSA©: SALSA© Forms A and B, each consisting of 72 items. Although the two forms yielded similar scores and possessed similar psychometric properties, the study had a few notable limitations. A number of these students spoke English as a second language. A somewhat small sample of leadership certificate program students (N=40) were administered the full-length assessment; their scores were used to construct the alternate forms. These issues may have influenced the categorization of items based on difficulty, as well as the actual construction of alternate forms to the extent that the

sample was not representative of those who will take the SALSA© in the future, the alternate forms developed based on this sample may prove unreliable.

The current study remedies some of these issues by using SALSA© scores collected from a sample of individuals with extensive leadership training to construct alternate forms of the assessment. This paper will first review relevant literature on assessment centers. The properties of SJTs will then be examined, and the construction and use of SALSA© will be discussed. Previous studies establishing alternate forms of SALSA© will then be described, along with their findings and limitations. Finally, the current study will be introduced and explained.

Assessment Centers

Assessment Centers have long been used by organizations to assess various desirable knowledge, skills, abilities, and behaviors. In a typical assessment center, multiple candidates are assessed on multiple constructs/dimensions by multiple assessors. Assessment centers may be comprised of a number of activities including interviews, performance or simulation exercises, paper and pencil exercises, and other activities. Individual activities may be tailored to meet the requirements of a specific organization (Guion, 1998).

A key feature of assessment centers is the use of assessors. Each candidate participating in the assessment center is rated by two or more assessors. Guion (1998) recommended at least a 2:1 ratio of assessors to participants. Individuals serving as assessors typically have some experience with or knowledge of subject matter relevant to assessment center exercises, and are commonly referred to as subject matter experts (SME's). Aside from requisite background knowledge, assessors must be trained to

attend to desired behaviors in each exercise, and rate participants accordingly in order to increase levels of rater agreement, and reduce bias in ratings. This practice is known as rater agreement training and/or calibration. Despite its obvious utility, this process adds considerable time and in certain cases, costs, to the use of assessment centers. Even with calibration/rater agreement training, biases and low rater agreement may still occur.

Benefits of using Assessment Centers.

Assessment centers have gained popularity in recent years for both developmental (e.g., training) and administrative (e.g., selection for hire or promotion) purposes (Arthur et al., 2003). Assessment centers may evaluate participants on a number of strong predictors of job performance, including cognitive ability, personality variables, job knowledge, etc. Assessment center scores based on these robust predictors have been shown to reliably predict candidate job performance (Arthur et al., 2003).To assess these predictors, assessment centers may include several different types of exercises including work samples, interviews, and other activities. The inclusion of multiple activities measuring desired characteristics provide more accurate measurement of relevant constructs (Arthur et al., 2003).

Issues with Assessment Center use.

Assessment centers face several technical and theoretical constraints. Assessment centers typically exhibit low to moderate levels of criterion validity. A meta-analysis conducted by Arthur et al. (2003) yielded an average value of r = .37, which was lower for corrected means. Evidence supporting criterion- and content-validity is strong (Gaughler, Rosenthal, Thornton, & Bentson, 1987; Sackett, 1987). Conversely, evidence of construct-validity (e.g., convergent and discriminant validity) has been consistently

weak. In assessment centers, participants are evaluated on the same set of constructs multiple times (by multiple raters). Strong convergent and discriminant validity (e.g., construct validity) would be expected among ratings in similar situations. Researchers hypothesize that this problem is associated with the tendency of assessors to focus on the specific factors of an exercise, instead of the global factors influencing overall participant performance.

Furthermore, assessment center activities typically do not simulate actual leadership situations that an individual may encounter in real life (Howard, 2008). Therefore, performance of participants may be affected by the type of exercise, and may not be a strong predictor of leadership behavior in a real-world setting. Assessment center exercises tend to have low face validity, which may impact individual performance (Moses, 2008). If the participant does not see the relevance of an exercise to his/her daily activities, he or she may not put forth maximum effort during participation.

Several other issues associated with assessment centers make them a disadvantageous method of evaluation. Assessment centers are often expensive to design and implement; hiring qualified personnel to develop and conduct the assessment center model requires considerable front-end investment. Each time the assessment center is conducted fees for facility rental, staff and assessor compensation, and materials contribute to the growing costs associated with assessment center use (Grant, 2009). Aside from costs, several difficulties are associated with administration of the actual assessment. Assessment centers are usually lengthy to administer, and actual administration may be difficult or complex, thus requiring additional staff training and/or qualifications.

Finally, assessment centers provide little convenience for attendees. Assessment centers must be held at a facility, with all raters and participants present. This requires considerable coordination between the two groups. The exercises, participants, and raters must all be coordinated to ensure that the center runs smoothly. Raters must undergo rater agreement/calibration training prior to the actual assessment in order to ensure that they are attending to the same behaviors and skills, thus requiring additional time and instruction. Additionally, only a small number of participants may take part in any individual assessment center. Costs to design, prepare, and conduct the center may range from a few hundred dollars to several thousand dollars per candidate assessed. Thus, organizations requiring the assessment of large numbers of individuals assume considerable time and resource costs if assessment centers are used as an evaluation method.

Despite their problems, assessment centers yield beneficial data and feedback to both organizations and participants. SJTs provide many of the same benefits, but circumvent some of the critical issues associated with assessment centers. The design, implementation, and properties of SJTs will now be discussed.

Situational Judgment Tests

SJTs are typically paper-and-pencil or computer-based assessments that in which participants are presented with hypothetical situations and asked them to select the best response. The format of the instructions and items determines whether the test examines average or optimal performance. Items asking participants to indicate what they "would do" (or have done) in a given situation are examining actual performance and are referred to as behavioral tendency instructions (McDaniel, Hartman, Whetzel, & Grubb, 2007).

Tests constructed in a "would do" format have been shown to correlate with measures of personality. SJTs asking participants to indicate what he/she "should do" in a given situation, or SJTs asking participants to select the "best" response, are examining optimal performance, and are referred to as knowledge instructions (Ployhart & Ehrhart, 2003). SJTs measuring optimal performance yield results and predictions comparable to those provided by other assessments of optimal performance, including work samples, cognitive ability tests, and tests of job knowledge.

Development of SJTs.

Typical SJT development follows the steps outlined by Motowidlo, Dunnette et al. (1990). This technique relies heavily on the construction and use of Critical Incidents. In order to generate critical incidents for later use as items, subject-matter experts (SME's) are asked to write short scenarios depicting events associated with the target job. To generate response options for each of the scenarios, a different group of SME's is asked to read each scenario and write descriptions of how they would respond. Following this, a group of more experienced SME's possessing extensive target domain knowledge rates the acceptability of each response option for a given item. These ratings are used to calibrate determine the best response for each item.

Based on the calibration and ratings, a scoring key is developed. Responses may be scored using one of six methods described by Bergman, Drasgow, Donovan, Henning, and Juraska (2006), which are empirical, theoretical, hybridized, expert-based, factorial, or sub-grouping methods. Of these methods, empirical and expert-based scoring are most frequently used for SJT applications (Lievens, Peeters, & Schollaert, 2008). Following the selection of a scoring method, a point assignment scheme is established. Respondents

may be awarded 1 point for a correct response, or assigned a -1 point value for choosing the least effective response. Special scoring keys must be developed for SJTs requiring test takers to rank the effectiveness of responses (Weekley & Ployhart, 2005). Several 'point' values (e.g., 0, 1, -1) may be assigned to each response option, depending on the applicability of each option to the scenario in question (Bergman et al., 2006).

Although the process described by Motowidlo et al. (1990) is used to develop most current SJTs, the process need not be rigidly followed. A number of variations on this basic method may be utilized, each of which will yield similar SJT products. During construction of SALSA ©, the same group of SME's was used to generate critical incidents and preliminary response options, even though Motowidlo et al. (1990) suggested that different groups of SME's be used for these tasks. Despite this change in method, the procedure resulted in the creation of a valid SJT (Shoenfelt, 2009).

Correlates and Psychometrics.

Scores on SJTs have been found to correlate with a number of strong predictors of job performance. Weekley and Jones (1999) reported that SJTs correlated strongly with cognitive ability (weighted average r = .45) and performance (weighted average r = .20). A later study by Weekley and Ployhart (2005) identified several additional correlates including job tenure (r = .13), training experience (r = .12), and cognitive ability (r = .36). The same study found that several personality dimensions relevant to job performance correlated significantly with SJTs: conscientiousness (r = .13), emotional stability/neuroticism (r = .17), and extraversion (r = .14).

A later meta-analysis of SJTs by McDaniel et al. (2007), reported that most studies of SJTs have reported reliability in terms of alpha coefficients. The authors noted

that this reliance on alpha coefficients may have yielded incorrect estimates of SJT reliability due to the fact that SJTs rarely assess singular constructs. Due to the multidimensionality of SJTs, test-retest reliability coefficients offer a better estimate of instrument reliability. A meta-analysis conducted by Ployhart and Ehrhart (2003) examining psychometric properties of various SJT formats and scoring practices reported that "would do" versions tended to show higher test-retest reliabilities (averaged $r = .83$) than "should do" versions (averaged $r = .36$). The authors also noted that the highest test-retest reliability occurred for SJTs asking respondents to indicate how likely they would be to perform each response ($r = .92$).

Advantages of using SJTs.

SJTs offer a number of technical and statistical advantages over assessment centers. SJTs generally have sound psychometric properties. Arthur et al. reported that assessment centers have an averaged criterion-related validity coefficient of $r = .37$; a meta-analysis by Christian et al. (2010) found that SJTs have validity coefficients ranging from $r = .58$ to $r = .67$, depending on the type of test and construct assessed. Furthermore, SJTs broaden the criterion domain by allowing a more varied sample of behaviors and responses to be examined (Lievens et al., 2008). SJTs typically cause less adverse impact than more cognitively-oriented methods of assessment. SJTs possess higher face validity than most cognitive measures due to their use of critical incidents depicting actual job situations instead of contrived activities or exercises (Lievens, Buyse & Sackett, 2005). As previously mentioned, SJTs are strong predictors of job performance (Motowidlo et al., 1990).

Due to the behavioral consistency principle, or the notion that past behavior predicts future behavior (Lievens et al., 2008), SJT responses will likely correspond to a respondent's future behaviors. The authors noted that SJTs evaluate the intentions and goals of respondents just as well as other established predictors of job performance. Additionally, meta-analyses by McDaniel et al. (2007) and Lievens, et al. (2008) indicated that SJTs provide incremental validity (.03 to .08) over cognitive and personality measures.

Aside from their robust psychometric characteristics, SJTs also offer several practical advantages over assessment centers. SJTs are typically administered in paper-and-pencil or computer-based formats, offering maximum convenience for test takers and organizations. This format also eliminates the need for raters or rater training, thus simplifying the evaluation and scoring processes for organizations. SJTs often require less time to administer and, because of their format, may be administered to a large number of applicants in a small amount of time. SJTs also cost less to design and administer than assessment centers over the life of the instrument. Although the design of SJTs requires considerable front-end investments of time and resources, administering the actual assessments is inexpensive.

SJT items are derived from critical incidents, and are hypothetical situations. Accordingly, SJT items may be tailored to measure specific constructs/dimensions associated with their intended use. Furthermore, subjectivity is virtually eliminated from the scoring process due to the creation of a scoring key and selection of a scoring methods following test construction, (McDaniel & Nguyen, 2001). Although SJTs offer a

number of advantages over assessment centers, there are several issues associated with their use.

Issues associated with SJTs.

Although internal consistency is generally high among SJT items (.43 to .94), it may be affected by many factors, including the length of the assessment, the response instructions used, or the multidimensionality of individual items (Lievens et al., 2008). Longer SJTs and those asking participants to rate the effectiveness of response options tend to have the highest internal consistency coefficients. Additionally, the use of factor analysis techniques to assess the internal consistency of SJTs may underestimate internal consistency coefficients, likely because of the multidimensional nature of SJTs. Thus, it may be best to use test-retest reliability to assess consistency of SJT assessments. Ployhart and Erhart (2003) noted that test-retest reliability for various forms of SJTs is adequate, ranging from $r = .20$ for SJTs asking participants to rate how effective each response option is, to $r = .92$ for SJTs asking participant to rate how likely they would be to do each response option.

SALSA©

Shoenfelt (2009) developed the Situational Assessment of Leadership: Student Assessment, or SALSA©, in response to the need expressed by the CLE for an SJT to replace the currently used assessment center model. The assessment evaluates seven common leadership assessment center dimensions reported by Arthur et al. (2003): Organizing/Visioning/ Planning; Consideration/Team Skills; Problem Solving/ Innovation; Influencing Others; Communication; Drive/Results Orientation; and Tolerance for Stress. Additionally, another dimension, Integrity/Ethics was included. The

instrument consists of a total of 130 items across eight dimensions, with 10-20 items per dimension.

SALSA© presents test takers with a number of hypothetical scenarios, and four response options for each scenario. Participants are instructed to select the response option that depicts the behavior they believe a leader should perform in order to obtain the most effective leadership response in each scenario. SALSA© assumes a "should do" format, which has been shown predict cognitive ability (McDaniel et al., 2007; Nguyen, Biderman, & McDaniel, 2005).

Completion of all 130 SALSA© items takes approximately one hour. One point is awarded for each correct response and respondents are not penalized for incorrect responses. Dimension scores are obtained by summing the correct responses for items in a given dimension. An overall score is obtained by summing the total number of correct responses across all dimensions. (Shoenfelt, 2009).

Test construction.

Individual items for the SALSA© were created using the critical incident technique advocated by Flanagan (1954). SME's were recruited from several sources within the university, including students enrolled in the Industrial/Organizational (I/O) Psychology Masters program, students in WKU's Honors Leadership program, members of the Dynamic Leadership Institute, and ROTC cadets. During critical incident workshops, SME's were asked to generate critical incidents depicting opportunities for leadership behaviors relevant to one of the eight identified dimensions. They were also asked to produce three to four responses for each scenario (Grant, 2009). The critical incidents and responses were edited and refined by Shoenfelt (2009).

Following the creation of items for the instrument, a scoring key was developed using the process described by Motowidlo et al. (1990). Seven WKU faculty members with substantial experience teaching leadership courses at the undergraduate and graduate level served as SME's. These individuals were provided with all test items and responses and asked to rate the effectiveness of each response option for a given test item. Only items with one correct response alternative, as rated by SME's, were included in the final version of SALSA©.

Psychometric properties.

The SALSA© has exhibited several strong psychometric properties in previous studies. Grant (2009) reported high internal consistency ($\alpha = .91$) for the full-length instrument. Convergent validity coefficients between scores in the CLE's assessment center and SALSA© scores were found to be low but significant. Validity coefficients for individual dimensions matched between the CLE assessment center and SALSA© ranged from $r = .28$ to $r = .44$, indicating low to moderate, but significant correlations (Grant). Composite assessment scores were significantly correlated with the composite SALSA© scores, ($r = .55$, $p < .01$). An analysis of item difficulty indicated a nearly even number of items previously categorized by SME's as easy, moderate, and difficult (Grant).

Alternate Forms Reliability

In order to facilitate the use of an instrument as a pre- and post-test, test items may be divided to create equivalent forms. Reliability of these new forms is assessed using the method of estimating alternate forms reliability advocated by Murphy and Davidshofer (1988). Alternate test forms are defined as two forms of the same instrument

with equivalent content, response process, and psychometric properties, but possessing different item sets.

In order to estimate alternate forms reliability, both forms of the instrument must be administered to a single group of participants spaced by a designated inter-test interval. Scores on the two test versions are then correlated to obtain the alternate forms reliability estimate. Stronger correlations between the two versions indicate high reliability. Alternate forms reliability approaches offer several advantages over test-retest approaches (e.g., administering the same test twice, spaced by a considerable inter-test interval) for pre-test/post-test applications. Given that the two forms of the instrument contain different but statistically equivalent item sets, practice and reactivity effects often observed with test-retest methods are virtually eliminated. Due to this, the lengthy inter-test intervals required by test-retest methods are not necessary (Murphy & Davidshofer, 1988)

Issues with Alternate Forms Reliability.

Although alternate forms reliability approaches offer several strong advantages over test-retest methods, there are several drawbacks associated with the approach. Developing several forms of a test requires considerable time and monetary resources, and alternate forms procedures may result in costs equal to or greater than those encountered with test-retest methods. This issue may be overcome by administering the entire, undivided instrument to participants, and then dividing scores into their respective forms. This method, as used in previous studies (Grant, 2009), eliminates the need for and costs associated with multiple test administrations.

Furthermore, arbitrarily splitting an instrument in half is likely to result in forms that are not equivalent on one or more psychometric or statistical properties. This issue may be averted through the use of grouping by pre-determined difficulty ratings and random assignment to forms. Scores on the resultant test forms should then be correlated to determine equivalency (Grant, 2009).

Previous Studies

Two previous studies created alternate forms of the SALSA©. Grant (2009) created alternate forms (SALSA© forms A and B) consisting of 72 items each. The forms were created using response data from students enrolled in the CLE's leadership certificate programs. The response sample consisted of 40 students, a number of whom did not speak English as a native language (ESL). Furthermore, the students received limited instruction in leadership skills and behaviors, compared to sample used in the current study. These sample attributes may have influenced the calculation of p-values and the resultant construction of alternate forms due to the main effects found for native vs. non-native English speaking respondents and for gender among non-native English speaking respondents.

Furthermore, the forms generated by Grant's (2009) study had unequal numbers of items across dimensions. The current study revisited the findings of Grant's study, and created alternate forms of the SALSA© with a total of 5 items per dimension, and 40 items per form. This process allowed items with the lowest item total correlations (ITC's) to be identified and eliminated, thus increasing reliability and providing a stronger overall instrument. The reduced length allows for further streamlining of the test administration process.

Slack (2010) also created alternate forms of SALSA© using a larger sample (N = 156) of undergraduate students enrolled in the CLE's leadership certificate program. Much as in Grant's (2009) study, the procedure yielded two alternate forms of SALSA©, but encountered similar limitations. Although the sample size was sufficient, it still consisted of a number of CLE leadership students who were non-native English speakers. Again, this quality may have influenced the results of the study, given the main effects for ESL respondents established in earlier an earlier study (Grant).

The Current Study

The current study re-evaluated and revised the procedures used by the previous researchers to create new alternate forms of the SALSA© instrument using a data set collected from individuals with considerable leadership training. Additionally, the study created abbreviated alternate forms of SALSA© with a total of 5 items per dimension, yielding two alternate forms of SALSA© with 40 total items each. The psychometric properties of the new forms were then assessed.

Hypotheses.

The current study used archival SALSA© scores taken from student athlete leaders, ROTC cadets, MBA, Ed.D., and M.A. in Industrial/Organizational Psychology students enrolled at Western Kentucky University to create alternate forms of SALSA© . A previous study (Grant, 2009) used SALSA© scores taken from undergraduate students enrolled in the leadership certificate program offered by the CLE. Due to the advanced training in leadership skills of student athlete leaders, ROTC cadets, MBA, Ed.D., and I/O Psychology students, as well as the increased experience of the graduate students, we

anticipated that the current archival sample should exhibit higher scores than the sample used in the previous study. Thus, we predicted that:

H1: The current sample will have higher overall and dimension scores on the SALSA© than undergraduate leadership certificate program students used in previous study/sample.

Furthermore, the current study sample is comprised of both undergraduate (student athlete leaders and some ROTC cadets) and graduate students (MBA, Ed.D., some ROTC cadets, and I/O Psychology). Due to the increased training, and experience levels of the graduate students, we also predicted that:

H2: Graduate respondent scores will be significantly higher than undergraduate respondent scores.

Previous re-translation and calibration ensured appropriate categorization of items, as well as measurement for all items included in the instrument. Since the two newly-constructed forms will be composed of equally difficult items, we anticipate that:

H3: SALSA Forms I and II scores will be positively correlated on each dimension, as will overall short-form SALSA scores.

Method

Participants

In a previous study (Tucker, 2011), 16 participants identified as M.B.A, or Ed.D. students at WKU completed either SALSA© form A or B. A second sample of 18 student athlete leaders completed either SALSA form A or B, or the full-length SALSA© assessment (Normansell, 2011). A third sample of 35 ROTC cadets, and 11 Industrial/Organizational (I/O) Masters-level students at Western Kentucky University completed the full-length SALSA© in two previous studies (Shoenfelt, 2009; Stroupe, 2010). All samples were combined to produce a sample consisting of 80 respondents; 36 were female and 44 were male. All participants were considered to be native English speakers.

Procedure

Construction of alternate forms.

Calibration data from the previous alternate forms study (Grant, 2009) were retrieved. During the calibration, SME's were asked to rate the effectiveness of each response option for its respective scenario. Mean ratings were then calculated for each response option, and the differences between mean ratings for the best and next-best response options for each scenario were determined. Items were classified into one of three difficulty categories based on these values. Items with a mean rating difference between the correct responses and the next highest rated response less than or equal to .5 were classified as "difficult" items; items with a mean rating difference between .5 and 1.0 were classified as "moderate" items; and those with differences greater than 1.0 were classified as "easy" items.

A second measure of item difficulty was determined by calculating a p-value (percentage of participants answering an item correctly) for each item. Archival response data collected from student athletes, MBA, Ed.D., and M.A. in I/O psychology students were used to calculate a p-value for each item of the SALSA© . Items were grouped into one of three difficulty categories: items with p-values less than or equal to .5 were categorized as "difficult," items with p-values between .51 and .74 were categorized as "moderate," and items with p-values of greater than .75 were categorized as "easy."

The difficulty categorizations derived from calibration and p-values were compared for each item, and items were categorized as either "difficult," "moderate," or "easy." P-value data and difficulty categorizations were then retrieved from both Grant (2009) and Slack (2010) and compared to those determined in the current study. Following this procedure, corrected item total correlations (ITC's) for each dimension were calculated for each form. For each dimension, the ten items with the highest ITC values were retained. Items in each dimension were paired by difficulty and ITC values. Items in each pair were randomly assigned to either SALSA© Form A or Form B. Corrected item total correlations and alphas were then calculated for each dimension and form.

The alpha level for each combination of items also was considered in determining which items to retain. Item sets that yielded matched (or very close) alphas for the two forms were retained on the final forms. The procedure produced two alternate forms of the SALSA© consisting of 5 items per dimension, for a total of 40 items per form, and possessing approximately equivalent scores and psychometric properties.

Following this process, composite scores, descriptive statistics coefficient alphas, and item total correlations were again calculated for each dimension and overall for each form.

Results

Item Classification.

Analyses of the SME data from calibration in Grant (2009) yielded 53 items classified as easy, 49 items classified as moderate, and 28 items classified as difficult. The results for each of the eight dimensions are presented in Table 2.

Table 2.

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

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

P-values (i.e., percent of respondents getting an item correct) were then calculated using SALSA© response data from student athletes, ROTC cadets, MBA, Ed.D., and M.A. in I/O Psychology students. This step yielded 54 items categorized as easy, 49 categorized as moderate, and 27 items categorized as difficult. The results for each dimension are presented in Table 3.

Table 3

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

Dimension	**Easy**	**Moderate**	**Difficult**	**TOTAL**
Organizing/Planning/Visioning	7	8	3	18
Consideration/Team Skills	9	6	6	21
Problem Solving/Innovation	7	6	6	19
Influencing Others	5	3	3	11
Communication	5	5	2	12
Drive/Results-Orientation	9	3	13	25
Tolerance for Stress	6	4	1	11
Integrity/Ethics	6	4	3	13
TOTAL	54	49	27	130

The results of the SME-based and P-value based methods of categorization were then compared to reach a final difficulty categorization for each item. The results from the SME and P-Value based difficulty analyses shared a moderate, positive correlation (r = .51, p =.00). A total of 67 items (51.5%) were classified into the same category by both methods. Items for which the two methods produced different classifications were ultimately classified using a rational decision process. P-values were typically used to make this decision, but if the difference between means was close to being classified as a different category, that was considered when determining final classification. The final difficulty categorization produced 46 easy items, 58 moderate items, and 26 difficult items. The results of the final categorization are presented in Table 4.

P-values and difficulty categorizations were then collected from Grant (2009) and compared to those obtained in the current study. A total of 94 out of a possible 130 items (72.3%) were classified into the same category in both studies. The results from the P-

Table 4.

Final Difficulty Categorization of Items by Dimension.

Dimension	**Easy**	**Moderate**	**Difficult**	**TOTAL**
Organizing/Planning/Visioning	3	6	1	10
Consideration/Team Skills	6	4	0	10
Problem Solving/Innovation	5	5	0	10
Influencing Others	3	4	3	10
Communication	5	4	1	10
Drive/Results-Orientation	6	4	0	10
Tolerance for Stress	4	5	1	10
Integrity/Ethics	4	4	2	10
TOTAL	36	36	8	80

value based difficulty analyses conducted in the current study and by Grant (2009) were significantly correlated ($r = .73$, $p = .000$). A table containing these values is presented in Appendix C.

Creation of Alternate Forms.

Following final difficulty categorization, item total correlations (ITC's) were calculated for items in each dimension. The ten items with the highest ITC values were retained. This process yielded a total of 36 easy items, 36 moderate items, and 8 difficult items. The results for each dimension are presented in Table 5.

In order to create alternate forms, each item in a given dimension was paired by difficulty categorization and ITC values. One item from each pair was randomly assigned to either SALSA© Form I or SALSA© Form II. This procedure was used for a total of 80 items. After this process was completed, each form contained 40 items. Item total correlations and alphas were then calculated for each dimension and form. The

Table 5

Difficulty of Top Ten Items in each Dimension

Dimension	**Easy**	**Moderate**	**Difficult**	**TOTAL**
Organizing/Planning/Visioning	3	6	1	5
Consideration/Team Skills	6	4	0	5
Problem Solving/Innovation	5	5	0	5
Influencing Others	3	4	3	5
Communication	5	4	1	5
Drive/Results-Orientation	6	4	1	5
Tolerance for Stress	4	5	1	5
Integrity/Ethics	4	4	2	5
TOTAL	36	36	8	80

distribution of items by difficulty for each form is presented in Table 6. A test map showing the item numbers that were retained and assigned to separate SALSA© forms may be found in Appendix C.

Cronbach's alpha was calculated as an estimate of internal consistency for the entire SALSA©, as well as for each new form. Internal consistency for the full-length SALSA© was $\alpha = .86$, SALSA© Form A was $\alpha = .73$, and SALSA© Form b was $\alpha = .75$. Cronbach's alpha was computed for each dimension on each of the assessments. These values are presented in Table 7.

Following form construction, alpha coefficients for some dimensions were found to be unacceptably low (Form A Organizing/Planning/Visioning, Form A Tolerance for Stress, Form B Problem Solving/Innovation, Form B Influencing others, and Form B Communication, and Form B Tolerance for Stress). For each form and dimension in question, the items not used on either form construction were added back to those

Table 6

Item Difficulty across dimensions for SALSA© Forms A and B

	Form A			Form B			
Dimension	**E**	**M**	**D**	**E**	**M**	**D**	**Total**
Organizing/Planning/ Visioning	0	5	0	3	1	1	10
Consideration/Team Skills	3	2	0	3	2	0	10
Problem Solving/Innovation	3	2	0	2	3	0	10
Influencing Others	2	2	1	1	2	2	10
Communication	3	2	0	2	2	1	10
Drive/Results-Orientation	3	2	0	3	2	0	10
Tolerance for Stress	2	3	0	2	2	1	10
Integrity/Ethics	2	2	1	2	2	1	10
TOTAL	18	18	4	18	18	4	40

Note: E = Easy, M = Moderate, D = Difficult

Table 7

Alpha Coefficients for SALSA©, Form A and Form B after Initial Form Construction

Dimension	**Form A**	**Form B**
Overall	.79	.78
Organizing/Planning/Visioning	.35	.45
Consideration/Team Skills	.58	.64
Problem Solving/Innovation	.47	.29
Influencing Others	.43	-.13
Communication	.59	.39
Drive/Results-Orientation	.58	.64
Tolerance for Stress	-.07	.07
Integrity/Ethics	.26	.26

discrepant item set. Alpha coefficients were then re-calculated, and the items found to provide the highest increase alpha upon removal were discarded from the dimension until only five items remained. This process was followed for three dimensions on Form A and three dimensions on Form B, and resulted in the replacement of a total of 10 items across forms. Difficulty level was taken into consideration during this process to ensure that the mean difficulty level within dimensions did not decrease, unless this decrease was slight and also mitigated by larger increases in the resultant alpha coefficients. Alpha coefficients, means and standard deviations for the full-length and revised forms are presented in Table 8.

Table 8

Final Alpha Coefficients, Means, and Standard Deviations for Full-Length SALSA©, Form A, and Form B after Revision

Dimension	**Overall Alpha**	**Overall Mean**	**Overall SD**	**Form A Alpha**	**Form A Mean**	**Form A SD**	**Form B Alpha**	**Form B Mean**	**Form B SD**
Overall	.86	85.33	12.96	.73	29.47	5.42	.75	26.95	5.43
Organizing/Planning/Visioning	.43	11.93	2.48	.35	3.38	1.21	.45	3.61	1.14
Consideration/Team Skills	.64	12.85	3.11	.58	4.07	1.14	.64	3.84	1.38
Problem Solving/Innovation	.40	11.84	2.41	.47	3.56	1.18	.36	2.84	1.16
Influencing Others	.40	6.87	1.87	.35	3.38	1.16	.20	2.97	1.09
Communication	.57	8.21	2.12	.59	3.95	1.20	.39	3.31	1.13
Drive/Results-Orientation	.64	16.97	3.63	.58	4.00	1.20	.64	3.69	1.39
Tolerance for Stress	.08	7.97	1.47	.11	3.78	.97	.22	3.41	1.07
Integrity/Ethics	.22	8.50	1.67	.26	3.35	1.10	.26	3.20	.95

The dimension of Tolerance for Stress was problematic across forms. The sub-scale produced unacceptably low alpha coefficients for the full-length SALSA© as well as for the newly created forms. Grant (2009) reported an alpha of .45 for the full-length

SALSA©, and coefficients of .07 and .46 for the short-form assessments for this dimension. Item replacement occurred on both forms for this dimension in the current study. Item difficulty totals for each finalized form are presented in Table 9.

Table 9.

Final Item Difficulty across dimensions for SALSA© Forms A and B

	FORM A			FORM B			
Dimension	**Easy**	**Mod**	**Diff.**	**Easy**	**Mod**	**Diff.**	**Total**
Organizing/Planning/Visioning	0	5	0	3	1	1	10
Consideration/Team Skills	3	2	0	3	2	0	10
Problem Solving/Innovation	3	2	0	2	3	0	10
Influencing Others	2	2	1	1	2	2	10
Communication	3	2	0	2	2	1	10
Drive/Results-Orientation	3	2	0	3	2	0	10
Tolerance for Stress	2	3	0	2	2	1	10
Integrity/Ethics	2	2	1	2	2	1	10
TOTAL	18	20	2	18	16	6	40

Hypothesis 1 predicted that the current sample of individuals with considerable leadership training would have higher overall scores on the SALSA© than the response set used by Grant (2009). T-tests were conducted to compare the previous and current samples. Significant differences between 2009 test takers and 2011 test takers were found for Tolerance for Stress. All other comparisons were not significant. Values for all t-test comparisons, as well as group means and standard deviations are reported in Table 10.

Hypothesis 2 predicted that graduate student (e.g. MBA, Ed.D. and M.A. I/O students) scores on the SALSA© would be significantly higher that undergraduate (e.g. student athlete leaders and ROTC cadets) scores on the full-length assessment. T-tests

were used to determine differences between the two groups, and seven out of eight dimensions yielded significant differences between groups. Overall graduate scores (M = 97.82, SD = 5.88) were significantly higher than undergraduate scores (M = 80.23, SD = 15.52; t (77) = -3.68, p = .000), thus supporting Hypothesis 2.

Table 10

T-test values comparing 2009 and 2011 SALSA© scores

Dimension	**2009 Means**	**2009 SD's**	**20011 Means**	**2011 SD's**	**t**	**p**
Organizing/Planning/Visioning	12.22	24.15	11.93	2.482	.63	.53
Consideration/Team Skills	13.03	3.11	12.85	3.11	.32	.75
Problem Solving/Innovation	12.20	2.75	11.84	2.41	.78	.44
Influencing Others	6.58	2.16	6.87	1.87	-.78	.44
Communication	7.73	1.92	8.21	2.12	-1.30	.20
Drive/Results-Orientation	17.10	4.19	16.97	3.63	.19	.85
Tolerance for Stress	7.22	1.99	7.97	1.47	-2.36	.02
Integrity/Ethics	8.03	2.06	8.50	1.67	-1.36	.18
Overall	84.12	15.74	85.33	12.96	-.46	.65

Table 11

T-test values comparing Undergraduate and Graduate SALSA© scores

Dimension	**Undergrad Means**	**Undergrad SD's**	**Grad Means**	**Grad SD's**	**t**	**p**
Organizing/Planning/Visioning	11.04	2.83	14.18	2.14	-3.48	.00
Consideration/Team Skills	12.00	3.37	15.27	1.19	-3.16	.00
Problem Solving/Innovation	11.11	2.67	13.64	1.75	-2.99	.00
Influencing Others	6.42	1.83	8.18	1.89	-2.89	.01
Communication	7.91	2.42	8.55	1.44	-.84	.40
Drive/Results-Orientation	16.06	3.84	19.64	1.63	-3.02	.00
Tolerance for Stress	7.53	1.71	9.00	1.00	-2.76	.01
Integrity/Ethics	8.04	2.06	8.04	.92	-2.08	.04
Overall	80.23	15.52	97.82	5.88	-3.68	.00

Hypothesis 3 predicted that there would be significant positive correlations between the scores on SALSA© Forms A and B (overall and for each dimension). Performance on the two forms was significantly correlated (r = .84 , p = .00). Correlation coefficients were also calculated between dimension scores from Form A to Form B. All correlations between forms except those for Problem Solving/Innovation and Tolerance for Stress dimensions were significant at the $p < .05$ significance level, demonstrating empirical support for Hypothesis 3. Correlations between dimensions and forms are presented in Table 12.

Table 12.

Correlations between dimensions of SALSA© Forms A and B.

Dimension	**r**	**p**
Organizing/Planning/Visioning	.48	< .01
Consideration/Team Skills	.60	.00
Problem Solving/Innovation	.19	> .05
Influencing Others	.39	< .01
Communication	.29	< .05
Drive/Results-Orientation	.64	< .01
Tolerance for Stress	.03	> .05
Integrity/Ethics	.30	< .01
Overall	.84	.00

Additional analyses. Final analyses were conducted to examine previously reported trends not included in the proposed hypotheses. Means and standard deviations for dimension and overall scores were calculated by gender and program, and are presented in Tables 13 and 14, respectively.

Table 13.

Mean SALSA© Total Scores by Gender

		OPV	CTS	PSI	IO	Com	DRO	TS	IE	OVR
Female	*M*	20.50	23.11	21.06	12.61	14.33	30.83	14.50	14.78	88.14
	SD	7.18	7.54	6.49	4.40	4.50	8.29	3.98	4.67	16.03
Male	*M*	22.18	23.00	21.00	12.22	15.14	30.68	13.77	15.21	87.56
	SD	5.79	7.48	6.62	3.96	5.24	9.32	4.63	4.94	13.29
TOTAL	*M*	21.42	23.05	21.03	12.40	14.78	30.75	14.10	15.01	87.23
	SD	6.46	7.46	6.52	4.14	4.90	8.80	4.34	4.79	14.08

Table 14.

Mean SALSA© Total Scores by Degree/Program

		OPV	CTS	PSI	IO	Com	DRO	TS	IE	OVR
MBA	*M*	15.25	13.50	12.75	8.50	9.25	21.25	8.25	9.25	98.00
	SD	2.82	4.24	2.82	2.07	2.12	6.04	2.92	2.38	12.83
Ed.D.	*M*	13.75	16.00	12.75	8.00	10.25	19.50	8.25	8.75	97.25
	SD	3.11	2.83	3.69	1.85	2.49	2.98	1.67	3.01	12.91
M.A. I/O	*M*	28.36	30.55	27.27	16.36	17.09	39.27	18.00	18.73	97.81
	SD	4.27	2.38	3.50	3.78	2.88	3.26	2.00	1.85	5.88
Athlete	*M*	19.00	22.33	20.11	12.33	14.56	28.67	14.67	15.00	80.44
	SD	6.55	8.57	6.70	4.19	5.26	7.13	3.63	4.67	15.77
ROTC	*M*	23.66	24.86	23.31	13.09	16.46	33.89	15.26	16.65	83.88
	SD	4.46	5.52	4.20	3.41	4.55	7.44	3.33	3.74	11.57
TOTAL	*M*	21.42	23.05	21.03	12.40	14.78	30.75	14.10	15.01	152.63
	SD	6.46	7.46	6.52	4.14	4.90	8.80	4.34	4.79	39.99

Note: OPV = Organizing/Planning/Visioning; CTS = Consideration/Team Skills; PSI = Problem Solving/Innovation; IO = Influencing Others; Com = Communication; DRO = Drive/Results-Orientation; TS = Tolerance for Stress; IE = Integrity/Ethics; OVR=Overall Score.
******MBA and Ed.D. respondents were given the short form SALSA© assessment. Equivalent scores were obtained by multiplying their overall dimensions by a conversion factor of 2. Their dimension scores were not converted.

Although no hypotheses were proposed regarding gender or program type, it was of interest to determine if SALSA© scores were moderated by either of these variables. Grant (2009) noted significant main effects for gender. In order to compare short and long-form scores, short form dimension scores were doubled. A T-test comparing adjusted overall SALSA© scores of males (M = 87.56, SD = 12.39) and females (M = 88.14, SD = 16.03) revealed no significant differences between sexes (t (77) = .181, p = .856).

A one-way ANOVA was conducted on adjusted SALSA composite scores to determine whether a main effect existed for degree type (F (4, 74) = 6.78, p = .000, η^2 = .27). Post-hoc Tukey's test revealed significant differences between groups. Athletes (M = 80.44, SD = 15.77) were found to have significantly lower overall scores compared to MBA (M = 98.00, SD = 12.83), Ed.D. (M = 97.25, SD = 12.91), and M.A. I/O (M = 97.81, SD = 5.88) students. Additionally, ROTC cadets (M = 83.88, SD = 11.57) were found to have significantly lower overall scores compared to MBA (M = 98.00, SD = 12.83), Ed.D. (M = 97.25, SD = 12.91), and M.A. I/O (M = 97.81, SD = 5.88) students. Scores among graduate students (e.g., MBA, Ed.D., and M.A. I/O) did not significantly differ, nor did scores between undergraduate students (e.g., student athlete leaders and ROTC cadets).

Discussion

Alternate Forms Reliability

The current study assessed alternate forms reliability of the SALSA© assessment. In order to evaluate the extent to which program participants have acquired leadership abilities, the Center for Leadership Excellence administers their Assessment Center to program participants before and after students complete the program. The SALSA© has been used by the CLE in lieu of assessment centers to assess leadership qualities of students. The creation of alternate forms of SALSA© by Grant (2009) enabled the SALSA© to be used as both a pre-and post-assessment for the CLE, while eliminating practice effects associated with employing the full-length assessment for this purpose. Original alternate forms of the assessment contained 72 items each, and possessed a strong coefficient of equivalence (r = .91), indicating that the two forms were relatively equivalent measures of leadership ability.

The current study revisited the data and procedures used by Grant to create the alternate forms, and produced new, abbreviated forms of the assessment; SALSA© Form A and Form B. Each form is comprised of 40 items, with a total of 5 items assessing each of eight dimensions. A calculation of the coefficient of equivalence indicates that the two forms are strongly correlated (r = .84, p = .000), and therefore also approximately equivalent measures of leadership ability. These new forms of SALSA© should adequately address the need for alternate forms for pre-and post-assessment. Furthermore, the abbreviated forms will likely streamline the assessment process by requiring less time for participants to complete. Correlations between individual dimensions on the two forms ranged from r = .30 to r = .64.

Despite the new forms having fewer items for each dimension, they exhibited reliability coefficients similar to both the full-length and previous alternate forms. Overall alpha coefficients for Forms A and B were $\alpha = .73$ and $\alpha = .75$, respectively. The full-length SALSA© alpha coefficient for the current sample was $\alpha = .86$. Alphas for the alternate forms established by Grant (2009) were $\alpha = .76$ and .78, and the full-length SALSA© alpha coefficient was $\alpha = .91$. Thus, despite the reduction in the total number of items on each form (130 to 72 to 40), minimal reliability was lost across the entire assessment. High overall alpha coefficients indicate that all dimensions included are effectively measuring the same construct.

Alpha coefficients for individual dimensions also encountered minimal loss of reliability for all dimensions except Tolerance for Stress and Influencing Others. The Tolerance for Stress sub-scale was problematic in Grant's (2009) original study in that it yielded a moderate coefficient alpha for the full form SALSA© ($\alpha = .41$), but an extremely low alpha level for SALSA Form A ($\alpha = .07$). Alphas for Forms A and B on this dimension were also quite low ($\alpha = .11$ and .22, respectively).

In the current study, two types of analyses were used to determine item difficulty: SME calibration ratings and p-values calculated from SALSA© response data. There was a 51.5% agreement on difficulty categorization of items between the two types of analyses. A combination of the two analyses resulted in the categorization of 130 items by difficulty. A total of 46 items were considered easy items, 58 were considered moderate items, and 26 were considered to be difficult items. Due to the fact that two different methods of difficulty categorization were used to determine the ultimate difficulty level of each item, we can expect that the final difficulty categorization of each

item is an accurate estimate. The high levels of agreement (72.3%) between the P-value based difficulty categorizations reported by Grant (2009) and those obtained in the current study further reinforce this supposition.

Out of 130 items, 80 were retained for the final alternate forms. Of these, 36 (45%) were categorized as easy, 36 (45%) were categorized moderate, and 8 (10%) were categorized as difficult items. An ideal test of leadership knowledge would be comprised of items assessing knowledge at both extremes of the distribution of leadership knowledge, in order to accurately assess and differentiate between students entering the program with presumably low levels of leadership knowledge, and those leaving the program, likely possessing high levels of such knowledge. Although the number of items categorized as difficult is rather low for the new forms, the high reliabilities and corrected item total correlations for each test serve as redeeming qualities for the new short forms of SALSA©.

Hypothesis 1 predicted that the current sample of individuals would have higher overall scores on the SALSA© than the response set used by Grant (2009) as individuals in the current sample possess considerably more leadership training than most of the individuals used in the previous sample. Significant differences between groups were found only for the dimension of Tolerance for Stress.

Hypothesis 2 predicted that overall SALSA© scores for graduate students would be significantly higher than overall undergraduate scores due to their advanced training and experience levels. Graduate student scores were found to be significantly higher than undergraduate student scores, which provided evidence in support of the hypothesis.

Hypothesis 3 predicted that SALSA© Forms I and II would be significantly correlated across dimensions and overall scores. Scores on the two forms were significantly correlated ($r = .84$, $p = .00$). Correlations between forms across dimensions ranged from .30 to .64; six out of eight correlations were significant at the $p < .05$ level, indicating support for this hypothesis.

Additional Findings

Although Grant (2009) reported significant main effects for gender, no such difference was found in the current sample. The results of a t-test comparing adjusted overall SALSA© scores of males and females revealed no significant performance effects for gender. One possible explanation for the lack of effects in the current sample may be due to the sample's composition. The sample used by Grant was comprised mainly of undergraduate students enrolled in the CLE Leadership Certificate Program, who had received some leadership instruction. In comparison, most of the individuals in the current sample had undergone extensive leadership training. It is possible that gender differences may exist among individuals with low levels of leadership knowledge, but may decrease as knowledge increases.

Significant differences were found between composite scores for student athletes, ROTC cadets and graduate (MBA, Ed.D., and M.A. I/O) students. The results of a one-way ANOVA and post-hoc tests comparing composite scores across groups revealed that athletes and ROTC cadets had significantly lower composite scores than their graduate student peers. These differences are likely due to two factors. First, the graduate students are enrolled in specialized programs providing them with extensive leadership training. Although the student athletes and ROTC cadets tend to have more leadership experience

and training than their undergraduate peers, it is likely that the training received by the graduate respondents provides leadership knowledge above and beyond that of the athlete and ROTC respondents. Second, the increased levels of general life experience possessed by the graduate respondents may have yielded positive performance effects on the assessment.

Limitations

The findings of the current study face several limitations. The initial limitation with all leadership SJT research is a lack of available qualified respondents. In order to determine optimal leadership knowledge, it is essential to recruit individuals with extensive leadership training and experience as participants. Unfortunately, it is often difficult to choose criterion for determining leadership experience. Aside from this hurdle, researchers may face difficulties in recruiting such individuals for participation due to a variety of reasons.

The current study used Grant's (2009) SME calibration data to determine item difficulty. The small number of SME's (6) used by Grant to calibrate item responses may serve as a limitation. A larger sample of SME's would have allowed a higher threshold of agreement to be attained, and the effects of extreme ratings would have been reduced. A small number of respondents (N = 80) were used to establish alternate forms in the current study. It is recommended that a sample size of at least several hundred respondents be used for such practices in order to ensure psychometrically sound results. Despite the agreement with the previous study and positive results obtained in this study, it is recommended that later studies attempt to establish alternate forms of the SALSA© using sample sizes of the recommended magnitude.

Difficulty estimates for each item were determined using one sample of SME's (calibration ratings) and one response set (p-values). Although SME and P-value based categorizations displayed acceptably high agreement with one another, as well as with the p-value estimates reported by Grant, it is still impossible to determine whether or not these findings will generalize to other samples. Furthermore, the same response set was used to calculate p-values and alternate forms reliability coefficients. Thus, no actual cross-validation has occurred using the newly constructed forms. As with the previous study, it is highly recommended that the new forms be cross-validated using a larger sample of respondents.

Finally, the current study produced two abbreviated alternate forms of the SALSA© consisting of 5 items on each of 8 dimensions, for a total of 40 items per form. This small number of items, although intended to streamline the assessment process, is likely to limit reliability estimates for the test. Using the current sample, acceptable reliability estimates were obtained for most dimensions on each form. Again, it is recommended that the new alternate forms be cross-validated using new, sufficiently large response sets. If the alternate forms are shown to lack sufficient reliability, the SALSA© may need to be administered in its full-length, 130-item form in order to ensure valid, reliable measurement of leadership knowledge.

Directions for Future Research

Significant differences were found among scores for graduate and undergraduate students. Future studies may attempt to examine the source(s) of these differences through the use of additional instruments and procedures. Although no significant differences between overall scores for gender were found in the current sample,

significant main effects were reported in Grant (2009). It is possible that the differences reported by Grant were due to characteristics of the sample used in that study. Future studies might attempt to determine the existence of such differences at both extremes of the distribution of leadership knowledge.

As previously mentioned, cross-validation of the newly created SALSA© Forms A and B also should be pursued in future studies. Ideally, such studies would incorporate adequate samples of respondents. The format of the SALSA© also may be examined in future studies. As previously discussed, SJTs with a "should do" format correlate with measures of cognitive ability, while SJTs administered in a "would do" format correlate with measures of personality. All forms of SALSA© use a "should do" format. Thus, it is of interest to determine whether SALSA© is actually examining leadership knowledge or general cognitive ability. It would be of interest to administer SALSA© in conjunction with an established measure of general cognitive ability, and then use those scores to control for cognitive ability. Such a study would likely determine whether SALSA© is indeed measuring leadership knowledge or general mental ability.

With regards to the use of the SALSA© for leadership assessment purposes by the CLE, future studies could examine the relationship between the number of LEAD courses taken or overall GPA in LEAD courses in order to determine the overall effectiveness of the CLE's leadership program in teaching leadership knowledge.

Conclusions

In summary, the SALSA© SJT was developed to assess eight dimensions of leadership. The current response sample was used to construct equivalent abbreviated alternate forms of the test, SALSA© Forms A and B, which are intended to be used as a

pre-and post-test of leadership knowledge. The forms appear to have acceptable distributions of easy, moderate, and difficult items, although the inclusion of a few more difficult items would improve this status. These forms yielded acceptable psychometric properties, and likely are appropriate for use in the intended manner. The high coefficient of equivalence suggests that the SALSA© Form A and Form B are acceptable alternate forms of the same instrument. Accordingly, test users should be confident in using SALSA© Form A and Form B as equivalent pre- and post- measures of leadership knowledge.

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.

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

Christian, M., Edwards, B., & Bradley, J. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*, 83-117.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-358.

Grant, K. (2009). The validation of a situational judgment test to measure leadership behavior. (Master's thesis). Western Kentucky University, Bowling Green, KY.

Gaughler, B., Rosenthal, D., Thornton, G., & Bentson, C. (1987). Meta-Analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511.

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

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

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*(3), 442-452.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of literature. *Personnel Review, 37*(4), 426-441.

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

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

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

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

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

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.

Normansell, D. (2011). The Situational Assessment of Leadership: Student Assessment (SALSA©): An evaluation of the convergent validity with multi-source feedback in Division I intercollegiate athletics (Master's thesis). Western Kentucky University, Bowling Green, KY.

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.

Sackett, P. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology*, *40*, 13-25.

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

Slack, P. (2010). A Situational Assessment of Student Leadership: An Evaluation of Alternate Forms Reliability and Convergent Validity. (Master's thesis). Western Kentucky University, Bowling Green, KY.

Stroupe, H. (2010). An Evaluation of the Convergent Validity of Multi-Source Feedback with Situational Assessment of Leadership: Student Assessment (SALSA©). (Master's Thesis). Western Kentucky University, Bowling Green, KY.

Tucker, J. (2011). An evaluation of the convergent validity the Situational Assessment of Leadership: Student Assessment (SALSA©) with of multi-source feedback with MBA and Ed.D. in Educational Leadership students (Master's Thesis). Western Kentucky University, Bowling Green, KY.

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

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

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

Appendix A
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 B
WKU Human Subjects Review Board Approval Form

WKU

A LEADING AMERICAN UNIVERSITY WITH AN INTERNATIONAL REACH
HUMAN SUBJECTS REVIEW BOARD

In future correspondence, please refer to HS11-231. March 24, 2011

Ashley N. Wade
Dr. Shoenfelt
Psychology
WKU

Ashley N. Wade:

Your research project, *An Evaluation of Alternate Forms Reliability and Convergent Validity of the Situational Assessment of Leadership Student Assessment© (SALSA©)*, 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 Manager
Office of Research
Western Kentucky University

cc: HS file number Wade HS11-231

INSTITUTIONAL REVIEW BOARD
WKU
APPROVED

Appendix C
Test Map for Alternate Forms

Item	SMEDiff	Difficulty 1	CurrentP	Difficulty 2	GrantP	Difficulty 3	Final Difficulty	Form
Org-1	1.17	1.00	0.81	1.00	0.902	1.00	1.00	
Org-2	1	2.00	0.67	2.00	0.623	2.00	2.00	
Org-3	0.67	2.00	0.61	2.00	0.607	2.00	2.00	A
Org-4	1.16	1.00	0.68	2.00	0.803	1.00	1.00	
Org-5	0.83	2.00	0.81	1.00	0.738	2.00	2.00	
Org-6	0.83	2.00	0.80	1.00	0.705	2.00	2.00	A
Org-7	1.33	1.00	0.54	2.00	0.721	2.00	2.00	A
Org-8	1	2.00	0.74	2.00	0.623	2.00	2.00	A
Org-9	2.33	1.00	0.90	1.00	0.869	1.00	1.00	B
Org-10	1.16	1.00	0.57	2.00	0.705	2.00	2.00	B
Org-11	0.5	2.00	0.46	3.00	0.459	3.00	3.00	B
Org-12	0.33	3.00	0.61	2.00	0.525	2.00	2.00	A
Org-13	0.5	2.00	0.42	3.00	0.443	3.00	3.00	
Org-14	1.66	1.00	0.92	1.00	0.902	1.00	1.00	B
Org-15	0.66	2.00	0.38	3.00	0.328	3.00	3.00	
Org-16	0.83	2.00	0.54	2.00	0.541	2.00	2.00	
Org-17	1.33	1.00	0.77	1.00	0.82	1.00	1.00	B
Org-18	1.34	1.00	0.79	1.00	0.738	2.00	1.00	
Con-1	0.8	2.00	0.58	2.00	0.574	2.00	2.00	A
Con02	0.67	2.00	0.67	2.00	0.836	1.00	2.00	
Con-03	0.84	2.00	0.86	1.00	0.705	2.00	2.00	A
Con-04	1.17	1.00	0.71	2.00	0.754	1.00	1.00	
Con-05	0.5	2.00	0.14	3.00	0.246	3.00	3.00	
Con-06	0.5	2.00	0.45	3.00	0.492	3.00	3.00	
Con-07	1.17	1.00	0.90	1.00	0.902	1.00	1.00	A
Con-08	1.17	1.00	0.88	1.00	0.787	1.00	1.00	
Con-09	1.83	1.00	0.90	1.00	0.902	1.00	1.00	A
Con-10	0.33	3.00	0.31	3.00	0.311	3.00	3.00	
Con-11	1.83	1.00	0.83	1.00	0.836	1.00	1.00	A
Con-12	1	2.00	0.12	3.00	0.197	3.00	3.00	
Con-13	1.17	1.00	0.69	2.00	0.656	2.00	2.00	
Con-14	1.34	1.00	0.79	1.00	0.738	2.00	1.00	B
Con-15	0.66	2.00	0.78	1.00	0.59	2.00	2.00	B
Con-16	0.5	2.00	0.32	3.00	0.459	3.00	3.00	
Con-17	1	2.00	0.54	2.00	0.41	3.00	2.00	
Con-18	1.17	1.00	0.71	2.00	0.705	2.00	2.00	B
Con-19	0.33	3.00	0.24	3.00	0.262	3.00	3.00	
Con-20	1.83	1.00	0.78	1.00	0.705	2.00	1.00	B
Con-21	1.66	1.00	0.79	1.00	0.852	1.00	1.00	B

Item	SMEDiff	Difficulty 1	CurrentP	Difficulty 2	GrantP	Difficulty 3	Final Difficulty	Form
Prob-1	1.5	1.00	0.58	2.00	0.656	2.00	1.00	
Prob-2	1.33	1.00	0.83	1.00	0.754	1.00	1.00	A
Prob-3	1.17	1.00	0.97	1.00	0.951	1.00	1.00	A
Prob-4	0.5	2.00	0.15	3.00	0.115	3.00	3.00	B
Prob-5	0.66	2.00	0.51	2.00	0.656	2.00	2.00	A
Prob-6	1.17	1.00	0.57	2.00	0.623	2.00	2.00	B
Prob-7	1.16	1.00	0.82	1.00	0.721	2.00	1.00	
Prob-8	0.84	2.00	0.39	3.00	0.41	3.00	3.00	
Prob-9	1.5	1.00	0.94	1.00	0.951	1.00	1.00	B
Prob-10	0.5	2.00	0.47	3.00	0.95	1.00	2.00	A
Prob-11	0.66	2.00	0.62	2.00	0.656	2.00	2.00	
Prob-12	0.5	2.00	0.41	3.00	0.475	3.00	3.00	
Prob-13	0.84	2.00	0.33	3.00	0.295	3.00	3.00	
Prob-14	1.16	1.00	0.81	1.00	0.639	2.00	1.00	A
Prob-15	1.17	1.00	0.86	1.00	0.918	1.00	1.00	
Prob-16	0.67	2.00	0.46	3.00	0.508	2.00	2.00	B
Prob-17	1	2.00	0.67	2.00	0.721	2.00	2.00	B
Prob-18	0.84	2.00	0.78	1.00	0.902	1.00	1.00	
Prob-19	0.67	2.00	0.64	2.00	0.623	2.00	2.00	
Influ-1	1	2.00	0.38	3.00	0.508	2.00	2.00	B
Influ-2	0.67	2.00	0.46	3.00	0.459	3.00	3.00	B
Influ-3	0.83	2.00	0.81	1.00	0.721	2.00	2.00	A
Influ-4	0.5	2.00	0.78	1.00	0.869	1.00	1.00	A
Influ-5	1.34	1.00	0.78	1.00	0.754	1.00	1.00	A
Influ-6	1.16	1.00	0.70	2.00	0.639	2.00	2.00	A
Influ-7	1	2.00	0.80	1.00	0.672	2.00	2.00	B
Influ-8	0.33	3.00	0.57	2.00	0.525	2.00	2.00	B
Influ-9	0.67	2.00	0.35	3.00	0.344	3.00	3.00	A
Influ-10	0.17	3.00	0.52	2.00	0.246	3.00	3.00	
Influ-11	1.5	1.00	0.81	1.00	0.803	1.00	1.00	B
Comm-1	0.83	2.00	0.69	2.00	0.672	2.00	2.00	A
Comm-2	1.84	1.00	0.82	1.00	0.82	1.00	1.00	A
Comm-3	1.33	1.00	0.62	2.00	0.475	3.00	2.00	B
Comm-4	0.83	2.00	0.63	2.00	0.525	2.00	2.00	A
Comm-5	0.67	2.00	0.58	2.00	0.721	2.00	2.00	B
Comm-6	0.84	2.00	0.33	3.00	0.377	3.00	3.00	B
Comm-7	2	1.00	0.90	1.00	0.754	1.00	1.00	A
Comm-8	0.5	2.00	0.49	3.00	0.393	3.00	3.00	
Comm-9	1.83	1.00	0.96	1.00	0.934	1.00	1.00	B
Comm-10	0.5	2.00	0.50	2.00	0.393	3.00	2.00	
Comm-11	1.17	1.00	0.91	1.00	0.836	1.00	1.00	A
Comm-12	1.17	1.00	0.78	1.00	0.754	1.00	1.00	B

Item	SMEDiff	Difficulty 1	CurrentP	Difficulty 2	GrantP	Difficulty 3	Final Difficulty	Form
Res-1	0.5	2.00	0.54	2.00	0.721	2.00	2.00	
Res-2	1.34	1.00	0.54	2.00	0.639	2.00	2.00	
Res-3	1.13	1.00	0.74	2.00	0.721	2.00	2.00	
Res-4	2.5	1.00	0.92	1.00	0.918	1.00	1.00	A
Res-5	1	2.00	0.47	3.00	0.361	3.00	3.00	
Res-6	0.5	2.00	0.49	3.00	0.492	3.00	3.00	
Res-7	1.5	1.00	0.81	1.00	0.836	1.00	1.00	A
Res-8	0.5	2.00	0.51	2.00	0.443	3.00	2.00	
Res-9	0.84	2.00	0.71	2.00	0.557	2.00	2.00	
Res-10	1.5	1.00	0.85	1.00	0.754	1.00	1.00	B
Res-11	0.84	2.00	0.72	2.00	0.721	2.00	2.00	
Res-12	0.5	2.00	0.41	3.00	0.426	3.00	3.00	
Res-13	0.83	2.00	0.68	2.00	0.852	1.00	2.00	A
Res-14	1	2.00	0.77	1.00	0.738	2.00	2.00	A
Res-15	2.17	1.00	0.90	1.00	0.918	1.00	1.00	A
Res-16	0.84	2.00	0.69	2.00	0.705	2.00	2.00	
Res-17	1.17	1.00	0.61	2.00	0.82	1.00	1.00	B
Res-18	1	2.00	0.80	1.00	0.738	2.00	2.00	B
Res-19	0.84	2.00	0.65	2.00	0.738	2.00	2.00	
Res-20	1.33	1.00	0.78	1.00	0.836	1.00	1.00	B
Res-21	1.16	1.00	0.84	1.00	0.721	2.00	1.00	
Res-22	1	2.00	0.71	2.00	0.672	2.00	2.00	B
Res-23	0.33	3.00	0.56	2.00	0.475	3.00	3.00	
Res-24	1	2.00	0.80	1.00	0.574	2.00	2.00	
Res-25	0.34	3.00	0.68	2.00	0.672	2.00	2.00	
Tol-1	0.33	3.00	0.76	1.00	0.475	3.00	2.00	A
Tol-2	0.5	2.00	0.45	3.00	0.541	2.00	3.00	B
Tol-3	1	2.00	0.92	1.00	0.787	1.00	1.00	A
Tol-4	0.33	3.00	0.57	2.00	0.525	2.00	2.00	A
Tol-5	0.67	2.00	0.83	1.00	0.803	1.00	1.00	B
Tol-6	1.34	1.00	0.83	1.00	0.836	1.00	1.00	A
Tol-7	0.66	2.00	0.55	2.00	0.541	2.00	2.00	B
Tol-8	0.67	2.00	0.76	1.00	0.721	2.00	2.00	
Tol-9	2.17	1.00	0.91	1.00	0.754	1.00	1.00	B
Tol-10	0.5	2.00	0.72	2.00	0.738	2.00	2.00	A
Tol-11	0.67	2.00	0.60	2.00	0.475	3.00	2.00	B
Int-1	0.67	2.00	0.77	1.00	0.656	2.00	2.00	A
Int-2	0.84	2.00	0.50	2.00	0.492	3.00	2.00	A
Int-3	0.83	2.00	0.31	3.00	0.393	3.00	3.00	
Int-4	0.34	3.00	0.38	3.00	0.475	3.00	3.00	A
Int-5	1.67	1.00	0.86	1.00	0.885	1.00	1.00	A

Item	SMEDiff	Difficulty 1	CurrentP	Difficulty 2	GrantP	Difficulty 3	Final Difficulty	Form
Int-6	1.83	1.00	0.74	2.00	0.639	2.00	2.00	
Int-7	1.34	1.00	0.82	1.00	0.738	2.00	1.00	A
Int-8	2	1.00	0.91	1.00	0.787	1.00	1.00	B
Int-9	2.5	1.00	0.97	1.00	0.934	1.00	1.00	B
Int-10	1.34	1.00	0.61	2.00	0.492	3.00	2.00	B
Int-11	0.5	2.00	0.58	2.00	0.443	3.00	2.00	B
Int-12	1.33	1.00	0.85	1.00	0.787	1.00	1.00	
Int-13	0.67	2.00	0.13	3.00	0.279	3.00	3.00	B

For Difficulty 1, Difficulty 2, Difficulty 3, and Final Difficulty Categories: 1= “Easy”, 2 = “Moderate”, 3 = “Difficult” Item Classification.

NOTE: Difficulty 1= Classification based on SME ratings, Difficulty 2 = Classification based on Current P-values, Difficulty 3= Classification based on P-values obtained by Grant (2009).