Running head: Situational Judgment Item Validity

Toward an Understanding of Situational Judgment Item Validity

Michael A. McDaniel

Virginia Commonwealth University

Joseph Psotka and Peter J. Legree

U.S. Army Research Institute

Paper presented at the 24th annual conference of the Society of Industrial and Organizational Psychology. New Orleans, April, 2009. This research was funded by the Army Research Institute through a contract awarded to Work Skills First, Inc. Correspondence on this paper may be directed to Michael McDaniel at mamcdani@vcu.edu.

Abstract

Consensually scored situational judgment tests using Likert scale response formats can be substantially improved with respect to validity, Black-White mean differences, resistance to faking, and test length. This improvement is achieved with two simple adjustments. The first adjustment is controlling for elevation and scatter (Cronbach & Gleser, 1953). This adjustment substantially improves item validity. Also, because there is a race difference in the preference for extreme responses on Likert scales (Bachman & O'Malley, 1984), these adjustments substantially reduce Black-White mean score differences. In addition, these adjustments eliminate the score elevation associated with the faking strategy of avoiding extreme responses (Cullen, Sackett, & Lievens, 2006). Item validity is shown to have a U-shaped relationship with item means. The second adjustment is to drop items with mid-range item means. This permits the scale to be shortened dramatically without harming validity.

Situational judgment tests (SJTs) present job applicants with written or video-based problem scenarios and a set of possible response options. Job applicants evaluate the effectiveness of the responses for addressing the problem described in the scenario. Although SJTs have been used in personnel selection for about 80 years (McDaniel, Morgeson, Finnegan, Campion & Braverman, 2001; Moss, 1926), and have been the subject of substantial research in the last two decades (McDaniel, Hartman, Whetzel & Grubb, 2007; Motowidlo, Dunnette & Carter, 1990; Weekley & Ployhart, 2006), there is very little research addressing how to best build and score SJTs (Schmitt & Chan, 2006; Weekley, Ployhart & Holtz, 2006). There is also little knowledge concerning the best approaches to build scales using SJT items to tap specific constructs. In the absence of this knowledge, many approaches have evolved for developing and scoring SJTs (Weekley, Ployhart & Holtz, 2006; Bergman, Drasgow, Donovan, Henning & Juraska, 2006) and the effectiveness of these methods for maximizing criterion-related and construct validity is largely unknown.

Unlike cognitive ability or job knowledge tests, response options in SJTs cannot easily be declared correct or incorrect. As such, items are typically scored using some form of consensus judgment (Legree, Psotka, Tremble, & Bourne, 2005). Expert judges are often asked to reach consensus concerning which responses are preferred (Weekley et al., 2006). Consensus may also be based on the responses of applicants, incumbents, or supervisors of incumbents. In such applications, the means of the respondents are considered the correct response.

There are several different response formats. When response instructions request that a respondent pick a single response (identify the behavior that you would most [or least] likely do, identify the most effective [or ineffective] response), the means are used to identify the response judged to be correct. Another format involves asking respondents to rate each response option using a Likert scale. Using this format, an applicant's score is often expressed as a deviation or a squared deviation from the mean. Alternatively, the mean is used to determine whether the item response is judged effective or ineffective, and the item is scored dichotomously (McDaniel, Whetzel & Nguyen, 2006).

Consensual scoring is a form of profile matching. One profile consists of the means of the items collected from some group (e.g., experts, applicants, incumbents). The other profile is the item responses of one respondent. A respondent's score on the SJT is a function of the degree of match between the respondent's answers and the group means. Cronbach and Gleser (1953) conceptualized profile matching with respect to elevation, scatter, and shape. Elevation is the mean of the items across a respondent. Scatter reflects the magnitude of a respondent's score deviations from the respondent's own mean. If one standardizes scores using a within-person *z* transformation, all respondents would have the same mean (zero) and the same standard deviation (one) across items. This transformation removes information from the scores related to elevation and scatter because all respondents have identical elevation and scatter. The remaining score information in the within-person standardized scores is called shape. Cronbach and Gleser argued that the investigator should consider whether elevation and scatter are important in their profile matching application. For SJTs, we suggest that elevation and scatter reflect response tendencies such as a preference for using one end of the Likert rating scale over another (e.g., rating most responses as effective, or rating most responses as ineffective) or preferences for extreme or more mid-scale Likert ratings (e.g., on a nine-point Likert scale preferring ratings of one and nine over ratings of three and seven). We assert that these response tendencies are primarily criterion-irrelevant noise in the ratings which damage the SJT item validity. Thus, we offer the following hypothesis:

Hypothesis 1: SJT scoring methods that control for elevation and scatter will yield higher response option validities than methods that do not.

Although seldom considered in the I/O and management literatures, there are Black-White differences in the use of Likert scales (Bachman & O'Malley, 1984). Specifically, Blacks tend to use extreme rating points (e.g., 1 and 7 on a 7-point scale) with greater frequency than Whites on average. Extreme rating points, on average, will have larger deviations from the consensual mean resulting in less favorable scores. The tendency of Blacks to use extreme ratings more so than Whites, will tend to increase Black-White differences. Controlling for elevation and scatter adjusts for individual differences in extreme responding. Thus, we offer the following hypothesis:

Hypothesis 2: SJT scoring methods that control for elevation and scatter will yield lower Black-White mean differences than methods that do not.

Cullen, Sackett and Lievens (2006) examined the coachability of two SJTs. In addition to evaluating a curriculum focused on strategies for improving scores, they simulated what would happen if a respondent was coached not to endorse extreme values. This simulation was done by changing the 1 and 2 responses to a 3 and by changing the 6 and 7 responses to a 5. They discovered that scores could be improved by 1.57 standard deviations if examinees did not endorse extreme answers (e.g., 1 or 2 and 6 or 7). In this paper, we refer to this strategy as *avoiding extreme responses*. Controlling for elevation and scatter adjusts for individual differences in extreme responding. Thus, we offer the following hypothesis:

Hypothesis 3: SJT scoring methods that control for elevation and scatter will reduce score elevation associated with a faking strategy of avoiding extreme responses.

Although intended to be a faking strategy, respondents who employ the avoiding extreme responses strategy should have reduced Black-White differences in extreme responding. When ratings have been adjusted for elevation and scatter, the mean Black-White difference in extreme responding should be reduced or removed and thus the use of the avoiding extreme responses strategy should have little impact on Black-White differences in SJT scale scores. However, for consensually scored SJTs using raw (i.e., unadjusted) Likert ratings, the data are expected to show Black-White differences in extreme responding and the avoiding extreme responses strategy should reduce the Black-White mean differences in the SJT scale score. Thus, we offer the following hypothesis:

Hypothesis 4: SJTs with consensus scoring based on raw (i.e. unadjusted) Likert ratings will show smaller Black-Whites differences in SJT scale scores for those tests completed with an avoiding extreme responses strategy.

For a seven-point rating scale such as described in the Cullen et al. (2006) study, the respondents who follow the avoiding extreme response strategy would only respond using three rating points: 4, 5, and 6. Although intended as a faking strategy, it could also be a scale construction strategy. That is, the researcher could recode more extreme ratings to be more moderate responses. Such a scale construction method largely controls for elevation and scatter. Thus, we suggest that tests completed with this strategy will show large magnitude correlations with other methods that control for elevation and scatter (e.g., within-person *z* transformations and dichotomous scoring). It would also follow that tests completed with this faking strategy

will have higher criterion-related validity than tests scored using raw (i.e., unadjusted) Likert scales. Thus, we offer the following hypotheses:

Hypothesis 5: SJT scale scores based on the avoiding extreme responses faking strategy will have large magnitude correlations with SJT scales that control for elevation and scatter.

Hypothesis 6: SJT raw consensus scale scores based on the avoiding extreme responses faking strategy will have larger criterion-related validity than SJT raw consensus scale scores based on raw (i.e., unadjusted) Likert scales.

Across respondents, the rated mean effectiveness on Likert scales varies across response options. Some response options have a mean indicating that the behavior, as rated by most respondents, is an effective solution to the problem described in the item stem and other responses have means indicating that most respondents believe that the behavior is ineffective. Response options also have variance. Some responses have low variance indicating that most respondents rated the response option near its mean rating. Other response options have high variance indicating that there was substantial disagreement among respondents concerning the effectiveness of the response option. This disagreement may reflect some ambiguity in the response option that requires the respondent to make inferences about the response option and/or the scenario. For example, if the situational judgment scenario concerns a miscommunication between a supervisor and a subordinate that has resulted in the subordinate feeling ill-treated, the response option "talk to your supervisor" is not informative concerning the content of the talk. Some might infer that the purpose of the talk is to resolve the miscommunication politely and the respondent might judge this to be an effective behavior. Another might infer that the purpose of the talk is to express anger at the supervisor and the respondent might judge this to be an ineffective behavior. When respondents disagree on the effectiveness of the response option, the variance of the ratings reflects this disagreement. We suspect that a response option with a larger than typical variance is more likely to have a mean near the midpoint of a Likert scale (e.g., near a 5 on a 9-point scale).

Item validity may be related to item means. We located two studies that examined the relationship between consensual means of experts and item validity. Both Waugh and Russell (2006) and Putka and Waugh (2007) reported U-shaped relationships between item validity such that items with low or high means had the highest validity. We argue that response options with means near the mid-point of the Likert scale have less informational value than response options near either the low or high end of the Likert scale. They might be less relevant to the scenario presented in the stem and thus provide little information on whether the respondent knows how to respond effectively. They might also be near the mid-point because the respondents show substantial variance (i.e., little agreement) on the effectiveness of the response option. Under either explanation, the response options have less information and may be less valid. Thus we offer the following hypothesis:

Hypothesis 7: There will be a U-shaped relationship between response option respondent means and item criterion-related validity such that items with low means and high means will be more valid than items with means near the mid-point of the Likert scale. This hypothesis applies to the raw consensus, standardized consensus, and dichotomous consensus scores.

Method

Measures

Situational judgment test. An SJT with 20 scenarios and 136 response options was developed as part of a U.S. Army Research Institute project designed to explore criterion-related and construct validity issues related to military attrition. Respondents were asked to read each scenario and rate the effectiveness of various response options on a nine-point Likert scale. Thus, the 136 response options result in 136 scorable items. However, each of the 136 response options is nested within scenario. This nesting has implications for the data analysis due to possible non-independence issues (Bliese, 2002). For most of our analyses, the item is the unit of analysis (N = 136).

Three consensual scoring strategies were used. The first strategy, which we label "raw consensus," is a common one in SJT applications. The raw consensus group means serve as the answer key and a respondent's score is the squared deviation from the mean. This score was then inverted so that high scores indicated a close match with the consensus means across respondents. The raw consensus score does not control for elevation or scatter. We refer to the rating data and consensus score as "raw" because the Likert ratings have not been adjusted or modified in any way. The second strategy, which we label "standardized consensus," first requires a within-person z standardization such that the mean across items for each respondent is zero and its standard deviation is one. As with the first strategy, a respondent's score is the squared deviation from the group mean which is then inverted such that high scores indicate a close match with the group's means. The rating data and the scores are referred to as standardized because of the z transformation that controls for elevation and scatter. The final scoring strategy, which we label "dichotomous consensus," uses the raw item mean across respondents to determine if a response option is effective (a group mean of 5.0 or above on the nine-point scale) or ineffective (a group mean below 5.0). Although 5.0 is the middle point on a nine-point scale and is neither effective or ineffective, no item mean was exactly 5.0 and thus all items could be classified as either effective or ineffective. If the group mean indicated that the response option was effective and the respondent indicated that the response option was effective (by giving a rating of 5 or above), the respondent received a score of one, otherwise a score of zero. Likewise, if the group mean indicated that the response option is ineffective and the respondent indicated that the response option was ineffective (a rating of 1 through 4), the respondent received a score of 1, otherwise a score of zero. This scoring method largely controls for elevation and scatter in that response options of one through four are treated identically and response options five through nine also are treated identically.

To evaluate hypotheses relevant to the avoiding extreme responses faking strategy, we developed alternative versions of our three scales in a manner similar to that of Cullen et al. (2006). For the alternate raw consensus data, we simply recoded Likert responses 1 through 3 to be 4. Likewise, we recoded Likert ratings of 7 through 9 to be 6. We then calculated the squared mean deviations from the sample mean based on the original data (the data that do not simulate avoiding extreme responses).

For the alternative standardized consensus scale, we took the recoded raw Likert items used to simulate avoiding extreme responses and subjected them to the within-person z transformation. When we calculated the squared deviation from the sample z transformed mean, we used the mean of the original z transformed data (i.e., the mean from the raw data that does not simulate the avoiding extreme responding strategy).

The dichotomous consensus scoring is not affected by avoiding extreme responses. The dichotomous consensus scoring treats all raw Likert scale ratings from 1 to 4 as identical, ineffective responses. Our simulation of the avoiding extreme response faking strategy recoded responses of 1 to 3 as 4 and thus did not change the ineffective classification of the responses for the dichotomous consensus scoring strategy. Likewise, the dichotomous consensus scoring strategy treats all raw Likert scale ratings from 5 to 9 as effective. Our simulation of the avoiding extreme responses faking strategy recoded ratings of 7 to 9 as 6 and thus did not change the effective classification of the responses for the dichotomy scoring strategy. Thus, we did not create an alternate form of the dichotomous consensus scale for the faking strategy because the faking strategy has no effect on the scale.

Some of the hypotheses are at the item level and some are at the scale level. To examine scale validity, we created nine composite scale scores. The first three composites are the sum of the 136 response option scores for the three scoring methods. The second six composites reflect our interest in determining the validity of scale composites where mid-range response options (i.e., response options with means near 5, the midpoint of the nine-point Likert scale), hypothesized to have weak validity, are discarded. We defined mid-range response options in two ways: a lenient exclusion of mid-range items and a stringent exclusion of mid-range items. For both the lenient and stringent exclusion rules, we rounded the item means to the nearest integer so that all item means were an integer from one to nine. This rounding is not needed for item exclusion but facilitates clear presentation of results. For the three lenient SJT score composites, we excluded items with rounded means of 4 through 6. These three composites were based on the 56 items with rounded means of 2 and 3 or 7 and 8 (no item means rounded to either 1 or 9) and corresponded to the raw consensus, standardized consensus, and dichotomized consensus scoring strategies. For the three stringent SJT composites, we excluded rounded means from 3 through 7. These scales consisted of 21 items with rounded means of 2 and 8 and corresponded to the raw consensus, standardized consensus, and dichotomized consensus scoring strategies. Table 1 summaries the nine item composite scale scores.

If our hypothesis concerning elevation and scatter is correct, the mean item validities should be lowest for the raw consensus scoring methods and higher for the standardized consensus scoring method and the dichotomous scoring method. The later scoring methods should have higher mean item validity because they control for elevation and scatter. Also, if our hypothesis concerning mid-range items is correct, within each scoring method, the mean item validities based on all 136 items should be the lowest, the mean validity of the items kept after applying the lenient item exclusion rule should be the higher, and the mean validity of the items kept after applying the stringent item exclusion rule should be the highest.

If Hypothesis 1 is correct, we would expect the scale validities to be lowest for the raw consensus data and higher for the standardized consensus and the dichotomous consensus methods. Predicting scale validity based on Hypothesis 1 is a tricky business. Although one can predict mean item validities based on Hypothesis 1, scale validity is a function of the item validity, the number of items, and the criterion-relevant redundancy of the items. Thus, it is possible for a scale composed of a large number of items with lower mean validity to have larger validity than a scale composed of a small number of items with higher mean validity. Because of this uncertainty, we offered no hypothesis regarding scale validity as a function of dropping mid-range items.

Extreme response score. We calculated an extreme response score by counting the number of raw Likert responses of 1 and 9 and summing the two counts. The extreme response

score is not the subject of any hypothesis but some hypotheses are based on the assumption that there is a Black-White mean difference in extreme responding so this score is calculated to check that assumption.

Biodata quitting criterion. Constructs associated with quitting behavior (Snell, Fluckinger & McDaniel, 2009) were measured using a scale consisting of the sum of 24 self-report biodata items. This is a self-report criterion. It was developed as a surrogate for the ideal criterion of military attrition, we refer to correlations with this criterion as criterion-related validities. We recognize that a measure of actual military attrition or a measure of job performance would have been a more ideal criterion.

Sample.

The sample consisted of 702 individuals. Some were college students who voluntarily participated for course credit. Others respondents were drawn from the community and participated for cash or gift cards. Of the 702 respondents, 510 were White and 111 were Black. Our race difference analyses are limited to the White and Black respondents. Respondents completed the survey package anonymously. This research was reviewed and approved by an Army Research Institute human research committee.

Results

Results relevant to Hypothesis 1

Hypothesis 1 holds that SJT scoring methods that control for elevation and scatter will yield higher response option validities than methods that do not. This hypothesis is at the item level of analysis. Table 2 shows mean item validity by scoring method. Hypothesis 1 is best addressed by examining the mean item validity for the 136 items scored using one of three scoring strategies. The raw consensus scoring method does not control for elevation and scatter but the standardized consensus and the dichotomous consensus do control for elevation and scatter. The raw consensus scoring mean item validity was .03, the standardized consensus scoring mean item validity was more than three times higher (.11), and the dichotomous consensus mean item validity was more than three times higher (.08). These results support Hypothesis 1.

Results relevant to Hypothesis 2

Hypothesis 2 holds that SJT scoring methods that control for elevation and scatter will yield lower Black-White mean differences than methods that do not. This hypothesis is at the scale level of analysis and rests on the assumption that there are Black-White mean differences in the use of extreme responses. Blacks had a mean extreme response count of 51.8 and Whites had an extreme response count of 41.3. These counts are based on 136 items and thus Blacks used an extreme rating, on average, for 38% of the items compared to 30% for Whites. The mean difference is statistically significant ($p < .0001$) and the standardized mean difference is .40. Thus the assumption underlying the hypothesis is supported.

Table 3 addresses the Black-White mean differences for each SJT scale. Of relevance to Hypothesis 2 are the standardized mean differences in the column labeled "Black-White *d*." The raw consensus scale yields a *d* of .42 favoring Whites. The two scoring methods that control for elevation and scatter show smaller *d*s. The standardized consensus scale yields a *d* of .30 and the dichotomous consensus scale yields a *d* of .18. These results support hypothesis 2. We do not

have an explanation for why the dichotomous consensus *d* is so much smaller than the standardized consensus.

Results relevant to Hypothesis 3

Hypothesis 3 holds that SJT scoring methods that control for elevation and scatter will reduce score elevation associated with a faking strategy of avoiding extreme responses. This hypothesis is at the scale level of analysis. Table 4 addresses this hypothesis. Cullen et al. (2006) found that the faking strategy raised scores by 1.57 standard deviations. For our data, the faking strategy raised scores by 2.20 standard deviations. Consistent with Hypothesis 3, when elevation and scatter are controlled, the faking strategy is ineffective. For the standardized consensus scale, the faking strategy *lowered* scores by .59 of a standard deviation. As noted earlier, the faking strategy has no impact on the calculation of the dichotomous consensus score and thus, there is no change.

Results relevant to Hypothesis 4

Hypothesis 4 holds that SJTs with consensus scoring based on raw (i.e., unadjusted) Likert ratings will show smaller Black-White differences in SJT scale scores for those scales completed with an avoiding extreme responses strategy. This hypothesis is at the scale level of analysis. Table 3 provides results relevant to this hypothesis. Table 3 shows that when the avoiding extreme ratings faking strategy is simulated in our data, the Black-White mean difference in scales score drops from .42 to .23, thus supporting Hypothesis 4. Although not relevant to the hypothesis, it is of interest to know that the faking strategy increased the Black-White *d* for the standardized consensus score. As noted earlier, the avoiding extreme ratings faking strategy does not alter the dichotomous consensus scale and thus can not effect a change in the Black-White *d*.

Results relevant to Hypothesis 5

Hypothesis 5 holds that SJT scale scores based on the avoiding extreme responses faking strategy will have large magnitude correlations with SJT scales that control for scatter. Table 5 presents intercorrelations for the three SJT scales and their counterparts that include the avoiding extreme ratings faking strategy. Note that the dichotomous consensus is not affected by the avoiding extreme ratings faking strategy and thus only one form of this scale is presented. Consistent with the hypothesis, the raw consensus scale based on item ratings adjusted to simulate the avoiding extreme ratings faking strategy is correlated .94 with the standardized consensus scale and .93 with the dichotomous consensus scale. The standardized consensus score based on item ratings adjusted to simulate the avoiding extreme ratings faking strategy is correlated .85 with both the standardized consensus scale and the dichotomous consensus scale. Thus, hypothesis 5 is supported.

Results relevant to Hypothesis 6

Hypothesis 6 holds that a SJT raw consensus scale score based on the avoiding extreme responses faking strategy will have larger criterion-related validity than SJT raw consensus scale scores based on raw (i.e., unadjusted) Likert scales. Table 2 (last column) shows that the correlation between the raw consensus scale and the criterion measure is .06. Although not presented in any table, the corresponding correlation for the race consensus scale based on the avoiding extreme response faking strategy is .35. Thus, Hypothesis 6 is supported.

Results relevant to Hypothesis 7

Hypothesis 7 holds that there will be a U-shaped relationship between response option respondent means and item criterion-related validity, such that items with low means and high means will be more valid than items with means near the mid-point of the Likert scale. This hypothesis is at the item level of analysis. This hypothesis incorporates the assumption that response options with means near the mid-point of the scale will have larger variances. Thus, we are assuming an inverted U-shaped relationship between response option means and variances. To test this assumption, we made the response option variance the dependent variable and used the response mean and its square as independent variables. The squared term is needed to evaluate whether the relationship varies from linearity. Because item responses are nested within stems, we conducted a random coefficient model analysis to see if the intercepts for the response option variance were fixed or random (Bliese, 2002). A fixed model was supported indicating that the variance of the item response did not differ significantly as a function of the stem associated with the response option. This permitted the test of Hypothesis 7 using ordinary least squares (OLS) regression. The *R* for the prediction of the item variance from the mean was .17 ($p < .05$) which yielded a multiple *R* of .81 (p for increment < .01; p for two predictor model < .01). Figure 1 shows a plot of the relationship and the average variance associated with each response option mean. There is a clear inverted U-shape relationship between response option means and variances. This inverted U-shape relationship is also seen in the second and third column of Figure 1 where one sees how the item variance varies with the item means in a nonlinear fashion.

Hypothesis 7 argued that there will be a U-shaped relationship between response option means and item criterion-related validity, such that items with low means and high means will be more valid than items with means near the mid-point of the Likert scale. We used a procedure similar to that described above to test this hypothesis. That is, we made the response options' validity the dependent variable and used the response mean and its square as independent variables. The squared term was needed to evaluate whether the relationship varies from linearity. Because item responses are nested within stems, we conducted a random coefficient model to see if the intercepts for the response option validities were fixed or random. A fixed model would fit the data if the mean item validities within a stem were the same across the twenty stems. A random model would fit the data if the mean item validities differed across the twenty stems.

We first examined the raw consensus scoring method. Following the procedure described by Bliese (2002), we determined that the random intercept model best fit the data (-2 log likelihood difference = 4.4; $p < .05$). Thus, we examined the hypothesized U-shape relationship using random coefficient modeling. The addition of the item mean and its square to the random intercept model increased the variance explained by 54% ($p < .01$). Figure 2 shows a clear U-shaped relationship. This U-shaped relationship is also seen in the second and third column of Figure 2 where one sees how the item validity varies with the item means in a nonlinear fashion. In summary, Hypothesis 3 using the raw consensus scoring method was supported.

We next examined the standardized consensus scoring method. Following the methods described by Bliese (2002), we determined that the fixed effect model best fit the data (-2 log likelihood difference = 2.5; $p > .05$). Thus, we examined the hypothesized U-shape relationship using OLS regression. The model *R* for the item mean as a predictor of item validity was .15 and was not statistically significant. The addition of item mean squared as a second predictor raised the multiple *R* to .52. Both the two variable model and the *R* increment from the one predictor model to the two predictor model were statistically significant ($p < .01$). Figure 3 shows a clear U shaped relationship. This U-shaped relationship is also seen in the second and third column of

Figure 3 where one sees how the item validity varies with the item means in a nonlinear fashion. In summary, Hypothesis 3 using the standardized consensus scoring was supported.

Finally, we examined the dichotomized consensus scoring method. Following the methods described by Bliese (2002), we determined that the fixed effect model best fit the data (-2 log likelihood difference = 0.5; $p > .05$). Thus, we examined the hypothesized U-shape relationship using OLS regression. The model *R* for the item mean as a predictor of item validity was .30 ($p < .01$). The addition of item mean squared as a second predictor raised the multiple *R* to .59. The two variable model and the *R* increment from the one predictor model to the two predictor model were statistically significant ($p < .01$). Figure 4 shows a clear U-shaped relationship. This U-shaped relationship is also seen in the second and third column of Figure 4 where one sees how the item validity varies with the item means in a nonlinear fashion. In summary, Hypothesis 7 using the standardized consensus scoring received support.

Discussion

Raw consensus scoring is likely the most common method of scoring SJTs. This research suggests that it is a substantially inferior scoring method when compared to a standardized consensus scoring and dichotomized consensus scoring. The mean item validity (see Table 2) for the raw consensus scoring was .03 as compared to .11 for standardized consensus scoring and .08 for dichotomized consensus scoring. The differences in scale validity are also dramatic. For the scales based on 136 items, the raw consensus scale validity was .06, compared to .33 for the standardized consensus scoring, and .34 for the dichotomized consensus scoring. We note that the standardized consensus scale and the dichotomized consensus scale also had lower Black-White mean differences. It is a rare situation where one can substantially improve validity while substantially reducing mean race differences. We attribute this situation to the removal of criterion-irrelevant individual differences in scale use (i.e., elevation and scatter) resulting in higher validity. The removal of the criterion-irrelevant differences in scale use also substantially reduced the Black-White mean difference in extreme responding (for the standardized consensus) or entirely removed the Black-White mean difference (for the dichotomized consensus).

Deletion of mid-range items also improves validity regardless of which consensus method is used. For the raw consensus scoring, the mean item validity for all items is .03 which rises to .08 with lenient exclusion of mid-range items, and to .12 with stringent exclusion of mid-range items. Similar effects are seen for standardized consensus items (mean item validities = .11, .14, and .16) and for dichotomized consensus items (mean item validities = .08, .13, and .14). The scale validity results are less clear. Although dropping mid-range items improves mean item validity, it also reduces the number of items available to build a scale. For raw consensus scoring, dropping mid-range items has desirable scale validity effects (scale validities = .06, .12, .17). However, this clear pattern is not found for standardized consensus scoring scales (scale validities = .33, .30, .23) and dichotomized consensus scoring (scale validities = .34, .33, and .31). One interpretation is that dropping mid-range items is most effective when one fails to control for elevation and range. Note in Figure 2 how the mean of the mid-range items for raw consensus scoring is negative (mean item validity = -.05 and -.01). However, scoring methods that control for elevation and range (Figures 3 and 4) have positive mean item validities even for the mid-range mean items. Thus, when item validities are negative for mid-range items, it is useful to drop them. However when the item validities are positive, even though

small, it is best to retain them if validity is the sole testing issue. Often, however, there are other considerations, such as the need to reduce testing time by shortening the measures. When using the standardized consensus scoring, we obtained a scale validity of .33 with 136 items and a scale validity of .30 with 56 items. For the dichotomous consensus method, we obtained a scale validity of .34 with 136 items and a scale validity of .33 with 56 items. In many settings, it may be worthwhile to lower validity by .01 or .02 if one can decrease the number of items by 60 percent (136 items to 56 items).

The Cullen et al. (2006) faking strategy of avoiding extreme responses is very effective (d = 2.2) in raising scores for the raw consensus scale. Whereas, we believe that this is one of the most common ways of scoring SJTs, this is a serious issue for those who use this scoring method. For standardized consensus scales, the faking strategy is counterproductive, *reducing* scores (d = -.59). The faking strategy has no effect on the dichotomous consensus scale. Although intended to be a faking strategy and not a scoring method, if it were used as a scoring (e.g., scale construction) method, it yields scale scores that are highly correlated with the standardized consensus scale and the dichotomous consensus scale (r = .94 and .93). As a scoring strategy, it is best classified as a method that largely controls for elevation and scatter. As such, relative to a raw consensus scale, it has much higher validity and lower Black-White mean differences.

Based on this research, we offer three recommendations for improved practice in using situational judgment tests. Then we identify limitations of the research and suggestions for future research.

Our first recommendation is that when using subject matter experts to establish a scoring key, one should screen items with respect to means and variances of subject matter effectiveness ratings. Responses with means near the center of the Likert scale ("mid-range means" for example three on a five-point scale) tend to have large variances (indicative of subject matter disagreement). Items with mid-range means and high variances tend to have very low validities. Consider removing such responses from the test prior to administration. We note that screening items for subject matter disagreement is not a new suggestion in the literature and has been a recommendation for good practice since Motowidlo et al. (1990). To get means and variance on subject matter expert ratings, ideally, the ratings need to be obtained independently prior to subject matter expert discussion. This practice does not preclude subsequent discussion among subject matter experts or a scoring key based on a consensus reached in discussion. Often the cause of disagreement among subject matter experts on the effectiveness of a response can be identified in discussion and the response can be edited to make it more clearly effective or ineffective.

Second, when scoring the SJT, screen items with respect to respondent means and variances. If one has screened items with respect to subject matter expert means and variances, one has likely removed many responses that would have had low validities. We also recognize the mean of applicant ratings is often similar to the mean of expert ratings (Legree et al., 2005) However, some items may have had desirable item properties in the subject matter expert sample, but may have less desirable properties in the respondent sample. Consider an insurance company that has a policy of not rushing customers in customer service calls because management believes that continuing the call as long as the customer wishes to talk establishes rapport and permits the call center agent to market additional products to the customer. Subject matter experts who work for the company are guided by this policy and uniformly rate such behavior as highly effective. In contrast, applicants may have widely varying experiences and

opinions on the benefits and drawbacks of lengthy customer service calls. This may result in substantial variance in effectiveness ratings on responses related to talkative customers. Our data indicate that high variance and/or mid-range mean items in respondent data have low or negative validities and are best not scored. We note that our second recommendation is primarily useful for situations where one does not have the criterion data and large samples needed for empirical screening of items. If one has stable empirical estimates of the validity of items, one does not need to use respondent sample means and standard deviations to predict validity that is already known.

Third, we recommend controlling for elevation and scatter in item responses. We examined two methods, within-person *z* transformations and dichotomization. Both methods substantially improved item validity, substantially reduced Black-White mean differences, and destroyed the effectiveness of the Cullen et al. (2006) faking strategy. Other methods such as trichotomization and correcting only for scatter but not elevation could be examined in future research. Note that if one scores the SJT based on a trichotomization strategy and drops the mid-range mean items, one will end up with something approximating a dichotomous scoring procedure.

Although this research has improved our knowledge concerning scoring of SJTs and factors associated with validity and mean racial differences, this research has some limitations and we offer suggestions for future research. Our criterion was not job performance but a self-report biodata measure of quitting. For the purpose of our research, the ideal criterion would be military attrition so our biodata measure is conceptually similar but clearly not the same as military attrition. The replication of our findings with actual military attrition and measures of job performance is clearly warranted. Our finding that one can increase validity and at the same time reduce Black-White mean differences is certainly in need of replication. Additional research is also needed on the best ways of controlling for elevation and scatter. A within-person *z* transformation is the sledge hammer of controlling for elevation and scatter. Other adjustments that control for just elevation or just scatter may be useful in understanding what aspects of individual differences in rating scale use control our effects. Perhaps some but not all individual differences in scale use are criterion-irrelevant. The Black-White differences in extreme responding are worthy of increased attention. The past research literature is not particularly helpful in understanding this effect. Although not presented in this paper, our analyses to predict this race difference using cognitive ability and the Big 5 yielded very weak prediction of this effect. More research on this race difference is clearly needed. Also, research is also needed to guide decisions concerning dropping mid-range items. Dropping some items is clearly advantageous but dropping too many items can reduce validity. Strategies are needed for finding the optimal method for dropping the midrange items.

References

Bachman, J.G. & O'Malley, P.M. (1984). Yea-saying, nay-saying, and going to extremes: Black-White differences in response styles. Public Opinion Quarterly, 48, 491-509.

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

Bliese, P.D. (2002). Multilevel random coefficient modeling in organizational research using SAS and S-Plus. In F.D. Drasgow & N. Schmitt (Eds). *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis.* San Francisco: Jossey Bass.

Cronbach, L.J. & Gleser, G.C. (1953). Assessing similarities between profiles. *Psychological Bulletin, 50,* 456-473.

Cullen, M.J., Sackett, P.R., & Lievens, F.P. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14,* 142-155.

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts, *Emotional Intelligence: An International Handbook.* (pp. 155-180). Berlin, Germany: Hogrefe & Huber.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A., & Braverman, E.P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730-740.

McDaniel, M.A., Hartman, N.S., Whetzel, D.L. & Grubb. W.L., III. (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60,* 63-91.

McDaniel, M.A., Whetzel, D.L., & Nguyen, N.T (2006). *Situational judgment tests in personnel selection: A monograph for the International Personnel Management Association Assessment Council.* Alexandria, VA: International Personnel Management Assessment Council.

Moss, F.A. (1926). Do you know how to get along with people? Why some people get ahead in the world while others do not. *Scientific American, 135,* 26-27.

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.

Putka, J.D. & Waugh, G.W. (2007, April). *Gaining insight into situational judgment test functioning via spline regression.* Paper presented at the 22nd annual meeting of the Society for Industrial and Organizational Psychology Conference. New York. April.

Schmitt, N. & Chan, D. (2006). Situational judgment tests: Method or construct? In J.A. Weekely & R.E. Ployhart (Eds.) *Situational judgment tests: Theory, measurement, and application.* Mahwah, NJ: Erlbaum. 135-155.

Snell, A.F., Fluckinger, C. D., & McDaniel, M.A. (2008). *Construct-oriented development of a biodata scale of quitting behaviors*. Paper presented at the 24th annual conference of the Society of Industrial and Organizational Psychology.

Waugh, G.W. & Russell, T.L. (2006). *The effects of content and empirical parameters on the predictive validity of a situational judgment test.* Paper presented at the 21st annual convention of the Society of Industrial and Organizational Psychology. Dallas

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

Weekley, J.A., Ployhart, R.E., & Holtz, B.C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J.A. Weekely & R.E. Ployhart

(Eds.) *Situational judgment tests: Theory, measurement, and application*. Mahwah, NJ: Erlbaum. 157-182.

Table 1. Summary of nine composite scores

	All items (Number of items = 136; no items are excluded).	Lenient exclusion rule (Number of items = 56; exclude items with rounded means of 4, 5, and 6)	Stringent exclusion rule (Number of items = 21; exclude items with rounded means of 3, 4, 5, 6 and 7)
Raw consensus	Does not control for elevation and scatter. Includes mid-range items.	Does not control for elevation and scatter. Excludes mid-range items leniently.	Does not control for elevation and scatter. Excludes mid-range items stringently.
Standardized consensus	Controls for elevation and scatter. Includes mid-range items.	Controls for elevation and scatter. Excludes mid-range items leniently.	Controls for elevation and scatter. Excludes mid-range items stringently.
Dichotomous consensus	Controls for elevation and scatter. Includes mid-range items.	Controls for elevation and scatter. Excludes mid-range items leniently.	Controls for elevation and scatter. Excludes mid-range items stringently.

Table 2. Mean item level validity and scale level validity as a function of controls for elevation, scatter and whether mid-range item are included.

Response option scoring	Controls for elevation and scatter	Mid-range items included	Mean Response Option Validity	Scale level Validity
Raw consensus (136 response options)	No	Yes	.03	.06
Raw consensus (56 items that exclude items with rounded mean values between 4 and 6)	No	No	.08	.12
Raw consensus (21 items that exclude items with rounded mean values between 3 and 7)	No	No	.12	.17
Standardized consensus (136 response options)	Yes	Yes	.11	.33
Standardized consensus (56 items that exclude items with rounded mean values between 4 and 6)	Yes	No	.14	.30
Standardized consensus (21 items that exclude items with mean values between 3 and 7)	Yes	No	.16	.23
Dichotomized consensus (136 response options)	Yes	Yes	.08	.34
Dichotomized consensus (56 items that exclude items with mean values between 4 and 7)	Yes	No	.13	.33
Dichotomized consensus (21 items that exclude items with mean values between 3 and 7)	Yes	No	.14	.31

Table 3. Black-White standardized mean differences in SJT scale scores

Scoring Method	Black-White *d*	Black-White *d* with simulation of the avoiding extreme ratings faking strategy
Raw consensus	.42	.23
Standardized consensus	.30	.36
Dichotomous consensus	.18	.18

Note: A positive *d* indicates that the White mean is higher (e.g., more favorable score) than the Black mean.

Table 4. Effect of the avoiding extreme score responses on SJT consensus scales.

Scoring Method	Change score change *d*
Raw Consensus	2.20
Standardized consensus	-0.59
Dichotomous consensus	0.00

Table 5. Intercorrelations of SJT scales

	1	2	3	4
1 Raw consensus score				
2. Raw consensus score with avoiding extreme ratings faking strategy	0.57			
3. Standardized consensus score	0.62	0.94		
4. Standardized consensus score with avoiding extreme ratings faking strategy	0.74	0.91	0.85	
5. Dichotomous Consensus Scale	0.51	0.93	0.85	0.88

Figure 1. Relationship between item variance and item mean.

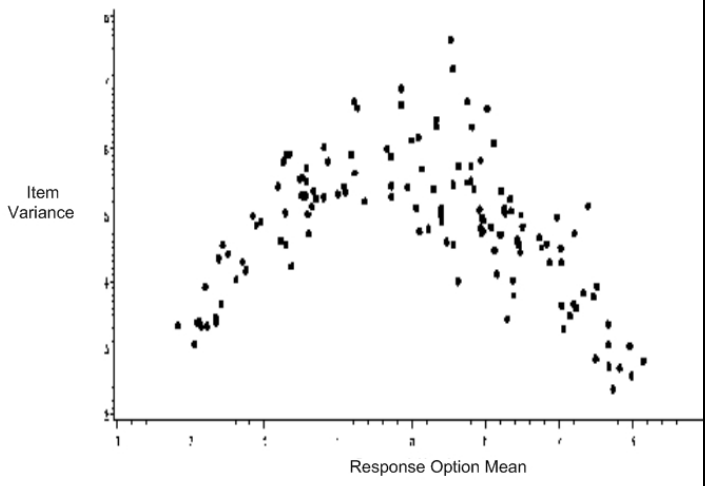
Plot of Item Variance by Item Mean	Raw item mean (rounded to an integer)	Item Variance
	2 3 4 5 6 7 8	3.6 4.9 5.5 5.7 5.2 4.1 3.0

Figure 2. Relationship between response option mean and response option validity using raw consensus scoring.

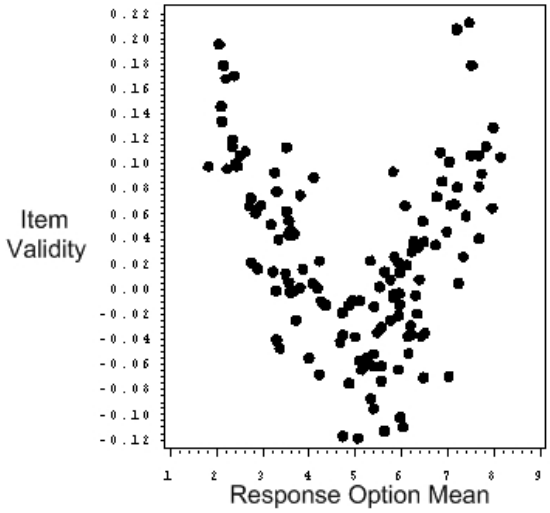
Plot of Item Validity by Item Mean	Raw item mean (rounded to an integer)	Item Validity
	2 3 4 5 6 7 8	.134 .043 .019 -.049 -.010 .071 .102

Figure 3. Relationship between response option mean and response option validity using standardized consensus scoring.

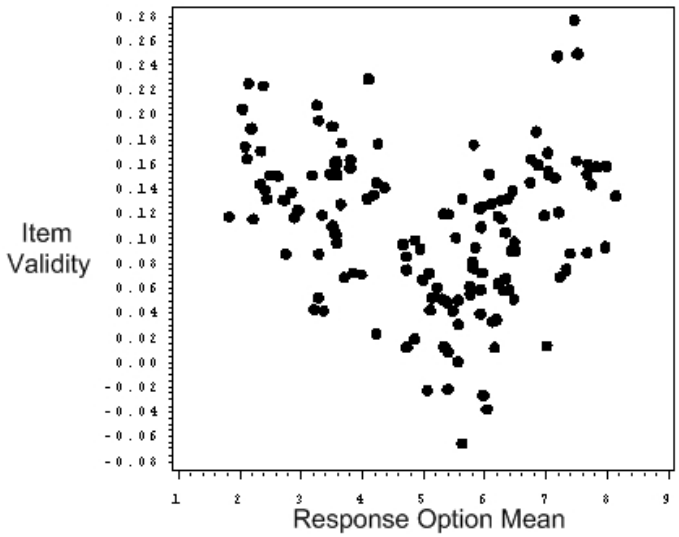
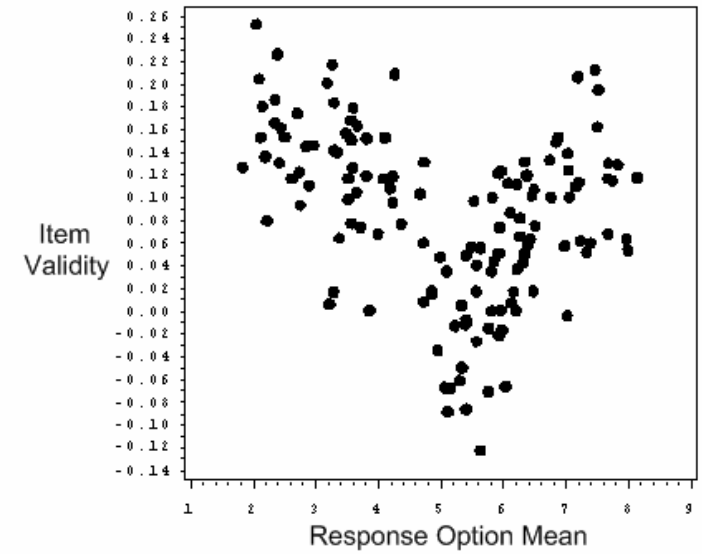
Plot of Item Validity by Item Mean	Raw item mean (rounded to an integer)	Item Validity
	2 3 4 5 6 7 8	.167 .122 .133 .054 .074 .142 .149

Figure 4. Relationship between response option mean and response option validity using dichotomized consensus scoring.

Plot of Item Validity by Item Mean	Raw item mean (rounded to an integer)	Item Validity
	2 3 4 5 6 7 8	.167 .129 .118 .002 .043 .112 .110