Journal of Applied Psychology

Toward an Understanding of Situational Judgment Item Validity and Group Differences

Michael A. McDaniel, Joseph Psotka, Peter J. Legree, Amy Powell Yost, and Jeff A. Weekley

Online First Publication, January 24, 2011. doi: 10.1037/a0021983

Journal of Applied Psychology

0021-9010/11/$12.00 DOI: 10.1037/a0021983

Toward an Understanding of Situational Judgment Item Validity and Group Differences

Michael A. McDaniel
Virginia Commonwealth University

Joseph Psotka and Peter J. Legree
U.S. Army Research Institute for the Behavioral and Social Sciences

Amy Powell Yost
Capital One

Jeff A. Weekley
Kenexa

This paper evaluates 2 adjustments to common scoring approaches for situational judgment tests (SJTs). These adjustments can result in substantial improvements to item validity, reductions in mean racial differences, and resistance to coaching designed to improve scores. The first adjustment, applicable to SJTs that use Likert scales, controls for elevation and scatter (Cronbach & Gleser, 1953). This adjustment improves item validity. Also, because there is a White–Black mean difference in the preference for extreme responses on Likert scales (Bachman & O'Malley, 1984), these adjustments substantially reduce White–Black mean score differences. Furthermore, this adjustment often eliminates the score elevation associated with the coaching strategy of avoiding extreme responses (Cullen, Sackett, & Lievens, 2006). Item validity is shown to have a *U*-shaped relationship with item means. This holds both for SJTs with Likert score response formats and for SJTs where respondents identify the best and worst response option. Given the *U*-shaped relationship, the second adjustment is to drop items with midrange item means. This permits the SJT to be shortened, sometimes dramatically, without necessarily harming validity.

Keywords: situational judgment test, extreme responding, racial differences, 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 eighty years (McDaniel, Morgeson, Finnegan, Campion & Braverman, 2001; Moss, 1926), there is very little research addressing how to best build and score SJTs (Schmitt & Chan, 2006; Weekley, Ployhart, & Holtz, 2006). In the absence of this knowledge, many approaches have evolved for developing and scoring SJTs (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; Weekley et al., 2006), but the effectiveness of these methods for maximizing criterion-related validity is largely unknown.

Michael A. McDaniel, School of Business, Virginia Commonwealth University; Joseph Psotka and Peter J. Legree, U.S. Army Research Institute for the Behavioral and Social Sciences, Arlington, Virginia; Amy Powell Yost, Capital One, Tampa, Florida; Jeff A. Weekley, Kenexa, Frisco, Texas.

This research was funded by the U. S. Army Research Institute for the Behavioral and Social Sciences through Contract W91WAW-07-C-0013, awarded to Work Skills First, Inc. The views, opinions, and/or findings contained in this article are solely those of the authors and should not be construed as an official Department of the Army or Department of Defense position, policy, or decision, unless so designated by other documentation.

Correspondence concerning this article should be addressed to Michael A. McDaniel, School of Business, Virginia Commonwealth University, 301 West Main Street, P.O. Box 844000, Richmond, VA 23284-4000. E-mail: mamcdani@vcu.edu

Unlike those in cognitive ability or job knowledge tests, response options in SJTs cannot easily be declared correct or incorrect. As such, items are typically scored with some form of consensus judgment (Legree, Psotka, Tremble, & Bourne, 2005). Typically, expert judges are asked to reach consensus concerning which responses are preferred (Weekley & Ployhart, 2006). Consensus may also be based on the responses of applicants, incumbents, supervisors of incumbents, or even customers. In such applications, the means of the respondents are considered the correct response (i.e., the test answer key).

Consensual scoring is a form of profile matching. One profile consists of the means of the items collected from some group (e.g., experts). The other profile is the item responses of a job applicant. A respondent's score on a SJT using a Likert format response scale 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 for a respondent. Scatter reflects the magnitude of a respondent's score deviations from the respondent's own mean. Legree (1995; Legree et al., 2005) suggested controlling for elevation and scatter. If one standardizes scores using a within-person z transformation, all respondents would have the same mean (0) and the same standard deviation (1) across items. This transformation removes information related to elevation and scatter from the scores, 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 investigators should consider whether elevation and scatter are important in their profile-matching application. We suggest that for SJTs, elevation and scatter primarily 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 midscale Likert ratings (e.g., preferring ratings of 1 and 9 over ratings of 3 and 7 on a 9-point Likert scale). These response tendencies can be expected to occur across items, and the difference between individuals in response tendencies can be viewed as a source of systematic error in ratings that will affect both elevation and scatter. We assert that these stable individual differences in response tendencies are primarily systematic, criterion-irrelevant noise in the ratings that 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 item validities than methods that do not.

Although seldom considered in the industrial–organizational and management literatures, there are stable White–Black differences in the use of Likert scales (Bachman & O'Malley, 1984). Blacks tend to use extreme rating points (e.g., 1 and 7 on a 7-point scale) with greater frequency than Whites on average. These differences have been found in multiple large, nationally representative samples in the United States (Bachman & O'Malley, 1984). In consensual scoring approaches used with Likert response format SJTs, extreme rating points, on average, will have larger deviations from the mean, resulting in less favorable scores. The greater tendency of Blacks to use extreme ratings will increase White–Black 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 White–Black mean differences than methods that do not.

This hypothesis is restricted to White–Black differences due to limitations of our data, but the finding may be more widely applicable to SJTs in international settings. Lau (2005) reviewed cross-cultural influences on extreme responding. However, the results of the literature are mixed concerning how countries and cultures vary in extreme responding. The applicability of our findings to settings outside the United States will require additional research.

Cullen, Sackett, and Lievens (2006) examined the coachability of two SJTs. They simulated what would happen if a respondent were coached not to endorse extreme values. This simulation was conducted using an existing data set in which the responses were made on a 7-point Likert scale by changing the 1 and 2 responses to a 3 and by changing the 6 and 7 responses to a 5. Cullen et al. discovered that scores could be improved by 1.57 standard deviations, on average, 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 coaching strategy of avoiding extreme responses.

Across respondents, the rated mean effectiveness on Likert scales varies across items. Some items have a mean indicating that the behavior described in the item, as rated by most respondents (whether applicants or subject matter experts), is an effective solution to the problem described in the scenario. Other items have means indicating that most respondents believe that the behaviors are ineffective. Items also have variance. Some items have low variance, indicating that most respondents rated the item near its mean rating. Other items have high variance, indicating that there was substantial disagreement among respondents concerning the effectiveness of the behavior described in the item. This disagreement may reflect some ambiguity in the item that requires the respondent to make inferences about the behavior described in the item 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 item "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. Others 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 behavior, the variance of the ratings reflects this disagreement. An item 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 validities may be related to item means. Two studies 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 means and item validities such that items with low or high expert rating means had the highest validities. We argue that items with means near the midpoint of the Likert scale have less informational value than items with means near either the low or high end of the Likert scale for three reasons. First, the items might be less relevant to the scenario and thus provide little information on whether the respondent can identify effective behavior. For example, with respect to the scenario concerning a miscommunication with a supervisor, a relatively irrelevant response might be "go out to dinner to try to feel better about the situation." The second reason why midrange items might have less informational value is because an item with a midrange mean (4 on a 7-point scale) is less sensitive to the assessment of a respondent's poor judgment. For example, if the answer key indicates that 4 is the best answer, a person with poor judgment can only be wrong by 3 points in that the worst response that the respondent could offer is a rating of a 1 or a 7. However, if the answer key states that 1 is the best answer, the respondent with poor judgment can be wrong by 6 points by providing a response of 7. Third, because midrange mean items tend to have substantial variance (i.e., little agreement) on the effectiveness of the item, the accuracy of the mean as an indicator of the correct answer is diminished, resulting in the scored item being less informative of the respondent's judgment (whether the respondent is an applicant or a subject matter expert). This third explanation is also relevant to SJTs with a best–worst response format. With respect to any of these three explanations, the midrange items may tend to be less valid. Thus, we offer the following hypothesis:

> *Hypothesis 4:* There will be a *U*-shaped relationship between item 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 midpoint.

This hypothesis applies to the raw consensus, standardized consensus, and dichotomous consensus scores in which the midpoint is the center of the Likert response scale. This hypothesis also applies to SJTs with the best–worst response format. For these items, a midpoint item is a scored item with a mean near the center of the distribution of item means. We note that being able to drop lower validity items may raise test validity. Dropping items also reduces the length of a test on subsequent administrations and thus reduces operational testing time.

The possibility of being able to drop items without impacting validity has benefits for applicants and for employers. Long testing times will discourage some desirable applicants from participating in testing. Likewise, testing time often requires applicants to take vacation or other leave from their current job. Shortening testing time makes it more convenient for applicants to be assessed. For employers, testing time affects the costs of the testing program. These costs include labor associated with the test administration as well as costs associated with the use of facilities. Employers can reduce their costs by reducing testing time. Alternatively, the ability to reduce testing time for one test frees up testing time that could be used for other employment tests.

Method

To test hypotheses, we analyzed four data sets that we label Study 1 through Study 4. The SJTs in Studies 1 and 2 were similar, and the SJTs in Studies 3 and 4 were similar. We first describe method information relevant to Studies 1 and 2 and then describe method information relevant to Studies 3 and 4.

Method for Study 1 and 2

The top five rows of Table 1 display method information for Studies 1 and 2. We used a Likert scale for recording respondent ratings in both studies. In Study 1, a SJT with 20 scenarios and 136 response options was used. In Study 2, the SJT had 31 scenarios and 192 response options. In each study, respondents were asked to read each scenario and rate the effectiveness of each response option on a Likert scale. Because each response option was rated, the number of items in the SJT is equal to the number of response options. Studies 1 and 2 were both concurrent studies. In Study 1 a self-report biodata measure assessing quitting tendencies (Fluckinger, Snell, & McDaniel, 2009) was used as the criterion, and in Study 2 a supervisory rating of task performance was used as the criterion.

Consensual scoring strategies were used in Studies 1 and 2. The first strategy, labeled *raw consensus,* is a common one in SJT applications. The raw consensus means (i.e., the mean ratings across respondents on each item) serve as the answer key, and a respondent's score is the squared deviation from the mean. This score was inverted, so that high scores indicated a close match with the group means (i.e., the item answer keys) 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, labeled *standardized consensus,* first requires within-person z standardization such that the mean across items for each respondent is zero with a standard deviation of one. As with the first strategy, a respondent's scale score is an aggregation of the items' squared deviations from the group mean. The scale score is 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, labeled *dichotomous consensus,* uses the raw item mean across respondents to determine if an item is effective. In Study 1, a 9-point Likert scale was used; a group mean of 5.0 or above was defined as effective, and group mean below 5.0 was defined as ineffective. In Study 2, a 6-point Likert scale was used, so that ratings of 4 to 6 were judged effective and ratings of 1 to 3 were judged ineffective. If the group mean indicated that the item was effective and the respondent indicated that the item was effective, the respondent received a score of one; otherwise, the respondent received a score of zero. Likewise, if the group mean indicated that the item was ineffective and the respondent indicated that the item was ineffective, the respondent received a score of one; otherwise, the respondent received a score of zero. This scoring method largely controls for elevation and scatter in that low ratings (1–4 for Study 1; 1–3 for Study 2) are treated identically and high ratings (5–9 for Study 1; 4–6 for Study 2) also are treated identically. The dichotomous scale does not require scale inversion in order for high scores to reflect favorable test performance.

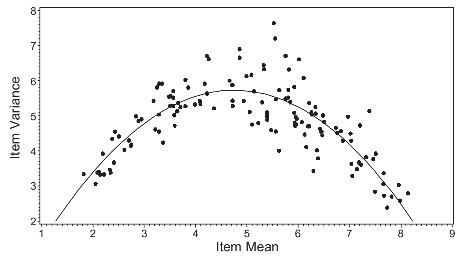
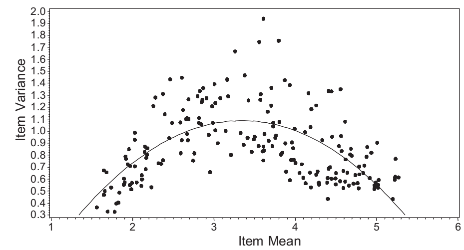
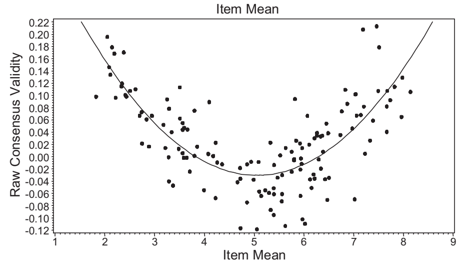
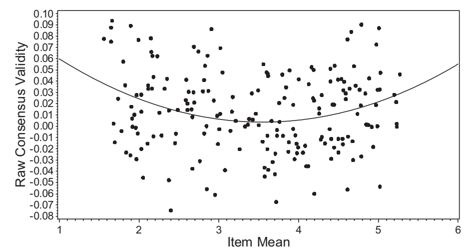
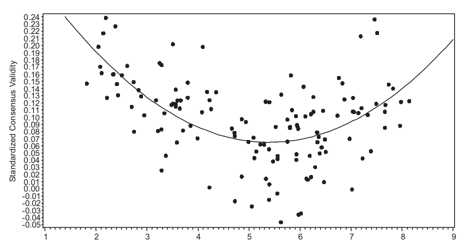
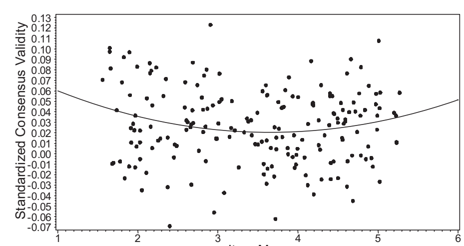
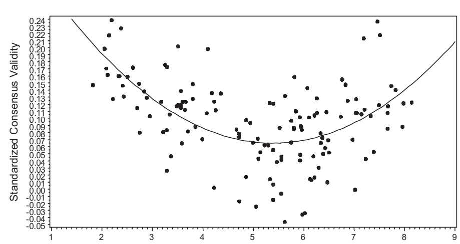
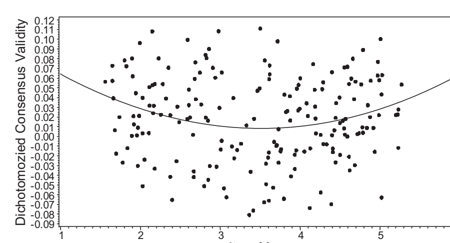
To evaluate hypotheses relevant to the coaching strategy of avoiding extreme responses, we developed alternative versions of our scales in a manner similar to that of Cullen et al. (2006). In Study 1, for the alternative raw consensus data, we recoded Likert responses 1 through 3 as 4. Likewise, we recoded Likert ratings of 7 through 9 as 6. In Study 2, we recoded Likert responses 1 and 2 as 3 and responses 5 and 6 as 4. 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 subjected the recoded raw Likert items used to simulate avoiding extreme responses 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 responses strategy).

The dichotomous consensus scoring is not affected by avoiding extreme responses. For example, in Study 1, the dichotomous consensus scoring treats all raw Likert scale ratings from 1 to 4 as identical, ineffective responses. Thus, recoding responses 1 through 3 as 4 did not change the ineffective classification of the responses. Likewise, recoding did not change the effective classification of the responses. Thus, even without the analysis of the data, it is known that the avoiding extreme responses strategy has no effect on dichotomous consensus scoring.

To examine scale validity, we created six composite scale scores in Studies 1 and 2. The first three composites are the sum of all the

Table 1
Summary of Method and Item-Level Results for Studies 1 and 2

Variable	Study 1	Study 2
Subjects and sample size	702 community and college respondents	997 incumbents
Criterion	Biodata quitting scale	Job performance
Study design	Concurrent	Concurrent
SJT response format	9-point Likert scale	6-point Likert scale
Hypotheses addressed by sample	1 through 4	1 through 4
Item variance by item mean		
Raw consensus scale Item validity by item mean		
Mean item validity	.03[a] (.08)[b]	.013[c] (.020)[d]
Standardized consensus scale Item validity by item mean		
Mean item validity	.10[a] (.13)[b]	.027 (.030)
Dichotomous consensus scale Item validity by item mean		
Mean item validity	.08[a] (.13)[b]	.018 (.026)

Note. SJT = situational judgment test.
[a] For Study 1, the differences in item mean validity for the 136 items across the three scoring methods were statistically significant ($p < .05$). For the 56 items remaining after midrange items were deleted, the raw consensus items had significantly lower mean item validity ($p < .05$) than the standardized consensus items and the dichotomous consensus items. The difference between the mean validity for the standardized and dichotomous consensus items was not statistically significant. [b] The mean item validity when excluding midrange items is in parentheses. We dropped 80 midrange items from Study 1 scales. The statistical significance of the item mean differences was calculated with an independent *t* test. One group was the 56 items that resulted from deleting the 80 midrange mean items. The second group was the 80 dropped items. The mean differences were statistically significant ($p < .05$). [c] For Study 2, the difference in item mean validity for the 192 items between the raw consensus items and the standardized consensus items was statistically significant ($p < .05$), but the difference between the mean raw consensus item validity and the mean dichotomous item validity was not. For the 87 items remaining after the midrange items were deleted, the item means for all three consensus methods were significantly different from each other ($p < .05$). [d] The mean item validity when excluding midrange items is in parentheses. We dropped 105 midrange items from Study 2 scales. The statistical significance of the item mean differences was calculated with an independent *t* test. One group was the 87 items that resulted from deleting the 105 midrange mean items. The second group was the 105 dropped items. The mean differences were statistically significant ($p < .05$) for the raw consensus scale and the dichotomous consensus scale but not for the standardized consensus scale.

item scores for the three scoring methods. The next three composites reflect our interest in determining the validity of scale composites where midrange items (i.e., in Study 1, items with means near 5, the midpoint of the 9-point Likert scale), hypothesized to have weak validity, are discarded. To define midrange items, we rounded the item means to the nearest integer, so that all item means were an integer from 1 to 9 for Study 1 and from 1 to 6 for Study 2. This rounding is not needed for item exclusion but facilitates clear presentation of results. In Study 1, we created the reduced item scale by excluding items with rounded means of 4 through 6, resulting in a scale with 56 items. In Study 2, we created the reduced item scale by excluding items with rounded means of 3 and 4, resulting in a scale with 87 items.

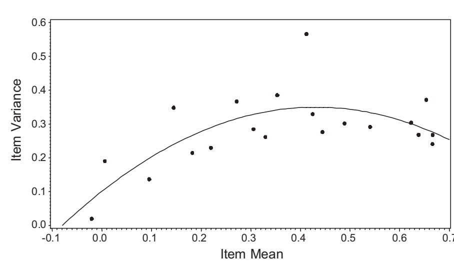
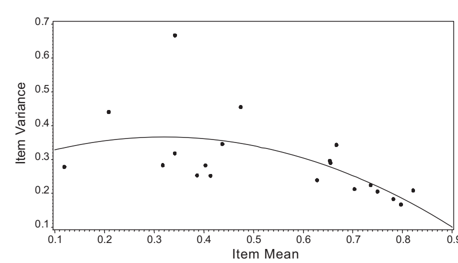
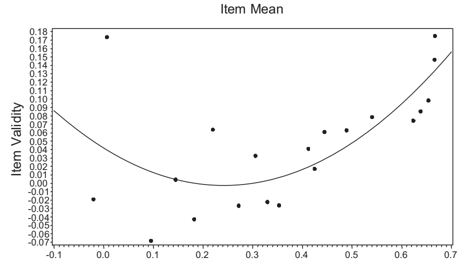
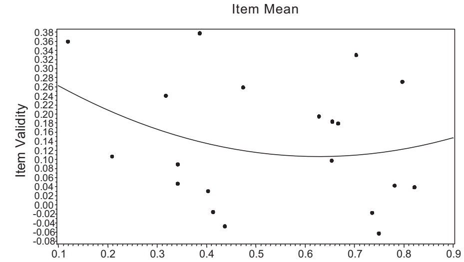
If Hypothesis 1 concerning elevation and scatter is correct, the mean item validities should be the lowest for the raw consensus scoring methods and higher for the standardized consensus scoring method and the dichotomous scoring method. The latter scoring methods should have higher mean item validity because they control for elevation and scatter. Also, if Hypothesis 4 concerning midrange items is correct, within each scoring method, the mean item validities based on all items should be the lowest and the mean validity of the items kept after discarding the midrange items should be higher.

Hypothesis 1 concerns item validity and not scale validity. However, if Hypothesis 1 is correct, we might expect the scale validities to be lowest for the raw consensus scoring method and higher for the standardized consensus and the dichotomous consensus scoring methods. However, predicting scale validity based on Hypothesis 1 is difficult. 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 higher 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 midrange items.

Method for Study 3 and 4

Table 2 summarizes method information for Studies 3 and 4. We note that in contrast to Studies 1 and 2, which were concurrent designs, Studies 3 and 4 used a predictive design. Both SJTs contained 10 scenarios, and respondents were instructed to pick two response options, one that is the best response and one that is the worst response. In the taxonomy of McDaniel, Hartman, Whetzel, and Grubb (2007), we note this is a knowledge instruction. The original Motowidlo, Dunnette, and Carter (1990) approach was a behavioral tendency instruction in that it asked for the most likely and least likely behavior. This resulted in 20 items (10 scenarios

Table 2
Summary of Method and Item-Level Results for Studies 3 and 4

Variable	Study 3	Study 4
Subjects and sample size	7,259 applicants of whom 377 had criterion data	8,667 applicants of whom 65 had criterion data
Criterion	Job performance	Job performance
Study design	Predictive	Predictive
SJT response format	Pick best and pick worst	Pick best and pick worst
Hypotheses addressed by sample	4	4
Item variance by item mean		
Item validity by item mean		
Mean item validity with 20 items (with 14 items)	.045 (.064)	.135 (.145)

Note. With reference to the last row of the table, in Studies 3 and 4, the statistical significance of the mean item validity differences was calculated with an independent t test. One group was the 14 items that resulted from deleting the six midrange mean items. The second group was the six dropped items. In Study 3, the t test of the 14 item mean to the mean of the six dropped items was statistically significant ($p < .05$), but the difference was not significant in Study 4. In Study 3, the Satterthwaite method was used because the standard deviation of the 14-item set was more than twice the standard deviation of the six-item set. SJT = situational judgment test.

each with two responses). The scoring key was provided by the consulting firm that permitted use of its data. The keyed values for each best and worst response have one of three values: −1, 0, 1. Because these two SJTs did not use Likert scales, their data are not relevant to Hypotheses 1 through 3 but provide an important test of Hypothesis 4 (i.e., we can test for a *U*-shaped relationship between item means and items validity). As for Studies 1 and 2, we created a scale based on all the items and a scale with midrange means excluded. Here, the item means and variances are the means and variances of the keyed items. For Studies 3 and 4, we created the reduced item scales by dropping the six items that had means in the middle of the distribution of the 20 item means. Thus, for both studies, the full-length scale had 20 items and the reduced-length scale had 14 items. Supervisory ratings served as the criterion in both studies.

Results

Hypotheses 1 through 3 were addressed with Studies 1 and 2. All four studies addressed Hypothesis 4.

Hypothesis 1 Results

Hypothesis 1 holds that SJT scoring methods that control for elevation and scatter will yield higher item validities than methods that do not. Results from Studies 1 and 2 address this hypothesis. This hypothesis is at the item level of analysis. Note that we are comparing mean item validities and not scale validities. Item validities have much smaller magnitudes, on average, than scale validities.

Table 1 shows mean item validity by scoring method. Hypothesis 1 is best addressed by examining the mean item validity for the scales with all items (Study 1, 136 items; Study 2, 196 items) using the three scoring strategies. The raw consensus scoring method does not control for elevation and scatter, but the standardized consensus and the dichotomous consensus methods do control for elevation and scatter. Studies 1 and 2 show consistent findings in support of Hypothesis 1. For Study 1, the raw consensus scoring mean item validity was .03, the standardized consensus scoring mean item validity was larger (.10), and the dichotomous consensus mean item validity was larger (.08). All the differences among these item means were statistically significant ($p < .05$). For Study 2, the raw consensus scoring mean item validity was .013, the standardized consensus scoring method mean item validity was larger (.027, $p < .05$), and the dichotomous consensus mean item validity was larger (.018) but not significantly different from the raw consensus mean.

Employment decisions are based on scale scores and not item scores, and so it is useful to examine scale-level validity. Higher mean item validities do not necessarily translate into higher scale validities, as a scale's validity is a function of the redundancy of the items in the prediction of the criterion. Scale validities for Studies 1 and 2 are shown in Table 3. Studies 1 and 2 show consistent findings. For Study 1, the validity for the raw consensus scale was .06 and the corresponding validity for both the standardized consensus scale and the dichotomous consensus scale was .34. Thus, the two scales that control for elevation and scatter have scale validities substantially larger than that for the raw consensus scale, which does not control for elevation and scatter. For Study 2, the validity for the raw consensus scale was .03. The validity for the standardized consensus scale was substantially larger (.16), and the validity of the dichotomous consensus scale was also larger (.10). We address the validity results for the scales with midrange item exclusion when presenting results relevant to Hypothesis 4.

Hypothesis 2 Results

Hypothesis 2 holds that SJT scoring methods that control for elevation and scatter will yield lower White–Black mean differences than methods that do not. This hypothesis is at the scale level

Table 3
Correlation Matrices for Study 1 (Above Diagonal) and Study 2 (Below Diagonal) With White–Black d and Change Score d

Variable	1	2	3	4	5	6	7	Black–White *d* (Study 1)	Black–White *d* (Study 2)	Change score *d* (Study 1)	Change score *d* (Study 2)
1. Criterion	—	.06	.15[a,b]	.34[a,c]	.32[a,c]	.34[a,d]	.33[a,d]				
2. SJT raw consensus	.03	—	.85	.62	.27	.51	.47	0.43[e]	0.56[e]	2.20[f]	1.07[f]
3. SJT raw consensus (midrange items deleted)	.03	.81	—	.76	.60	.66	.67	0.42[e]	0.36[e]	0.03	−1.08[f]
4. SJT standardized consensus	.16[a,c]	.43	.51	—	.80	.90	.87	0.29[e]	0.36[e]	−0.30[f]	−1.08[f]
5. SJT standardized consensus (midrange items deleted)	.10[a,b,c]	.21	.56	.79	—	.77	.79	0.13	0.08	0.30[f]	0.42[f]
6. SJT dichotomous consensus	.10[a]	.26	.32	.49	.45	—	.92	0.18	0.12	0.00	0.00
7. SJT dichotomous consensus (midrange items deleted)	.14[a,d]	.42	.61	.75	.76	.59	—	0.22[e]	0.07	0.00	0.00

Note. Study 1 has 702 observations. The White–Black *d* for Study 1 is based on the 621 respondents who were White or Black. All observations had criterion data. Study 2 has 997 observations. The White–Black *d* for Study 2 is based on the 772 respondents who were White or Black. All observations had criterion data. Statistical tests concerning the differences between correlations in the same sample are based on a dependent *t* test, which relies on the sample size, the mean correlation differences, and the correlation between the two scales. SJT = situational judgment test.
[a] The criterion-related validity of the SJT scale is statistically significant ($p < .05$). [b] The difference between the criterion-related validity of the SJT scale with all items and the criterion-related validity of the same scale with midrange items removed is statistically significant ($p < .05$). [c] The difference between the criterion-related validity of the raw consensus scale and the standardized consensus scale with the same number of items is statistically significant ($p < .05$). [d] The difference between the criterion-related validity of the raw consensus scale and the dichotomous consensus scale with the same number of items is statistically significant ($p < .05$). [e] The White–Black *d* is statistically significant from zero ($p < .05$). [f] The change score *d* is statistically significant from zero ($p < .05$).

of analysis. Table 3 addresses the White–Black mean difference for each SJT scale. Of relevance to Hypothesis 2 are the standardized mean differences in the columns labeled "White–Black *d*." Studies 1 and 2 show consistent results. For Study 1, the raw consensus scale yields a *d* of 0.43 favoring Whites. The two scoring methods that control for elevation and scatter show smaller *d*s. The standardized consensus scale yields a *d* of 0.29, and the dichotomous consensus scale yields a *d* of 0.18. In Study 2, the raw consensus scale yields a *d* of 0.56, the standardized consensus scale yields a *d* of 0.36, and the dichotomous scale yields a *d* of 0.12. These results support Hypothesis 2. We note that the standardized consensus and dichotomous consensus scales have substantially higher validities and substantially lower White–Black mean differences than the raw consensus scale.

Hypothesis 3 Results

The results of Studies 1 and 2 are relevant to Hypothesis 3, which holds that SJT scoring methods that control for elevation and scatter will reduce score elevation associated with a coaching strategy of avoiding extreme responses. This hypothesis is at the scale level of analysis. Table 3 (see the last two columns) displays results for this hypothesis. Cullen et al. (2006) found that the coaching strategy raised scores by 1.57 standard deviations. Studies 1 and 2 show similar findings in support of Hypothesis 3. In Study 1, the coaching strategy raised scores by 2.20 standard deviations for the raw consensus scale. However, consistent with Hypothesis 3, when elevation and scatter are controlled, the coaching strategy is ineffective for the standardized consensus scale, in which the coaching strategy lowered scores by 0.30 of a standard deviation. The coaching strategy has no impact on the calculation of the dichotomous consensus score, and, thus, there is no change. In Study 2, the coaching strategy raised scores by 1.07 standard deviations for the raw consensus scale but lowered scores for the standardized consensus score by 1.08 standard deviations.

This pattern of results is not consistent when one drops midrange items. For the raw consensus scale, if one drops midrange mean items (dropping items whose means round to 4–6), the coaching strategy lowers scores by 1.08 standard deviations. For the standardized consensus scale, the coaching strategy improves scores by 0.42 standard deviations for the scale with midrange mean items dropped.

Hypothesis 4 Results

Results from all four studies are relevant to Hypothesis 4, which holds that there will be a *U*-shaped relationship between item means and item-criterion-related validities. Thus, items with low means and high means will be more valid than items with means near the midpoint of the scale for Likert responses (or near the midpoint of the scored item mean distribution for the best–worst responses). This hypothesis is at the item level of analysis and incorporates the assumption that items with means near the midpoint will have larger variances. Thus, we are assuming an inverted *U*-shaped relationship between item means and variances. To test this assumption, we made the item variance the dependent variable and used the item mean and its square as independent variables. The squared term is needed to evaluate whether the relationship varies from linearity.

Results for Studies 1 and 2 are presented in Table 1, and results for Studies 3 and 4 are displayed in Table 2. For each of the four studies, the plots of item variance by item means show clear inverted *U*-shaped relationships. The regression models for the four curves were evaluated with a two-step regression in which the item mean was entered in Step 1 and the square of the item mean was entered in Step 2. The squared term is needed to evaluate the extent of nonlinearity. In the four samples, the regression models were statistically significant ($p < .05$) and the variance increments associated with Step 2 were also statistically significant ($p < .05$). Because the items are nested within stems, we conducted random coefficient model analyses to see if the intercepts for the item variance were fixed or random (Bliese, 2002). On the basis of that test, we evaluated the models with ordinary least squares regression or random coefficient modeling, depending on whether the intercepts were fixed or random. The details of these analyses are available from the first author.

Hypothesis 4 argued that there will be a *U*-shaped relationship between item means and item validities, such that items with low means and high means will be more valid than items with means near the midpoint. We used a procedure similar to that described above to test this hypothesis. That is, we made the item validity the dependent variable and used the item mean and its square as independent variables. The *U*-shaped curves summarizing the relationship between items' means and items' validities are shown in Table 1 (for Studies 1 and 2) and Table 2 (for Studies 3 and 4). All regression models evaluating the curves were statistically significant ($p < .05$), and the incremental variance of the second step where the squared item mean was entered was statistically significant ($p < .05$; these results are available from the first author). These results show that, on average, item validities are higher for items with low or high means and lower for items with midlevel means. For example, as seen in Study 1, the mean item validity for the raw consensus scale was .03, but when the midrange mean items were dropped the mean item validity rose to .08.

With respect to scale validities (see Table 3 for Studies 1 and 2 and Table 4 for Studies 3 and 4), some increased when midrange items were dropped and some decreased when midrange items were dropped. In Study 1, the raw consensus scale, based on 136 items, yielded a validity of .06. However, when midrange items were dropped, resulting in only 56 items, the validity increased to .15. Also in Study 1, the standardized consensus scale validity was .34 for the 136-item scale and dropped to .32 for the 56-item scale.

Table 4
Correlation Matrices for Study 3 (Above Diagonal) and Study 4 (Below Diagonal)

Variable	1	2	3
1. Criterion	—	.16	.18[a]
2. SJT best–worst	.44	—	.87
3. SJT best–worst (extreme mean items removed)	.40[a]	.92	—

Note. SJT = situational judgment test.
[a] For Studies 3 and 4, all criterion-related validities for the SJT scales were statistically significant ($p < .05$). For Studies 3 and 4, the difference between the criterion-related validities of the two scales was not significant.

Not all the changes in scale validities with the deletion of midrange items were statistically significant. As we note in the introduction, the ability to drop items from a test without harming validity (or without harming validity very much) has implications for testing times. Shorter testing times can make the application process more appealing to applicants and potential applicants. Shorter testing times for employers are also beneficial. Employers can reduce costs or can use the freed-up testing time for other assessments.

Discussion

Although usage frequency data are not available, raw consensus scoring is likely the most common method of scoring SJTs with Likert scales. This research suggests that it is a substantially inferior scoring method than standardized consensus scoring and dichotomous consensus scoring. In particular, the raw consensus method yields the lowest validities and the largest White–Black mean differences, and the scores can be substantially inflated by rating responses near the midrange of the Likert scale. We speculate that a scoring method that raises validity while reducing mean race differences might have relevance to the search for alternative measures with the same or higher validity with less adverse impact. The *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978) encourages such a search for alternative measures.

Controlling for elevation and scatter by using the standardized or dichotomous consensus methods substantially increases validity over that obtained from the raw consensus method and simultaneously reduces White–Black mean differences. Meeting the goals of raising validity and reducing mean racial differences is a challenge (Sackett, Schmitt, Ellingson, & Kabin, 2001). Research in this area often examines trade-offs between validity and mean racial differences (De Corte, Lievens, & Sackett, 2008; Ployhart & Holtz, 2008; Sackett, De Corte, & Lievens, 2010). Our results differ from this trade-off literature because we both raised validity and reduced mean White–Black score differences. Although demonstrated empirically in two samples, our findings are sufficiently uncommon that additional replication is clearly warranted, particularly with applicant samples. Because the tendency toward the use of extreme responses varies across cultures and countries (Lau, 2005), the use of standardized and dichotomous consensus scoring methods should be explored with SJTs administered across countries and cultural/ethnic groups.

A reviewer suggested that Asians often refrain from using extreme responses. This assertion may or may not be correct. Bachman, O'Malley, and Freedman-Doan (2010) found that Asian Americans and White Americans have a similar frequency of extreme responses. They noted that Black Americans have the most extreme responding. Hispanic Americans have less extreme responding than Blacks but more than White and Asian Americans. The Lau (2005) review is a narrative review, and one can find citations to support most any position. For example, Lau, citing Stening and Everett (1984), reported that Filipino, Malaysian, Indonesian, and Thai managers provided extreme responses significantly more often than did U.S. and Japanese managers. We do not believe that clear conclusions can be drawn from this literature in its current state and encourage more research and a quantitative review of this literature.

In addition, we offer that the scoring methods that adjust for scatter and elevation will remove the group differences in response tendencies regardless of whether a demographic subgroup favors or avoids extreme responding. Just as the adjustments removed the White–Black extreme response differences, the same adjustments remove the response differences between groups that favor midrange item ratings and groups that do not.

A reviewer noted that the changes in scale validity and changes in White–Black score differences may be due to differences in reliability. We first consider this alternative hypothesis with respect to analyses comparing full-length scales to scales with reduced items. We find that when validities increase but mean differences decrease, it is very hard to attribute differences between full-scale and reduced-item scales (i.e., scales with midrange items excluded) to reliability differences. Consider the raw consensus scale for Study 1 (see Table 3). The White–Black mean differences for the full scales was .43 and dropped to .42 in the reduced item scale. One could argue that the change in White–Black mean differences is a result of the lower reliability of the reduced item scale. However, the validity for the full scale was .06, and the validity for the reduced item scale was .15. One would not expect the validity to more than double when the reliability of the scale was reduced. The change in reliability hypothesis also lacks explanatory value when one considers validity and mean racial differences for the three versions of the full item scales. Although the number of items in each scale is constant, the differences across the three scales concerning whether and how item ratings are transformed might affect reliability. Again, consider Study 1 (see Table 3). The White–Black d for the raw consensus scale was 0.43, compared to 0.29 for the standardized consensus and 0.18 for the dichotomous consensus scale. One might argue that we have damaged the reliability of the measure when calculating the standardized consensus and dichotomous consensus scales, and it is the lower reliability that is responsible for the reductions in White–Black mean differences. However, the damaged scales should also have lower validity. That is not the case. The raw consensus scale has a validity of .06, compared to .34 for the standardized consensus and for the dichotomous scale.

Still, the reliability hypothesis is an explanation for the results when the White–Black mean differences and the validities vary in the same direction. This happens for some of our data (see Table 3, Study 2 results for the standardized consensus scale and the reduced item standardized consensus scale). Thus, it would be informative to correct the White–Black d and the scale validities for measurement error and then examine the pattern of differences.

Unfortunately, as noted by several scholars, SJTs and, often, SJT items are construct heterogeneous (Clause, Mullins, Nee, Pulakos, & Schmitt, 1998; McDaniel et al., 2007; McDaniel & Whetzel, 2005). This makes coefficient alpha an inappropriate statistic for assessing reliability. The data sets in this study lack test–retest reliabilities, which would be an appropriate reliability for SJTs. Thus, this study cannot evaluate whether the validity and subgroup differences may be, at least in part, due to scale reliability differences. We encourage research to address this alternative hypothesis to some of our findings.

We also recommend a closer examination of controlling for elevation and scatter. A within-person z transformation is the sledgehammer method 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. Not all factors that influence elevation and scatter may be criterion irrelevant. For example, very low scatter across items (i.e., the respondent gives the same Likert ratings to almost every item) may indicate reading comprehension problems or some decrement in judgment. The effects of controlling elevation and scatter on the constructs assessed by the SJT also warrant additional investigation.

We found that controlling for elevation and scatter can nullify a coaching strategy of using midrange responses on Likert format SJTs. Both this study and the previous work of Cullen et al. (2006) have been based on simulations. This coaching strategy and methods of minimizing or eliminating its impact should be examined in instructed faking studies.

Deletion of midrange items also tends to improve mean item validity regardless of which consensus method is used. This finding also holds for SJTs with best–worst response formats. However, scale validities do not necessarily increase as the number of midrange items decreases. This occurs because the scale validities are a function of the mean item validities, and the redundancy of the criterion overlap with the items. One can obtain higher validity with a scale with lower mean item validity with more items than with a scale with higher mean item validity but fewer items. Research is needed to identify strategies to capitalize on the relationship between item means and item validity in order to increase scale validity.

Although the items with the high or low item means in our two data sets had the highest validity, in operational selection settings, these items may be the most easily faked. Thus, it is important to replicate these findings in applicant samples. Instructed faking studies could also be used to determine whether the items' means predict the extent to which the item responses can be improved through coaching.

Research is needed to examine how test content and constructs being assessed might change as one drops midrange mean items. If one has a lengthy and very homogeneous vocabulary test and one drops items, both the original test and the reduced item test still measure vocabulary. However, SJTs are construct heterogeneous at both the item and the scale level (McDaniel & Whetzel, 2005). If one has a 136-item SJT test and drops 105 items, it is not necessarily true that the full-length and reduced item tests measure the same content and constructs.

Science progresses in many ways. A key way is through replication. Studies 1 and 2 were consistent in showing that controlling for elevation and scatter simultaneously raises validity and lowers White–Black mean differences. This is an uncommon finding that should be replicated, particularly in applicant samples and in cross-cultural samples. Studies 1 through 4 were consistent in finding a *U*-shaped relationship between items means and item validity. We note this finding permits the deletion of items without loss (or much loss) of validity. Replication in more applicant samples with both Likert response formats and best–worst response formats is encouraged.

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. doi:10.1086/268845

Bachman, J. G., O'Malley, P. M., & Freedman-Doan, P. (2010). *Response styles revisited: Racial/ethnic and gender differences in extreme responding* (Monitoring the Future Occasional Paper No. 72). Ann Arbor, MI: Institute for Social Research. Available from http://www.monitoringthefuture.org/

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. doi:10.1111/j.1468-2389.2006.00345.x

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* (pp. 401–445). San Francisco, CA: Jossey-Bass.

Clause, C. S., Mullins, M. E., Nee, M. T., Pulakos, E. D., & Schmitt, N. (1998). Parallel test form development: A procedure for alternative predictors and an example. *Personnel Psychology, 51,* 193–208.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarities between profiles. *Psychological Bulletin, 50,* 456–473. doi:10.1037/h0057173

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. doi:10.1111/j.1468-2389.2006.00340.x

De Corte, W., Lievens, F., & Sackett, P. R. (2008). Validity and adverse impact potential of predictor composite formation. *International Journal of Selection and Assessment, 16,* 183–194. doi:10.1111/j.1468-2389.2008.00423.x

Equal Employment Opportunity Commission. (1978, August 25). Uniform guidelines on employee selection procedures. *Federal Register, 43,* 38290–393150.

Fluckinger, C. D., Snell, A. F., & McDaniel, M. A. (2009, April). *Construct-oriented development of a biodata scale of quitting behaviors.* Paper presented at the meeting of the Society of Industrial and Organizational Psychology, New Orleans, LA.

Lau, M. Y. (2005). *Extreme response style: A review of the evidence, assessment and potential solutions to the problem.* Columbia University.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a Likert-based testing procedure. *Intelligence, 21,* 247–266. doi:10.1016/0160-2896(95)90016-0

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: An international handbook.* (pp. 155–179). Berlin, Germany: Hogrefe & Huber.

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. doi:10.1111/j.1744-6570.2007.00065.x

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. doi:10.1037/0021-9010.86.4.730

McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33,* 515–525. doi:10.1016/j.intell.2005.02.001

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., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640–647. doi:10.1037/0021-9010.75.6.640

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61,* 153–172. doi:10.1111/j.1744-6570.2008.00109.x

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

Sackett, P. R., De Corte, W., & Lievens, F. (2010). Decision aids for addressing the validity–adverse impact trade-off. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 453–472). New York, NY: Routledge/Taylor.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56,* 302–318. doi:10.1037/0003-066X.56.4.302

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

Stening, B. W., & Everett, J. E. (1984). Response style in a cross-cultural managerial study. *Journal of Social Psychology, 122,* 151–156. doi:10.1080/00224545.1984.9713475

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

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. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 157–182). Mahwah, NJ: Erlbaum.

Received August 17, 2009
Revision received October 5, 2010
Accepted October 5, 2010 ■