Controlling for elevation and scatter in situational judgment test scoring: A replication

Michael A. McDaniel
Virginia Commonwealth University

Jeff A. Weekley
Kenexa

Paper presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology. San Diego. April, 2012.

Abstract

This paper sought to replicate McDaniel, Psotka, Legree, Yost, and Weekley (2011). That research showed that situational judgment test scoring methods that control for elevation and scatter yield larger validities than those that do not. McDaniel et al. also showed that dropping mid-range items (items with mean scores near the midpoint of a Likert scale) resulted in higher validity. The current study replicated the results of the McDaniel et al. (2011) study, although the magnitude of the validities was low.

Situational judgment tests (SJTs) present job applicants with scenarios describing challenges faced at work. Applicants are also presented with possible response options to each situation and asked to evaluate each option for either effectiveness or likelihood of performing each action. Although SJTs are frequently used, there is relatively little research addressing best practices in building and scoring SJTs (Schmitt & Chan, 2006; Weekley, Ployhart, & Holtz, 2006).

McDaniel, Psotka, Legree, Yost, and Weekley (2011) evaluated two adjustments to common scoring approaches for SJTs. In multiple samples, McDaniel et al. demonstrated that these scoring adjustments resulted in increased levels of validity and reduced White-Black mean differences. These results may be understood by considering how SJTs are scored. In contrast to cognitive ability or job knowledge tests, SJT response options cannot easily be declared correct or incorrect. As a result, SJT items are most frequently scored using consensus judgment (Legree, Psotka, Tremble, & Bourne, 2005), typically that of incumbents or their supervisors (Weekley & Ployhart, 2006). Consensus may also be based on the means of effectiveness ratings provided by subject matter experts.

The first scoring adjustment affects the elevation and scatter of the respondents' Likert ratings of the response options. Thus, this adjustment only applies to SJTs in which the respondent evaluates response options using a Likert scale. Elevation and scatter can best be understood in the context of profile matching (Cronbach & Gleser, 1953). The primary profile is the answer key. Typically, the key is the mean effectiveness rating of the item responses collected from some group (e.g., experts). Thus, if the SJT has 10 scenarios each with five responses, the primary profile is a vector of means of the 50 items (10 scenarios times 5 responses) . The remaining profiles are the Likert item responses of applicants. The match

between an applicant's ratings and the answer key yields the applicant's score on the test. Thus, an applicant whose Likert ratings closely match the mean ratings used as the answer key will have a favorable test score.

Cronbach and Gleser (1953) discussed profile matching with respect to elevation, scatter, and shape. Elevation is defined as the mean of the response ratings across items for a respondent. Thus, an applicant who rates most response options as effective (ratings of 6 or 7 on a 7-point Likert scale) would have a high elevation score. An applicant who rates most response options as ineffective (ratings of 1 or 2 on a 7-point scale) would have a low elevation score. Scatter reflects the variance around the respondent's own mean. Applicants whose ratings cluster tightly in small range of the rating scale (e.g., an applicant who only gives ratings of 4 or 5 on a 7-point scale) have less scatter than an applicant whose ratings span the full range of the rating scale. The final characteristic of a profile is shape, which expresses the pattern of the profile controlling for elevation and scatter.

To illustrate these concepts, consider Figures 1a, 1b, and 1c. Figure 1a displays the profile of the answer key for a five item SJT. Figure 1b displays the profile of the answer key plus the profiles of the responses of two applicants (Manny and Moe). The elevation of Manny's profile is higher than the elevation of Moe's profile. Manny, whose mean rating across the five items is 4.4, always rates the response options one point higher than Moe, whose mean rating across the five items is 3.4). However, the shape and the scatter of the profiles for Manny and Moe are identical. The identical shape of Manny and Moe's profiles is evident by observing the pattern of the ratings across the items. The scatter can be expressed as the standard deviation of the ratings across the five response options. For both Manny and Moe, the standard deviation is

1.14. Figure 1c is the same as Figure 1b expect that the profile of a third applicant, Jack, has been added. Jack uses a wider range of the 7-point rating scale than either Manny or Moe and thus has more scatter (the standard deviation of Moe's five ratings is 2.41). The shape of Jacks' profile also differs from Manny and Moe's profile. However, the mean of Manny's ratings and the mean of Jack's ratings are identical (means = 4.4). Thus, Manny and Jack have identical elevation.

Legree (1995; Legree et al., 2005) argued that consensus scoring will be more accurate when controlling for elevation and scatter. One way of achieving this is to perform a within-person *z*-score transformation such that all applicants will have the same mean (0) and the same standard deviation (1) across items. Thus, all applicants would have identical elevation (identical mean ratings across items) and scatter (identical variance across items). The remaining score information in the transformed scores is shape.

McDaniel et al. (2011) suggested that, for SJTs, elevation and scatter primarily reflect response tendencies. Examples of response tendencies that effect elevation are preferences for the high end of the ratings scale (i.e., rating most response options as effective), or preferences for the low end of the ratings scale (i.e., rating most response options as ineffective). Examples of response tendencies that affect scatter are preferences for extreme ratings (i.e., rating many response options as either 1 or 7 on a 7-point Likert scale). This would result in high rating scatter. Low scatter would result from a preference for a more limited range of the Likert scale (e.g., preferring ratings of 4 or 5 on a 7-point Likert scale). Consistent with Legree (1995; Legree et al., 2005), McDaniel et al. hypothesized that these response tendencies are best viewed as a source of systematic error in ratings, which is criterion irrelevant.

Controlling for elevation and scatter would remove this criterion-irrelevant systematic error and improve the validity of the items.

McDaniel et al. (2011) also noted that there are stable White–Black mean differences in the preference for extreme responses on Likert scales (Bachman & O'Malley, 1984). On average, Blacks tend to endorse extreme rating points (e.g., 1 and 7 on a 7-point Likert scale) more often than Whites. This finding has been replicated in several large, nationally representative samples in the United States (Bachman & O'Malley, 1984). When a SJT is scored consensually, the scoring key for an item is seldom an extreme response (e.g., for the mean of a 7-point scale to equal 7, all judges would have rated the option 7; this is unlikely). As such, the preference for extreme responses among Blacks will, on average, result in lower mean scores for Blacks relative to Whites. After correcting for scatter, the mean SJT scores between Blacks and Whites will become smaller, on average. In brief, correcting SJT responses for elevation and scatter may increase validity and simultaneously reduce Black-White mean test score differences. For a very long time, higher validity with lower Black-White mean differences has been the "holy grail" of personnel selection.

The second scoring adjustment offered by McDaniel et al. (2011) relates to the relationship between item validities and item means. Both Waugh and Russell (2006) and Putka and Waugh (2007) explored the relationship between consensual means of experts and item validity. They reported *U*-shaped relationships between item means and item validities. Specifically, items with low or high means had the highest validities. McDaniel et al. suggested that items with means near the midpoint of the Likert scale may have less informational value than items with means near either the high or low end of the Likert scale (see McDaniel et al. for a discussion of three reasons for why this occurs). McDaniel et al. demonstrated that dropping

items with mid-range means tends to increase or approximately maintain the validity of the test. Tests with fewer items but with similar or higher validity than longer tests are advantageous in testing situations. Shorter tests can result in a better applicant experience and shrink time to hire or permit more assessments in a fixed period of time.

McDaniel et al. (2011) demonstrated substantial improvements in SJT scoring by yielding higher validity and lower White-Black score differences. Such positive findings warrant attempted replications. The purpose of the present study was to replicate the McDaniel et al. findings with respect to validity.

Method

Measures

The SJT contained 52 stems (i.e., scenarios) with 5 responses each. Thus, there were 260 items (52 stems times 5 responses per stem). The SJT scenarios concerned challenges faced in sales occupations.

One consensus scale was based on subject matter expert (SME) means and was calculated by summing the squared difference between the each item's answer key (i.e., SME mean rating) and the respondents' rating. Because this scoring makes zero the highest possible score, the score was inverted such that higher scores reflect better performance on the SJT. We labeled this scale the SME consensus scale. Other consensus scales used the mean of the respondents as the answer key. Using the means of the respondents, we calculated a raw consensus scale, a standardized consensus scale, and a dichotomized consensus scale consistent with the McDaniel et al. (2011) methods. The raw consensus scale was calculated by summing the squared difference between the each item's mean rating across respondents (the raw

consensus answer key) and the respondent's rating. This scale was inverted such that higher scores reflected better performance. Because the Likert ratings of the respondents were not transformed for the SME consensus scale and the raw consensus scale, neither scale corrected for elevation and scatter.

Two additional scales were created to correct for elevation and scatter in the respondents' Likert ratings. To create the standardized consensus scale, we first transformed the respondent ratings using a within-person *z* transformation and then summed the squared difference between each item's mean rating across respondents (the standardized consensus scale answer key) on the transformed items and the transformed respondent's rating. As with the raw consensus score, the scale was inverted such that higher scores reflect better performance. To create the dichotomized consensus scale, respondent-derived means were dichotomized at the mid-point of the rating scale such that each response was judged either effective or ineffective. Thus, the answer key for the dichotomized consensus rating was a dichotomy (e.g., effective or ineffective). Likewise, the ratings of the respondents were dichotomized such that the response options were judged as effective or ineffective. Respondents obtained a score of 1 if their dichotomized evaluation of the response option matched the dichotomized answer key, and otherwise received a score of zero. The standardized consensus scoring completely removed the effects of elevation and scatter. The dichotomized consensus scoring largely removed the effects of elevation and scatter.

To permit an evaluation of the validities of reduced items scales, for all four consensus scales (i.e., SME, raw, standardized, dichotomized), we created a reduced length scale by dropping the 135 items with mid-range means on the Likert rating. In this study, midrange items were defined as those with a mean Likert rating between 2 and 4. This reduced the number of items in the scales from 252 to 117. This is a reduction of over 50% of the items.

To summarize, we evaluated the validity of eight SJT scales. All the scales were created using consensual scoring. Four of the scales were composed of 252 items and four of the scales contained only 117 items. Half of the scales (SME consensus and raw consensus) did not correct for elevation and scatter. The other half of the scales (standardized consensus and dichotomized consensus) did correct for elevation and scatter. Two of the scales relied on SME means as the answer key (SME consensus scale with 252 items and SME consensus scale with 117 items). The remaining six scales relied on the mean of the respondents as the answer key.

The criterion was an objective sales performance measure composed of four components: (1) percent of quota obtained; (2) year-over-year growth; (3) percent of standard activities achieved; and (4) percent of pipeline to quota. The performance criterion was provided to the researchers as a composite and thus the subcomponents of the criterion could not be considered separately. Nor could internal consistency reliability be calculated for the criterion because we did not have data on the four components that created the composite.

Sample and Design

The validity of the SJT was evaluated in a concurrent design. The respondents were 184 incumbents. Because there were only 12 Black respondents, we could not evaluate McDaniel et al.'s (2011) results with respect to reducing mean racial differences in SJT scores.

Results

Table 1 shows the validities of the various SJT scales. The first column lists the scales. Column two shows the validity of the scales based on all 252 items. In column 3, the validity of a scale with the mid-range items is presented.

Discussion

Table 1 offers several findings. We highlight three of them. First, none of the scales resulted in large magnitude validities. Second, despite the low magnitude of the validities, the study replicated McDaniel et al.'s (2011) findings by showing that scales that do not control for elevation and scatter (SME consensus and raw consensus scales) have lower validities than scales that do control for elevation and scatter (standardized consensus and dichotomized consensus). For example, in column two, although the SME consensus and raw consensus SJT scales had validities of .01 and .00, the validity for the standardized consensus scale (.03) and the dichotomized consensus scale (.06) were higher. Thus, although none of the validities were large, the validity of the SJT scales that controlled for elevation and scatter were larger than the SJT scales that did not.

Third, the study replicates McDaniel et al. in showing that one can drop a large number of mid-range items and either increase or approximately maintain validity. In column two, the SME consensus scale increased from .01 to .10, the raw consensus scale validity increased from .00 to .10, the standardized consensus validity increased from .03 to .10, and the dichotomized consensus validity increased from .06 to .07. Thus, in this replication, dropping mid-range items always increased the validity of the scale.

We explored several reasons for the low validities for all scales. First, we inspected the SJT. Nothing seemed abnormal about the test. All items focused on challenges encountered in sales work and the items and the scales had reasonable variances. Second, we considered that the SME consensus ratings were flawed, but the SME consensus scales yielded no worse validities than the raw consensus scales in which the answer key was based on the item means generated from the respondents. Third, we considered the possibility of random responders in the sample.

Because the respondents were incumbents and may have been unmotivated to complete the long test, we considered whether respondents were answering without reading the questions. We conducted a disjoint cluster analysis (Anderberg, 1973; nearest centroid sorting) to form clusters of respondents. If a respondent is the sole member of a cluster, the respondent answered like no other and is likely an outlier. Some advocate dropping such respondents. There were no one-respondent clusters.

Fourth, we considered the possibility that consensus means based on subgroups of the respondents might work better than consensus means based on all respondents. Our reasoning was that some subsets of respondents likely make better judgments than other subsets of respondents. The three largest clusters in the data set consisted of 100, 66, and 17 respondents, respectively. Keys based on the means of those respondent groups performed no better than keys based on the full sample. Fifth, we considered the possibility that the objective sales criterion was flawed. However, we had no way of evaluating the criterion because we had only a composite criterion and had no other information on the appropriateness of the criterion. Sixth, the SJT data may have been subject to substantial range restriction. However, we did not have access to the standard deviation of the scales in an applicant pool and thus could not evaluate the range restriction hypothesis. Seventh, the results could be due to sampling error. McDaniel, Hartman, Whetzel, and Grubb (2007) reported that the mean observed validity for 96 knowledge response instruction SJTs was .20. For the same size in this study, the 95% confidence interval of a correlation of .20 ranges from .06 to .34. The lower bound of the confidence interval is near the observed validities in this study. Although we considered seven possibilities for the low magnitude validities, picking among them is a speculative endeavor.

In summary, the validities of all SJT scales were low. However, the merit of the scoring strategies offered by McDaniel et al. (2011) was supported. We recommend additional replications of these scoring strategies.

References

Anderberg, M.R. (1973). *Cluster analysis for applications*. New York: Academic Press.

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

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.

McDaniel, M.A., Psotka,J., Legree, P.J., Yost, A.P., Weekley, J.A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology, 96,* 327-336.

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.

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.

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., & 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.

Table 1. Validities for eight SJT scales

SJT Key	Validity for a scale with all 252 items	Validity for a scale with 117 items (135 mid-range items were deleted)
SME Consensus	.01	.10
Raw Consensus	.00	.10
Standardized Consensus	.03	.10
Dichotomized Consensus	.06	.07

Figure 1. Illustration of SJT items as profiles.[1]

Figure 1a. The profile of the answer key

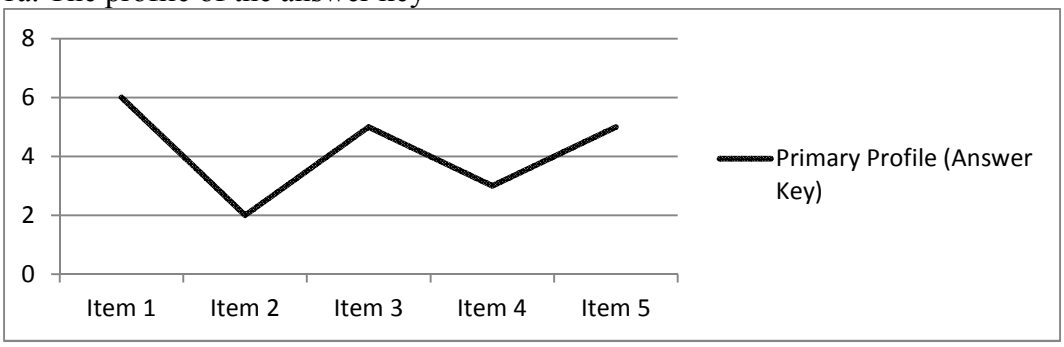


Figure 1b. The profile of the answer key with profiles for two applicants. Applicants Manny and Moe differ in elevation but have identical scatter and shape.

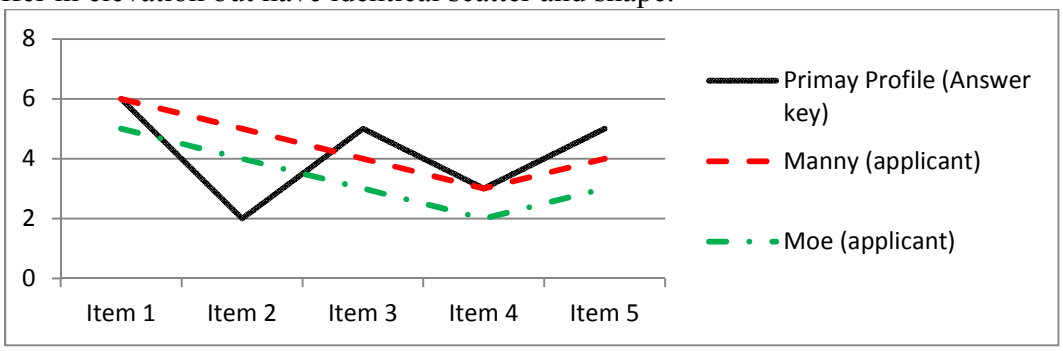
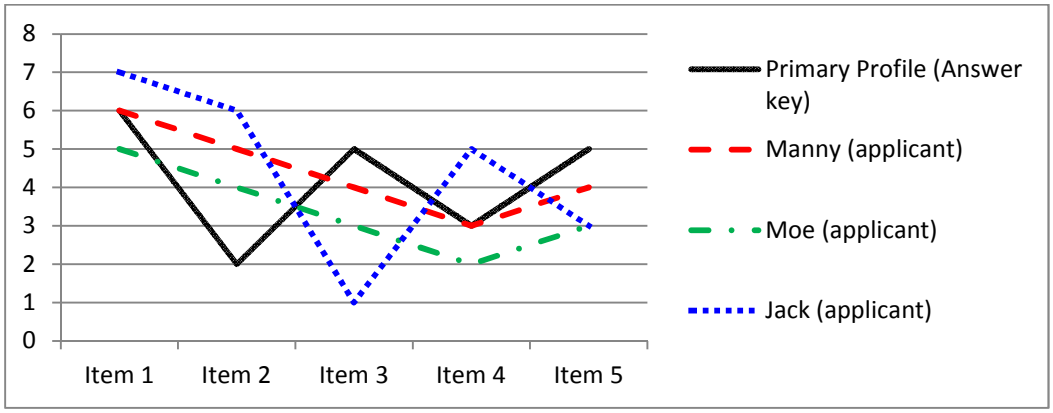


Figure 1c. The profile of the answer key with profiles for three applicants. The ratings of Jack have greater scatter than the ratings of Manny and Moe. Jack's shape also differs from the others. Many and Jack have identical elevation.



[1] The graphs in Figure 1 are best viewed in color.