Symposium/Forum
TITLE
Situational Judgment Tests: Research from an interactionist perspective

ABSTRACT
This symposium describes SJT research from an interactionist perspective where behavior is a function of the interaction between a person and the situation. One paper provides the theoretical framework; two papers describe characteristics of examinees that affect SJT performance; two papers describe features of SJTs that affect performance/validity of SJTs.

PRESS PARAGRAPH
This symposium describes recent SJT research from an interactionist perspective in which behavior is a function of the interaction between a person (P) and the situation (S): (B = f (P, S)). One paper provides the theoretical framework of interactionist theory; two papers describe characteristics of examinees that affect performance (demographic match between actors in a video and examinees themselves and knowledge/experience of examinees); two papers describe features of SJTs (content and scoring) that affect SJT performance/validity of SJTs.
Situational judgment tests (SJT) are widely used to assess a variety of constructs ranging from judgment to interpersonal skills. Recently, their use has become more prevalent and many vendors offer SJTs via the internet. Much research has been conducted to assess characteristics of examinees and features of SJTs that affect how people respond to such items. These papers review such research from an interactionist perspective. Campion and Ployhart provide an overarching framework for the papers in this symposium by describing SJTs through the lens of interactionist psychology, which states that the behavior is a function of the interaction between a person (P) and the situation (S): \( B = f(P, S) \). Although not explicitly, the remaining papers in this symposium are tied to this unifying framework. The next two papers in this symposium describe characteristics of examinees that affect performance on SJTs. The second paper describes the effect of demographic similarity on SJT performance. The third paper describes how knowledge/experience affects how people respond to SJTs. The remaining two papers focus on features of SJTs. The fourth paper describes the identification of factors underlying an SJT in which the situations developed to measure particular competencies. The fifth paper focuses on various scoring algorithms and how they affect the validity of SJTs.

Regarding characteristics of examinees, Golubovich and Ryan describe the influence of demographic similarity between actors in video-based SJTs and examinees and the influence of race and sex match on performance and perceptions of the SJT. Specifically, they investigated the extent to which race and sex “matches” between the actors and respondents affect performance on two common SJT content areas, helping and conflict. Overall, the findings indicate that individuals do attend to the race and gender of those they are “interacting with” in SJT scenarios and this influences their performance. For items related to helping, individuals performed differently when the actor was not of the same race as they were. For items related to
conflict, individuals did not perform differently when the actor was of a different racial background or a different gender from themselves.

Grand, Allen and Pearlman describe the influence of examinee knowledge and experience on SJT performance by comparing a mediation model (in which individuals with more accurate ITPs accrue more accurate job knowledge, which subsequently influences SJT responses) with a moderation model (in which respondents might generally rely on heuristics (ITPs) to make decisions unless more situationally relevant information (specific job knowledge) is available). Respondents were asked to respond to an SJT by selecting the most and least effective behavioral response for each item, data were also collected about respondents’ behavioral/temperamental tendencies, experience with specific military activities across various areas of expertise (used to calculate job knowledge), general cognitive aptitude (ASVAB), and job performance (supervisor ratings). Preliminary analyses suggest that respondents may differentially rely on knowledge/heuristic-based versus cognitive/rational-based reasoning depending upon how they are responding to an SJT scenario (i.e., identifying most versus least effective option).

Regarding features of SJTs (i.e., content) that influence performance, Waugh, Russell and Whetzel describe an SJT in which item writers wrote scenarios to measure five dimensions: adaptability to changing conditions, effective self-directed learning, effective self-management, relating to and supporting peers, and teamwork. Psychologists also rated the traitedness of response options using seven personality traits (e.g., achievement orientation and self-reliance). An exploratory factor analysis of the seven trait scales indicated an interpretable three-factor solution. The traits loading highly on the first factor were related to accomplishing tasks independently. The second and third factors were related to interacting with people.
McDaniel and Weekley describe methods for scoring SJTs while controlling for elevation and scatter. Elevation concerns within respondent means (e.g., whether the respondent’s ratings tend to be high or low). Scatter concerns within respondent variance (whether respondents rate the behaviors with similar scores or use the full range of the Likert scale). Although the validities for predicting an objective sales criterion were low, they replicated McDaniel et al. (2011) by showing that scales that do not control for elevation and scatter have lower validities than scales that do. Further, one can drop a large number of mid-range items and either increase or maintain validity.

In sum, these papers represent current research regarding the extent to which characteristics of examinees affect performance on SJTs. They also describe features of SJTs by describing the content of SJTs and how scoring algorithms can affect the validity of SJTs. These papers have a unifying theme in that they address components of interactionist theory.
In contrast to continued empirical research on situational judgment tests (SJTs), this presentation calls for a greater conceptual understanding of SJTs. Specifically, with rare exceptions (e.g., Motowidlo & Beier, 2010), current SJT development methods lack theory-driven approaches for structuring situational content. However, modern interactionist psychology is the result of over 40 years of research seeking to understand the joint relationship between persons and situations, and hence provides many important implications for SJTs.

Space constraints obviously limit us from going into greater detail, but our presentation will focus on (a) specific implications from the interactionist literature and (b) concrete examples of how SJT research and practice should change as a result of these implications.

**Job-Oriented SJT Development**

The situational content of SJT items is usually derived from a critical incidents approach. This approach produces SJT scores capable of achieving high criterion-related validity, but treats situations in a rather descriptive manner. That is, SJTs sample from a job domain but not a construct domain, and hence are unlikely to be homogenous (Arthur & Villado, 2008). We propose that SJT scores can achieve higher criterion-related and construct validity by adopting principles from interactionist psychology.

**Modern Interactionist Psychology**

Interactionist psychology states that the behavior is a function of the person (P) and the situation (S): (B = f (P, S)). Modern interactionism is based upon the following core principles:
The emphasis is on behavior from an intra-individual perspective. There are ideographically predictable patterns in an individual’s behavior across situations and through time. This requires an understanding of inter-individual differences in intra-individual behavior across similar situations (Mischel, 2004).

The focus is not on statistical interactions between P and S. Instead, modern interactionism suggests that behavior is a result of a dynamic reciprocal relationship between the person and situation (Krahe, 1992).

Stable individual differences (e.g., personality, ability) produce characteristic ways of perceiving, interpreting, and responding to situations. Thus, there is a mediated model linking trait-like individual differences to state-like individual differences to situations. Objective situational features are less important than perceived situational features (Mischel & Shoda, 1995); a feature shared by recent work on SJTs (e.g., Motowidlo & Beier, 2010) and personality (Tett & Guterman, 2000).

There are two ways to understand the psychological meaning of situations. The stimulus-analytical approach is a process by which researchers attempt to classify situations according to their perceived meaning to the individual. The response-analytical approach refers to the process by which researchers attempt to classify situations according to the responses elicited by the individual. Notice SJT research has adopted only the response-analytical approach, but both are necessary (Magnusson & Endler, 1977).

Situations differ in the demands they place on different knowledge, skill, ability, and other characteristics (KSAOs). KSAOs are relevant for some situations but not others.
• By presenting multiple psychologically similar situations repeatedly to participants, intra-individual behavioral profiles can be obtained. These mappings provide Type I (elevation or deviation from the mean) and Type II (shape) consistency measures (see Mischel & Shoda, 1995). These profiles provide a deeper understanding of an individual’s processing system and lead to better prediction of the person’s future behavior.

**Implications for SJT Development and Measurement**

The core principles from modern interactionist theory have many important implications for the development of a more construct-valid SJT. The following are among the most important:

• SJT development must begin with the careful construction of a detailed situational-dispositional taxonomy wherein the construct being assessed is explicitly linked to the situational content and the behavioral context (see Murtha, Kanfer, & Ackerman, 1996, for an example). In this way, situations are derived from constructs, not critical incidents, and behavioral outcomes can be assessed with higher construct validity. Remember, it is the relationship between the situation’s psychological meaning and the individual that produces behavior; this is why the situational component is so vital.

• Repeated presentation of psychologically similar situations provides the opportunity to develop intra-individual behavioral profiles (Mischel & Shoda, 1995). By dummy coding situations according to situational-dispositional dimensions and correlating those items with responses, a profile stability coefficient can be obtained without having to present identical items. These profile stability coefficients can then be correlated with performance (in a manner analogous to using growth curves as predictors).
Both stimulus-analytical and response-analytical approaches are necessary for developing construct-valid SJT items.

In conclusion, we contend that (a) more focus on the structure of situations is necessary to enhance construct validity, and (b) interactionist psychology provides direction on how to structure situations.

References


Demographic Cues In Video-based Situational Judgment Items

Juliya Golubovich and Ann Marie Ryan
Michigan State University

Selection researchers and practitioners have long been interested in reducing any adverse impact associated with assessment methods, and such research has extended to SJTs (e.g., Chan & Schmitt, 1997; Lievens & Sackett, 2006). One unexamined issue is whether actor race and gender in a video-based situational judgment test (SJT) affect the performance and reactions of test takers from various demographic groups. Relational demography theories suggest demographic similarity between individuals influences the quality of their interactions with one another and their individual outcomes (e.g., Goldberg, Riordan, & Zhang, 2008; Tsui, Egan, & O’Reilly, 1992; Tsui & O’Reilly, 1989). Thus, we might expect that individuals assuming the perspective of someone interacting with actors in a video might be influenced by their similarity to those actors. This is particularly true in helping and conflict contexts – two common types of SJT scenarios. Research has shown that individuals’ willingness to help others can be influenced by the helper’s race and gender relative to the victim’s (e.g., Eagly & Crowley, 1986; Kuntsman & Plant, 2008). This research suggests, for instance, that White individuals may be more willing to help other Whites than they are to help Blacks, and that men may more readily help women than other men. Research on conflict suggests that individuals may show lower assertiveness and lower cooperativeness when responding to conflict with someone from a different racial group (Davidson, 2001) and that both men and women may be less willing to display aggressive behavior and may be more likely to act positively and cooperatively when interacting with women (Carli, 1989; Russell & Owens, 1999; Shute & Charlton, 2006). If individuals do respond differently to helping and conflict SJT scenarios based on their own relative to the actor’s
demographics, these response differences may translate into group performance differences on the test.

To examine the implications of actor race and gender in a video-based situational judgment test relative to the test taker’s, a lab study with a 3x2x2x2 design (race of respondent: White, Black, Asian; race match between respondent and actor: yes, no; sex of respondent: male, female; sex match between respondent and actor: yes, no) was conducted. Participants (N = 374) were instructed to imagine themselves in the role of job applicants for a plant technician job asked to complete a situational judgment test. They viewed one of the six versions of the test.

Respondents’ most and least likely responses to the SJT items were scored according to a scoring key previously validated for the plant technician position. Overall, the findings indicate that individuals do attend to the race and gender of those they are “interacting with” in SJT scenarios and this influences their performance.

For items related to helping, individuals performed differently when the actor was not of the same race as them. For one of the items, lower performance occurred for all groups (Asian, Black and White) when the actor was of a different race than they were. For the other helping item, Black test takers performed considerably more poorly when the actor was not Black. The gender of the actor did not have a strong influence, except that female test takers did score better on one of the items than males when the White actor was male rather than female.

For items related to conflict, individuals did not perform differently when the actor was of a different racial background or a different gender from themselves. However, Asian and Black participants did perform better when the actor was female rather than male.

Relative to White respondents, minorities (Asians, Blacks) perceived higher job relatedness of the test. Women perceived less compatibility with the organization relative to male
respondents. Similarity to the actors in the video did not have much effect on reactions. Minorities (Asians, Blacks) perceived better opportunity to perform on the test when they viewed minority as opposed to White non-focal actors, but otherwise similarity in race or gender had no effects.

Although these findings are based on a limited sample of items and research participants rather than actual job applicants, the fact that the demographic characteristics of actors in SJT videos influenced performance on the test and minority respondents’ perceptions of opportunity to perform, points to the importance of organizations attending to cues in SJT items. Organizations may want to systematically consider demographic cues of actors in developing and validating SJT items. More generally, greater attention to how the demography depicted in assessment materials affects test taker responses would be a useful direction to take in development of new tools.

References


Considerable debate exists regarding the construct validity of situational judgment tests (SJTs) (McDaniel & Nguyen, 2001; Weekley & Jones, 1999). Arguments surrounding SJT measurement have generally adopted one of two perspectives—1) SJTs are unitary measures of intuitive and generalizable procedural knowledge (Sternberg et al., 1995), or 2) SJTs are multidimensional measures that reflect numerous constructs relevant to job knowledge and performance (cf., Chan & Schmitt, 1997). However, alternative conceptualizations that attempt to integrate the perspectives have also been pursued (e.g., Motowidlo, Hooper & Jackson, 2006). For example, Motowidlo and Beier (2010) propose that SJT performance reflects two facets of an individual’s knowledge-base: implicit trait policies (ITPs), defined as beliefs about the general effectiveness of specific behavioral expressions/personality traits across situations (i.e., “Being agreeable is generally good”), and specific job knowledge about effective actions unique to a particular context (i.e., “Being agreeable is good in this job situation”). According to the authors, “...[ITPs] represent general domain knowledge because accurate ITPs can be learned before people enter any particular job situation and are not dependent on specific job experience” (p. 323).

Current Study

Motowidlo and Beier (2010) thus propose that individuals use generalized knowledge (ITPs) and job-specific knowledge when responding to SJTs. In doing so, these authors draw
from the knowledge acquisition literature to propose a mediation model in which individuals with more accurate ITPs accrue more accurate job knowledge, which subsequently influences SJT responses (Figure 1A). Despite promising evidence for the relations among these components, there is also an alternative, equally plausible account for the relation between ITPs, knowledge, and SJT performance supported by Bayesian decision-making approaches (Figure 1B); respondents might generally rely on heuristics (ITPs) to make decisions unless more situationally relevant information (specific job knowledge) is available (Kahneman & Tversky, 1973). This suggests a moderation model in which these two knowledge components demonstrate a significant interactive effect on SJT performance.

The current research examines the role of knowledge and experience in SJT performance using data collected from 1,893 Army Non-Commissioned Officers (NCOs) to directly test these competing models. Data were gathered from NCOs participating in a concurrent validation effort for a 40-item interpersonal skills SJT where most and least effective behavioral responses were recorded for each item. Data were also collected about respondents’ behavioral/temperamental tendencies in interpersonal situations (i.e., agreeableness, openness, and conscientiousness; used to calculate ITPs), experience with specific military activities (used to calculate job-relevant knowledge), general cognitive aptitude (ASVAB), and job performance (supervisor ratings).

Analyses and Implications

Regression analyses revealed that although the ITP composite was significantly related to job-relevant knowledge and SJT performance, there was no change in the regression weights of the ITP-SJT performance relationship when job knowledge was included in the model, failing to support a mediation model (Baron & Kenny, 1986). In contrast, a small but statistically significant interaction between ITP and job knowledge on SJT performance was observed for
respondents selecting the most effective SJT option (Figure 2). Furthermore, Table 1 shows that ITPs (Z=2.20, p<.05) and specific job knowledge (Z=4.27, p<.01) were both more strongly correlated with correctly identifying the most effective course of action for an SJT item than the least effective course of action, whereas cognitive aptitude was more strongly correlated with correctly identifying least effective responses (Z=3.27, p<.01).

Together, these preliminary analyses suggest that the influence of general and specific knowledge on SJT performance may be more complicated than either the simple mediation or moderation models would predict. Of note, it appears that respondents may differentially rely on knowledge/heuristic-based versus cognitive/logic-based reasoning depending upon how they are responding to an SJT item (i.e., identifying most versus least effective option). Such findings may hold a promising direction for future investigations of SJT construct validity. For example, a great deal of research exists on the manner by which systematic cognitive biases emerge from heuristic-based decisions (e.g., representativeness biases, anchoring, etc., Tversky & Kahneman, 1974)—how might such response tendencies affect the validity of SJT measurements? These findings may hold possible practical implications as well. For example, the stronger relationship between job knowledge-performance (versus ITP-performance) suggests that SJTs where respondents are asked to select the most effective response option may be better suited for instances where employers wish to distinguish among individuals with more versus less job-relevant experience. Alternatively, requiring respondents to select most and least effective response options within the same instrument could potentially improve the practical utility of SJTs as it may enable administrators to assess and/or control for differences between experiential-based and cognitive/logic-based reasoning skills.
References


Figure 1. Proposed model of knowledge, experience, and SJT test performance.

Figures adapted from Motowidlo and Beier (2010). The lower boxes are displayed to highlight the difference between Motowidlo and Beier’s (2010) mediation model (Figure 1A) and the moderation model (Figure 1B) suggested by the current study.
Figure 2. ITP by Job Knowledge Interaction in Predicting “Most Effective” SJT Performance
Table 1.

Means, Standard Deviations, and Intercorrelations

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Job Specific Experience</td>
<td>4.05</td>
<td>1.07</td>
<td>(.91)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Openness</td>
<td>3.37</td>
<td>0.49</td>
<td>.05*</td>
<td>(.66)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Conscientiousness</td>
<td>1.22</td>
<td>0.23</td>
<td>.06**</td>
<td>−.01</td>
<td>(.59)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Agreeableness</td>
<td>1.26</td>
<td>0.23</td>
<td>.03</td>
<td>.10**</td>
<td>.52**</td>
<td>(.64)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. ITP Composite</td>
<td>0.00</td>
<td>0.71</td>
<td>.07**</td>
<td>.55**</td>
<td>.74**</td>
<td>.79**</td>
<td>(.74)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. General Cognitive Aptitude</td>
<td>58.11</td>
<td>18.57</td>
<td>−.15**</td>
<td>.18**</td>
<td>.01</td>
<td>.02</td>
<td>.09**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. SJT Most Effective Score</td>
<td>5.20</td>
<td>0.29</td>
<td>.23**</td>
<td>.05</td>
<td>.37**</td>
<td>.30**</td>
<td>.34**</td>
<td>.15**</td>
<td>(.74)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. SJT Least Effective Score</td>
<td>3.93</td>
<td>0.39</td>
<td>.09**</td>
<td>.02</td>
<td>.29**</td>
<td>.27**</td>
<td>.25**</td>
<td>.57**</td>
<td>(.84)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. SJT Total Score</td>
<td>2.13</td>
<td>0.60</td>
<td>.17**</td>
<td>.04</td>
<td>.36**</td>
<td>.32**</td>
<td>.34**</td>
<td>.23**</td>
<td>.85**</td>
<td>.92**</td>
<td>(.85)</td>
<td></td>
</tr>
<tr>
<td>10. Overall Effectiveness</td>
<td>5.09</td>
<td>1.12</td>
<td>.06</td>
<td>.00</td>
<td>.03</td>
<td>.01</td>
<td>.02</td>
<td>.05</td>
<td>.11**</td>
<td>.12**</td>
<td>.13**</td>
<td>(.47)</td>
</tr>
<tr>
<td>11. Soldier Rank</td>
<td>5.06</td>
<td>0.73</td>
<td>.58**</td>
<td>−.08**</td>
<td>.15**</td>
<td>.11**</td>
<td>.09**</td>
<td>−.05</td>
<td>.38**</td>
<td>.22**</td>
<td>.32**</td>
<td>.16**</td>
</tr>
</tbody>
</table>

* \( p < .05 \), ** \( p < .01 \). \( N = 962-1882 \). Reliability estimates are presented on the diagonal in parentheses. All are coefficient alphas except for Bisunally’s (1978) composite reliability and \( \text{ICC}[C,k] \); McGraw & Wong, 1996).

\(^a\)Score calculated by averaging keyed effectiveness rating of SJT options selected as “most effective” course of action

\(^b\)Score calculated by averaging keyed effectiveness rating of SJT options selected as “least effective” course of action

\(^c\)Score calculated by subtracting keyed effectiveness rating of SJT options selected as “least effective” from those identified as “most effective

\(^d\)Soldier rank ranges from 4 (E4) to 6 (E6)
The construct validity of situational judgment tests (SJTs) has been an on-going research question since interest in these assessments returned in the early 1990s with Motowidlo, Dunnette and Carter’s (1990) article on low-fidelity simulations. Although SJTs typically have good criterion-related validity when used for selection, very little is known about which constructs are measured by an SJT. Most factor analyses have yielded factors that account for minimal amounts of variance and solutions that are largely uninterpretable.

Developers often design an SJT to measure several constructs. Typically, the score for each construct is computed by summing the scores on the items designed to measure that construct. However, an item might not be measuring the construct it was intended to measure. It is difficult to write an SJT item whose response options all measure the same construct. In fact, hererogeneity can be a problem even at the option level—let alone at the item level. Two examinees might judge the same action to be effective for quite different reasons. Thus, the same action could be measuring different constructs for different examinees.

Purpose

The current study compares two different rational approaches to measuring constructs in an SJT. In one approach, the item writers were asked to write each scenario to measure the following five dimensions: adaptability to changing conditions, effective self-directed learning, effective self-management, relating to and supporting peers, and teamwork. In the other approach, 5 industrial-organizational psychologists rated the traitedness of each response option on seven personality traits: achievement orientation, self-reliance, dependability, sociability, agreeableness, social perceptiveness, and team orientation. For each trait-option combination, the
psychologists rated the degree to which the action and trait were related. Response options were dropped if they had moderate or high traitedness ratings on multiple traits or if there was substantial disagreement among the psychologists. Factor analyses will be used to assess the extent to which the items tap the five intended dimensions and the extent to which the personality traits are supported by factor analysis at the response option level.

Instrument and Sample

Each SJT item consisted of a description of a problem situation—typically three sentences—and several alternative actions. The examinee was asked to rate the effectiveness of each action on a seven-point scale, where higher numbers represent higher levels of effectiveness. The SJT was administered to 319 Soldiers in U.S. Army reception battalions. These Soldiers had just entered the Army but had not been assigned to training at that time. Each Soldier took 1 of 4 different forms. Thus, the sample size for each form was about 80. Each form had 14–16 items. Most items had seven response options.

The score on a trait for a specific option was computed as shown in Equation 1 below. The trait score for an entire test is the mean of the trait scores across the options linked to the trait.

\[
\text{optionTraitScore} = (\text{SoldiersRating} - \text{keyedEffectiveness}) \times \text{traitedness} \tag{1}
\]

The judgment score for an option was computed as shown in Equation 2 below. The difference between the rating and keyed effectiveness values is subtracted from 6 so that higher values represent better scores. The judgment score for an entire test form was the mean of the option scores.

\[
\text{optionEffectivenessScore} = 6 - |\text{SoldiersRating} - \text{keyedEffectiveness}| \tag{2}
\]
Analyses and Results

Several factor analyses will be done to examine the constructs assessed by the SJT. Factor analyses will be conducted at both the item level and response-option level. Factor analyses will be done to see (a) how well the items fit the five dimensions used by the item writers, (b) how well the options fit the seven traits used in the traitedness ratings, (c) whether the items comprise an interpretable factor structure, (d) whether the options comprise an interpretable factor, and (e) whether the seven traits comprise an interpretable factor structure.

An exploratory analysis of the seven trait scales has already been done. Parallel analyses clearly indicated a three-factor solution. The orthogonal model was more interpretable than the oblique model. The traits loading highly on the first factor are related to accomplishing tasks independently. Factors 2 and 3 are related to interacting with people. Factor 2 appears to involve working with other people to accomplish tasks. Factor 3 appears to be almost equivalent to the Agreeableness trait. The content of the factor structure can be loosely interpreted using the dimensions of initiating structure (Factor 1) and consideration (Factors 2 and 3) unearthed in the Ohio State Leadership studies (e.g., Stogdill, 1950).

References


Controlling for elevation and scatter in SJT scoring: A replication

Michael A. McDaniel  
*Virginia Commonwealth University*

Jeff A. Weekley  
*Kenexa*

This paper is intended to replicate McDaniel, Psotka, Legree, Yost and Weekley (2011), who found that situational judgment test (SJT) scoring methods that control for elevation and scatter are superior to methods that do not. Elevation concerns within respondent means (e.g., whether the respondent's ratings tend to be high or low). Scatter concerns within respondent variance (whether respondents rate the behaviors with similar scores or use the full range of the Likert scale). McDaniel et al. also demonstrated that the validity of a SJT scale can be maintained or enhanced by dropping items with mean Likert ratings near the middle of the scale (e.g., near 3 on a 5 point scale).

**Method**

The SJT contained 52 stems with 5 responses each resulting in 260 items. The criterion was an objective sales performance criterion composed of four factors: (1) percent of quota obtained; (2) year-over-year growth; (3) percent of standard activities achieved; and (4) percent of pipeline to quota.

One consensus key was based on SME means. We also calculated a raw consensus scale, a standardized consensus scale, and a dichotomized consensus scale consistent with the McDaniel et al. (2011) method. For the last three keys we also explored the validity of the scales based on means from 3 subsets of respondents formed with a disjoint cluster analysis (Anderberg, 1973, nearest centroid sorting).
Results and Discussion

Table 1 shows the validities of various SJT scales. The first column describes the methods for creating the key. Column two shows the validity when the mean is based on the SME mean (for the SME key) or on all respondents (for the raw, standardized, and consensus key). In this study, midrange items were defined as those with a mean Likert rating between 2 and 4. In columns 2-5 the validity of a scale with the mid-range items is deleted shown in parentheses. The scales based on all items are built from 260 items. The scales created after dropping mid-range items are based on 117 items.

In Table 1, the validities in the third column were based on the mean of the 100 observations in the largest cluster. The validities in the fourth column were based on the mean of these 66 observations in the second largest cluster. Finally, in the fifth column, the validities are based on the third largest cluster (the mean of these 17 observations).

Table 1 offers several findings. First, none of the scales resulted in large prediction of the objective sales criterion. Second, despite the low magnitude of the validities, the study replicated McDaniel et al. (2011) by showing that scales that do not control for elevation and scatter (raw consensus scales) have lower validities than scales that do (standardized consensus and dichotomized consensus). For example, in column two, although the raw consensus SJT scale had a validity of zero, the validity for the standardized consensus scale (.03) and the dichotomized consensus scale were higher (.06). Third, the study replicates McDaniel et al. in showing that one can drop a large number of mid-range items and either increase or maintain validity. In column two, the SME based scale validity 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.
When the means of the largest cluster of respondents (N = 100) was used in deriving the SJT key, the validity of the SJT scales were larger than the means based on all the respondents. The validities improved further when the mid-range items were dropped for the raw consensus scale (.02 vs .09) and the standardized consensus scale (.04 vs .10), but not for the dichotomized consensus scale (.11 vs. .07).

When the means of the second largest cluster (N = 66) were used in building the SJT scoring keys, the validities went negative but changed to larger magnitude positive validities when midrange items were dropped. For the keys based on the means of third largest cluster (N = 17), the validities showed the expected pattern of validities with respect to controlling for elevation and scatter and increases in validity when min-range items were dropped.

References


Table 1. Validities as a function of type of consensus key, sample from which means are drawn, and retaining all items or dropping mid-range items

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SJT Key</strong></td>
<td>Validity based on all observations</td>
<td>Validity of key based on means of 100 obs. in largest cluster</td>
<td>Validity of key based on means of 66 obs. in 2nd largest cluster</td>
<td>Validity of key based on means of 17 obs. in 3rd largest cluster</td>
<td></td>
</tr>
<tr>
<td>SME</td>
<td>.01 (.10)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Raw Consensus</td>
<td>.00 (.10)</td>
<td>.02 (.09)</td>
<td>-.04 (.09)</td>
<td>-.01 (.01)</td>
<td></td>
</tr>
<tr>
<td>Standardized Consensus</td>
<td>.03 (.10)</td>
<td>.04 (.10)</td>
<td>-.01 (.09)</td>
<td>.03 (.08)</td>
<td></td>
</tr>
<tr>
<td>Dichotomized Consensus</td>
<td>.06 (.07)</td>
<td>.11 (.07)</td>
<td>-.05 (.07)</td>
<td>.09 (.08)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Validities for scales with mid-range items dropped are in parentheses.