Journal of Vocational Behavior 104 (2018) 199–209

Contents lists available at ScienceDirect

Journal of Vocational Behavior

journal homepage: www.elsevier.com/locate/jvb

Optimizing the validity of situational judgment tests: The importance of scoring methods

Qingxiong (Derek) Weng[a,*], Hui Yang[a], Filip Lievens[b], Michael A. McDaniel[c]

[a] School of Management, University of Science and Technology of China, China
[b] Ghent University, Belgium
[c] Virginia Commonwealth University, United States

ARTICLE INFO

Keywords:
Assessment
Situational judgment test
Extreme response tendency
Scoring method
Criterion-related validity

ABSTRACT

In recent years, situational judgment tests (SJTs) have made strong inroads in assessment practices. Despite the importance of scoring for the validity of SJTs, little attention has been paid to different SJT scoring methods. This study investigated the influence of scoring methods on the criterion-related validity of SJTs. We examined five different consensus scoring methods (i.e., raw, standardized, dichotomous, mode, and proportion scoring) and several integrated scoring methods for scoring the same SJT. Results showed that one of the most popular scoring approaches (raw consensus scoring) is associated with an extreme response tendency and yields the lowest scale validity of all scoring approaches examined. Moreover, the mean item validity of midrange items was good only when they were scored by the mode consensus method. Thus, this study extends previous work (McDaniel et al., 2011) by deepening our understanding of how different scoring methods improve the validities of SJTs. Our findings suggest that using scoring methods that control the influence of extreme response tendency on the scores of SJTs yields higher validities. Finally, this study is the first to suggest that scoring SJTs with integrated methods yielded higher mean item validities than using any single method.

Throughout people's career, assessment instruments play important roles in evaluating their individual differences. Hence, assessment instruments are omnipresent for both career guidance/counseling (typically a within-person assessment) and career decision-making (typically a between-person assessment) (Watson & McMahon, 2014). For example, in early career stages (exploration and establishment stages), people complete assessment instruments for exploring career opportunities, getting a first full-time job, being promoted or for identifying one's strengths/weaknesses. In mid-career and later career stages, assessment instruments are also used to determine new task assignments/challenges or to even reassess people's careers.

Some career assessment instruments (e.g., Holland's RIASEC inventory) were specifically developed for career guidance/counseling/development, whereas others were adopted from existing selection tools (e.g., cognitive ability tests, personality inventories, assessment centers, situational judgment tests) and were thus originally developed for selection purposes. However, many of these traditional selection procedures have been widely accepted as useful career assessment instruments (Jansen & Vinkenburg, 2006; Lent, Brown, & Hackett, 1994; Tokar & Fischer, 1998; Volodina, Nagy, & Köller, 2015).

In the last decade, situational judgment tests (SJTs), as measures of people's procedural knowledge in specific domains such as interpersonal and leadership domains (Lievens & Sackett, 2012; Motowidlo, Dunnette, & Carter, 1990), have made particularly strong inroads in selection practices throughout the world. Although SJTs have typically been used as selection devices by organizations for

* Corresponding author.
E-mail address: wqx886@ustc.edu.cn (Q.D. Weng).

https://doi.org/10.1016/j.jvb.2017.11.005
Received 11 April 2017; Received in revised form 10 November 2017; Accepted 11 November 2017
Available online 13 November 2017
0001-8791/

making selection and career decisions about applicants and employees, SJTs are also useful for vocational purposes. This is because SJTs provide people with realistic job situations that they might encounter in their life, thereby assessing how they would react to these situations. Given these features, SJTs, have been increasingly used for instance, in college admissions, either as a mandatory test to be admitted to college or as a non-mandatory self-assessment tool (e.g., Lievens, 2013).

The criterion-related validity of SJT scores was established in dozens of primary studies and in several meta-analyses (e.g., Chan & Schmitt, 1997; Christian, Edwards, & Bradley, 2010; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; Motowidlo et al., 1990; Smith & McDaniel, 1998). Consistent with other assessment methods, the validity of SJTs is likely influenced by the scoring method (Arthur et al., 2014; Campion, Ployhart, & MacKenzie, 2014). Only when the scoring is valid can an accurate portrayal of people on relevant characteristics be obtained. There exist many ways of scoring SJTs and importantly a few studies indicate that validities vary by scoring method (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; McDaniel, Psotka, Legree, Yost, & Weekley, 2011). Unlike assessments with clearly correct answers, SJT responses cannot readily be identified as correct or incorrect. As such, items are typically scored with some form of consensus judgment (Legree, Psotka, Tremble, & Bourne, 2005). Once a consensually-derived scoring key is defined, the score of a respondent is a function of the degree of match between the respondent's ratings and the scoring key.

Raw consensus scoring, a most common and traditional scoring method for SJTs, however, has two serious problems. First, those who tend to use extreme ends of a Likert rating scale (i.e., extreme responders) are likely to lower scores on the SJT. Those who provide extreme ratings, will on average, have larger deviations from the scoring key mean, resulting in less favorable scores. To the extent that extreme response tendencies are unrelated to job performance, score differences caused by differential use of extreme responding would constitute test bias, thus affecting vocational and selection decisions. The second major problem with consensus scoring as Cullen, Sackett, and Lievens (2006) demonstrated is that a coaching strategy of avoiding extreme rating points on a Likert scale can substantially increase SJT scores. Therefore, one approach to improve the validity of SJTs is to use a method could minimize the influence of extreme ratings on the personal scores.

To address the problems exist in raw consensus method, McDaniel et al. (2011) suggested two alternative methods (*standardized consensus* and *dichotomous consensus*), and found that using either method, one can control elevation and scatter (Cronbach & Gleser, 1953), leading to higher item validity and scale validity. These two methods provide possible solutions for the two serious problems associated with raw consensus scoring. Other consensus methods, such as *mode and proportion consensus* (see the description of these methods in Table 1), which have been widely used for emotional intelligence (EI) tests (Barchard et al., 2013; Barchard & Russell, 2006; MacCann et al., 2004) have not been widely adopted in SJTs. With respect to EI tests, the mode and proportion consensus methods may be promising approaches because they purportedly offer unidimensional scores and demonstrate convergent validity (Barchard & Russell, 2006; MacCann et al., 2004). More importantly, both mode and proportion consensus are non-distance

Table 1
Descriptive summary of five consensus scoring methods.

Method	Representative articles	Description	Example
Raw consensus	Legree (1995) Legree et al. (2005) Sacco, Schmidt, and Rogg (2000) McDaniel et al. (2011)	A respondent's score on one item is the inversion of the squared deviation between the scoring key of this item and his/her rating; the scale score is an aggregation of the scores on all items.	If the scoring key of an item is 3.5, a respondent's score for a rating of 1 is the inversion of the squared deviation from 3.5, i.e., $1/(3.5–1)^2$.
Standardized consensus	Wagner (1987) Legree (1995) McDaniel et al. (2011)	Each respondent' ratings for all items are transformed to *z*-scores such that the mean across items is zero with a standard deviation of one. A respondent's score on an item and the whole scale is calculated as the approach in raw consensus to the *z*-scores of the ratings.	If the scoring key of an item is 3.5 and the *z* transformation of a respondent's rating is 1.2, the score on this item is the inversion of the squared deviation from 3.5 to 0.8, i.e., $1/(3.5–1.2)^2$
Dichotomous consensus	Lievens, Buyse, and Sackett (2005) McDaniel et al. (2011) Crook et al. (2011) Motowidlo, Martin, and Crook (2013)	This method uses the scoring key (i.e., raw item mean across respondents) to determine if an item is correct. If the group mean indicates that item is incorrect and the respondent indicates that the item is incorrect, the respondent receives a score of one; otherwise, the respondent receives a score of zero.	On a 5-point Likert scale, a group mean of 3 or above for an item is judged as correct, and group mean of below 3 is judged as incorrect. Thus, a respondent's rating of 3, 4, or 5 receives a score of 1, and a rating of 1 or 2 receives a score of 0.
Mode consensus	Geher, Warner, and Brown (2001) MacCann, Roberts, Matthews, and Zeidner (2004) Barchard and Russell (2006)	The mode, or the rating chosen by the largest proportion of the respondents, is judged as correct. If a respondent's rating is consistent with the mode, the respondent receives a score of one; otherwise, the respondent receives a score of zero.	If 4 in a 5-point Likert scale is the mode, a rating of 4 receives a score of 1, and ratings of 1, 2, 3, and 5 receive a score of zero.
Proportion consensus	Mayer, Caruso, and Salovey (2000); Mayer, Salovey, Caruso, and Sitarenios (2003) MacCann et al. (2004) Barchard, Hensley, and Anderson (2013)	A rating is scored by the proportion of respondents who have the same rating.	If 45% of respondents choose 1, a rating of 1 receives a score of 0.45.

consensus scoring methods because the SJT scores are not determined by the distance from the applicants' rating to the scoring key. Considering the advantages of the two consensus methods suggested by McDaniel et al. (2011) and the two non-distance methods drawn from the emotional intelligence scoring literature, we aim to compare the criterion-related validity of these four scoring methods with the traditional scoring method – raw consensus.

In addition to controlling for the influence of extreme response, another factor influencing SJT scoring and validity is that the means and standard deviations of response-option-level characteristics can predict each item's criterion-related validity on an SJT (Putka & Waugh, 2007). Waugh and his colleagues (Putka & Waugh, 2007; Waugh & Russell, 2006) reported U-shaped relationships between item means and item validities, such that items with low or high expert rating means had the highest validities. This suggests that items with means close to the midpoint of a Likert scale may have less informational value than items with means near the extremities (see McDaniel et al., 2011). Given the heterogeneity of items and the respective advantages of different scoring methods, a final objective of this study was to examine for the first time whether an integrated scoring method (i.e., combining different scoring depending on the characteristics of SJT items to be scored) provides better item validity than any single scoring method.

In summary, the purposes of this research are threefold: (a) to investigate the criterion-related validity of five different scoring methods; (b) examine how response tendency (extreme responding) and SJT item characteristics (midrange items) influence the relationship between these scoring methods and SJT validity; and (c) determine whether integrated scoring methods for items with different characteristics enhance the criterion-related item validity of an SJT. We conduct these examinations with the aim of optimizing scoring methods for SJTs that use Likert type scales. The following section details the hypotheses and research questions related to these objectives.

1. Hypotheses

1.1. Extreme response tendency and SJT score

An extreme response tendency refers to a tendency to select the endpoints of a rating scale (e.g., strongly disagree and strongly agree), leading to a larger variance that directly affects the strength of item intercorrelations (Bolt & Johnson, 2009). In particular, extreme responses typically result in lower scores on SJTs because the scoring key for an item is seldom at the endpoint of a scale (e.g., for the mean of a 5-point scale to equal 5, all experts/judges would have to select option 5, which is unlikely). Rather, mean scores of experts/judges[1] would be closer to the center of the distribution (e.g., between 2 and 4 on a 5-point Likert scale). Extreme response tendencies can be expected to occur across items and thereby influence individuals' overall SJT scores. This issue may be particularly dramatic for participants who use only a part of the rating scale (Legree, 1995), that is, who present an extreme response tendency.

In raw consensus scoring, respondents' scores depend on the sum of the absolute or squared difference between their answer and the “correct” answer. As such, when using raw consensus scoring for SJTs, participants' extreme response tendencies detrimentally affect scores. Following McDaniel et al.’s (2011) finding that two adjusted raw consensus methods (i.e., standardized and dichotomous consensus) can control for elevation and scatter of item ratings, we expect that scoring SJT by standardized and dichotomous consensus could result in less excessive correlations between participants' extreme response tendencies and their overall scores than by raw consensus.

Using the mode consensus method, the rating chosen/agreed upon by the largest proportion of the sample (i.e., the mode) is treated as the correct answer (i.e., scored as 1), while all other responses are treated as incorrect (i.e., scored as 0). The scoring key for the mode consensus method is dichotomous because respondent scores are judged as “correct” or “incorrect.” Dichotomizing the items removes individual differences in response tendency (Whetzel & McDaniel, 2009) and negates (McDaniel et al., 2011) the effects of coaching approaches mentioned by Cullen et al. (2006).

Proportion consensus scoring can be used for a variety of response formats (e.g., categorizing, rating scales), which is considered as a valuable scoring method for EI tests (e.g., MacCann et al., 2004; Mayer et al., 2000). Proportion consensus scoring refers to scoring items based on the proportion of respondents who gave the same response as the test-taker (Barchard et al., 2013). When there is no consensus scoring key generated either by norms or experts, or when one scoring key is not necessarily better than other keys, proportion scoring can be used to produce a solution. Compared to the distance consensus methods, a respondent's score on any item with proportion method could also be less influenced by the extreme response tendencies, as the score is not determined by the distance between the rating and the key answer. So, the following hypotheses were proposed:

Hypothesis 1. Compared to the raw consensus method, using standardized, dichotomous, mode and proportion consensus methods for SJTs will result in weaker relations between extreme response tendencies and overall SJT scores.

We now discuss how these scoring methods affect criterion-related validity. Scoring SJTs with the standardized and dichotomous consensus methods has been found to yield greater validity scores than when scoring using the raw consensus method (McDaniel et al., 2011). We sought to replicate these previous findings and add to them by clarifying whether scoring SJTs using the mode and proportion consensus methods could also yield greater validity values than raw consensus scoring.

The proportion consensus method identifies optimal answers (Legree, 1995) to a test by assessing individuals' convergence with

[1] In case the mean responses of the sample of respondents are chosen as the scoring key, Cullen et al. (2006) reasoned that it was also unlikely that sample means for any given item would be extreme.

the mean appraisals in non-experts,[2] in this case with SJT respondents. Proportion consensus is considered logically plausible for scoring tests that evolve within a social context (such as EI tests) because group consensus should accurately reflect the correct answers in such cases (Barchard et al., 2013; Mayer et al., 2003). In applying proportion consensus scoring method to SJTs, we can infer that choosing a response option with a higher consensus has a high probability of being related to the criterion (e.g., responding adequately to job demands). Mode consensus method works in a similar way as the proportion method; it treats the rating on the point chosen by the most as 1 and the rating on other points as 0. Thus, we expect that using the mode or proportion consensus method for scoring SJTs could also lead to higher criterion-related validity than using the raw consensus method. Scale validity is a function of the item validity, the number of items, and the criterion-relevant redundancy of the items (McDaniel et al., 2011), therefore, higher item mean item validity does not necessarily lead to high scale validity. Accordingly, we offer the following hypothesis:

Hypothesis 2. Compared to the raw consensus method, using standardized, dichotomous, mode and proportion consensus scoring methods for an SJT will yield higher criterion-related validity at both the item and scale level.

1.2. Scoring items with different characteristics

As mentioned previously, the relationship between item means and item validity is a U-shaped curve (e.g., Putka & Waugh, 2007; Waugh & Russell, 2006): items with low or high expert rating means exhibit the highest validities, whereas items with means near the midpoint of a Likert scale have less informational value than items with means near the extremities (McDaniel et al., 2011). Midrange items (e.g., 3 on a 5-point scale) likely have less informational value because they are less sensitive to poor judgment by respondents (see McDaniel et al., 2011, for more detail).

However, if we score the SJT using the mode or proportion consensus methods, the problems with raw consensus scores are minimized. Thus, we expect that scoring SJTs using the mode consensus or proportion consensus methods deals more effectively with these midrange items than the raw consensus method for two reasons. First, when using either method, the midrange items remain sensitive to assessments of poor judgment. For example, consider the use of mode consensus scoring on a 5-point Likert scale. If the scoring key indicates that 3 is the best answer for the item, then responses of 3 for that item will be scored as 1 and all other responses will receive a score of 0. This suggests that the sensitivity to respondents' poor judgment is equal between midrange and non-midrange items. Second, in using these two scoring methods for an SJT, the correct answer is judged by the proportion of respondents' responses rather than the item mean. Thus, contrary to distance consensus scoring, midrange items have insufficient informational value and no longer exert detrimental effects. As such, we suggested the following hypothesis:

Hypothesis 3. Scoring midrange items using the mode and proportion consensus scoring methods will yield higher item criterion-related validity than the raw consensus method will.

Furthermore, following McDaniel et al.'s (2011) argument, we further differentiated SJT items according to their means and variance. Fig. 1A, B, C, and D display these various item types. The first item type (in Fig. 1A) has a mean (across respondents) close to the midpoint and a low variance in responses, whereas the second item type (Fig. 1B) has a mean close to the midpoint and a high variance. The third and fourth item types (Fig. 1C and D) both have non-midrange means, but the third item type has a low variance while the fourth has a high variance. For items with large variances, such as the second and fourth types, the disagreement among respondents is substantial.

In view of these various item characteristics and the possible advantages of different scoring methods for certain item types, we combined the various scoring methods for rating SJTs to examine whether an integrated method yields higher validity than does a single method. The integrated method is to score the mid-range items and non-midrange items with scoring methods that prior research found to be the most effective for dealing with these specific item types. Given the early state of our theoretical understanding of the relative advantage of the integrated strategy, we framed our aim as the following research question:

Research Question 1: Does scoring an SJT with an integrated method yield higher item and scale validities than scoring with a single method?

2. Method

2.1. Sample and procedures

Managers were recruited from five companies in China. Data were collected during these companies' annual management development programs. About one week before the program, the HRM department of each company asked all managers to complete a training needs survey that also contained a battery of items on job performance. The survey was framed as a training needs survey to allow participants to objectively report their performance. Initially, 151 managers participated in the survey. Several days later, coworkers and supervisors were also asked to rate these participants' job performance after being told that the results would be used as the basis of the annual performance appraisal. Then, at the beginning of the program, all participants received an announcement

[2] Research found a high correlation between scoring keys developed using general norms and those developed by experts (r = 0.908, N = 703; Mayer et al., 2003), suggesting that consensus scoring methods can be highly accurate (Barchard et al., 2013).

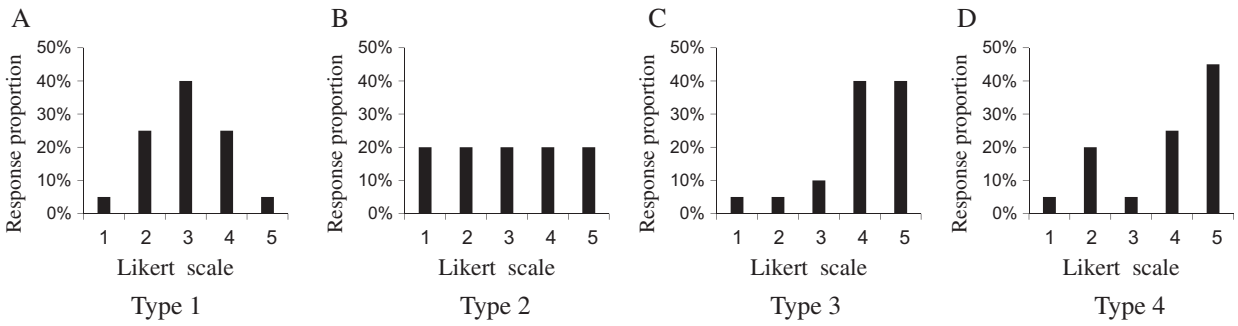

Fig. 1. Illustration of different item types by elevation and scatter.

that the company would organize an SJT in the next day and that their performance on the SJT would be used as a basis for making annual promotion decisions. During the SJT administration, nobody reported emotional distress, and only three participants were absent due to personal matters.

Resulting from this process, we obtained self-report (M = 4.16, SD = 0.26), supervisor-rated (M = 4.14, SD = 0.44), and peer-rated (M = 4.41, SD = 0.52) job performance ratings for all participants. The spread in ratings across the various performance appraisals for the 142 participants was acceptable (SD < 0.30), but nine participants received very diverse performance appraisals from different raters (SD > 0.55). Interrater agreement on the performance appraisals of each participant was also assessed by $r_{WG(J)}$ (James, Demarree, & Wolf, 1984). The average $r_{WG(J)}$ statistic across participants was 0.52, indicating on average at best moderate agreement levels of job performance ratings among different raters (Putka & Waugh, 2007).

After deleting the data of the nine participants with highly variable performance appraisal ratings and the three participants who did not take the SJT, 139 observations with SJT and performance ratings remained. Of this eligible sample, there were 60 men and 79 women. The mean age was 32.6 (range 26–52 years). The sample was diverse in terms of age, with 22% in the 26–30 age group, 38.5% in the 31–35 group, 22% in the 36–40 group, and 16.5% in the > 40 group. On average, participants had 2.5 years of organizational tenure.

2.2. Measures

2.2.1. SJT

The SJT used in this study was provided by McDaniel (personal communication, August 18, 2012). This SJT was comprised of 32 scenarios (item stems), all of which were management problems that managers might encounter in the workplace. Each item stem was followed by five to seven response options yielding 157 items. Participants were instructed to rate each response option in terms of its effectiveness from 1 (*very ineffective*) to 5 (*very effective*). The Cronbach's alpha of the SJT scores in this study was 0.72, which suggests that our SJTs have acceptable reliability (cf. Catano, Brochu, & Lamerson, 2012; McDaniel et al., 2001; Schmitt & Chan, 2006).

2.2.2. Criterion measures

Job performance was measured using Williams and Anderson's (1991) 7-item task performance scale (Cronbach's alphas = 0.81). All items were rated on a five-point Likert scale (1 = *very strongly disagree*, 5 = *strongly agree*). Sample items were “adequately completes assigned duties,” and “fulfills responsibilities specified in the job description.” Confirmatory factor analysis showed that the one-factor model of job performance fit the data well (χ^2/df = 1.24, CFI = 0.97, GFI = 0.95, RMSEA = 0.05, RMR = 0.04). The mean of job performance for participants was 4.22 (SD = 0.46).

2.3. Data analysis

2.3.1. Scoring approaches

Following McDaniel et al.'s (2011) methods, we used the means of each item across respondents as the scoring key and then calculated respondents' scores using the raw, standardized, and dichotomous consensus methods separately (see examples in Table 1). In particular, for raw consensus scoring, respondents' scores for a single item were calculated by summing the squared difference between each item's scoring key (i.e., the raw item mean across respondents) and their ratings. The score was then inverted so that higher scores reflected better performance on the SJT. For the standardized consensus method, the raw ratings were transformed using within-person z standardization; then, respondents' scores were calculated using the standardized ratings in a similar way as the raw consensus method. For the dichotomous consensus method, we used the scoring key to determine whether an item was correct or incorrect. If the group mean indicated that the item was incorrect and the respondent also indicated that the item was incorrect, the respondent received a score of 1 for that item; otherwise, the respondent received a score of zero (see example in Table 1).

We also scored the SJT using the mode and proportion consensus methods that have been commonly used in EI tests (Barchard et al., 2013; Barchard & Russell, 2006; MacCann et al., 2004, see examples in Table 1). As noted above, for the mode consensus

Table 2
Correlations between response tendencies and SJT scores (N = 139).

Variable	1	2	3	4	5
1. Extreme response tendency					
2. SJT scores (raw)	− 0.65**				
3. SJT scores (standardized)	0.02	0.63**			
4. SJT scores (dichotomous)	− 0.05	0.54**	0.79**		
5. SJT scores (mode)	0.23*	0.36**	0.85**	0.73**	
6. SJT scores (proportion)	0.16	0.44**	0.86**	0.87**	0.92**

Note. SJT, situational judgment test.
* $p < 0.05$.
** $p < 0.01$.

method, the mode was judged as the correct response (i.e., this served as scoring key). If the respondent's rating was consistent with the mode, the respondent received a score of one; otherwise, the respondent received a score of zero. Finally, using the proportion consensus, the respondents' ratings were scored on the basis of the proportion of all respondents who had the same rating.

To answer the research question, we calculated the validities of various integrated scoring methods. Specifically, rather than using a single scoring method for all SJT items, the integrated scoring methods utilized relatively superior consensus scoring methods for different type of items (i.e., midrange and non-midrange items). The strategies of integrating scoring methods will be discussed later.

2.3.2. Determination of midrange items

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 (McDaniel et al., 2011). As a 5-point Likert scale was used for the SJT, the midrange items were defined as those with a mean (across all respondents) between 2.5 and 3.5. We identified 46 midrange items, 41 of which had a high variance ($SD > 1.20$), and 5 of which had a low variance ($SD \leq 1.20$). Following the midrange item classification, the 111 remaining non-midrange items were also divided into two types: 91 items with low variance ($SD \leq 1.20$) and 20 items with high variance ($SD > 1.20$).

3. Results

3.1. Tests of hypotheses

Hypothesis 1 predicts that the negative relationship between extreme response tendency and individual scores is weaker for the standard consensus, dichotomous consensus, mode consensus, and proportion consensus methods than it is for the raw consensus method. The results reported in Table 2 appear to support this hypothesis: specifically, the association between extreme response tendency and individual score was significantly weaker for the standardized ($r = 0.02$, *ns*; $z = 6.56$, $p < 0.001$), dichotomous ($r = -0.05$, *ns*; $z = 5.98$, $p < 0.001$), mode ($r = 0.23$, $p < 0.01$; $z = 8.32$, $p < 0.001$), and proportion consensus methods ($r = 0.16$, *ns*; $z = 7.72$, $p < 0.001$) than it was for the raw consensus method ($r = -0.65$, $p < 0.01$). Following Holm's (1979) suggestions, we controlled for the family-wise error rate using the step-down Bonferroni procedure and the results showed that the relationship between extreme response tendency and individual scores for raw consensus is stronger than it is for other four consensus methods at $p < 0.05$.

Hypothesis 2 posits that scoring an SJT with the standard, dichotomous, mode, and proportion consensus methods yields higher criterion-related validities (both item and scale level) than does the raw consensus method. Results in Table 3 show that the mean *item* validity for raw consensus scoring (0.013) was lower than that for the standardized (0.020), dichotomous (0.028), mode (0.035), and proportion (0.044) consensus methods. The *scale*-level validities were consistent with the mean item validities by scoring strategy—namely, the validities for standardized (0.224), dichotomous (0.150), mode (0.263), and proportion consensus scoring (0.253) were substantially greater than that for the raw consensus scoring method (0.134). Although the mean item validity and scale validity for raw consensus scoring were consistently lower than the validities for the other scoring methods, the differences of the validities are not significantly. Therefore, Hypothesis 2 was not supported.

Hypothesis 3 proposes that rating mid-range items using the mode and proportion scoring methods yields higher item criterion-related validity than do other consensus methods. For midrange items with large variances (see Table 3), the mean item validities of the SJT scored by mode (0.024) and proportion consensus scoring (0.012) were greater than were those obtained by raw (− 0.011), standardized (− 0.007), and dichotomous consensus scoring (− 0.009). Similarly, the item validities for midrange items with low variance scored by mode (0.026) and proportion consensus scoring (0.004) were greater than were those obtained via raw (− 0.017), standardized (− 0.024), and dichotomous (− 0.027) consensus scoring. Importantly, by using mode consensus scoring the mean item validity of the midrange items were improved to an amount (0.024 and 0.026, for midrange items with large and low variance, respectively) that was even greater than the mean item validity of all items scored by raw (0.013) and standardized (0.020) consensus. Although the mode and proportion scoring methods yields higher item validity than other consensus methods, the differences of the item validity across three scoring methods were not significant. Thus, Hypothesis 3 was not supported. Surprisingly, the mean item validity remained relatively constant for mode consensus scoring across the two types of midrange items.

Table 3
Mean item validities and scale validities for various consensus scoring methods.

	Raw	Standardized	Dichotomous	Mode	Proportion
Scale validity					
Single scoring method	0.134	0.224**	0.150	0.263**	0.253**
Excluding midrange items[a]	0.108	0.185*	0.190*	0.250**	0.250**
Scoring midrange items with mode consensus[b]	0.109	0.186*	0.196*	0.263**	0.238**
Mean item validity					
Single scoring method	0.013	0.020	0.028	0.035	0.044
Excluding midrange items[a]	0.023	0.028	0.040	0.040	0.059
Scoring midrange items with mode consensus[b]	0.023	0.028	0.038	0.035	0.048
Midrange items (small variance)	− 0.017	− 0.024	− 0.027	0.026	0.004
Midrange items (large variance)	− 0.011	− 0.007	− 0.009	0.024	0.012
Non-midrange items (small variance)	0.025	0.039	0.044	0.019	0.081
Non-midrange items (large variance)	0.058	0.038	0.097	0.073	0.083

* $p < 0.05$.
** $p < 0.01$.
[a] In the SJT, 46 items were midrange items, of which 41 had high variance and 5 had low variance; 111 items were non-midrange items, of which 91 had low variance and 20 had high variance. Excluding midrange items means that the 46 midrange items were deleted from the scale.
[b] This indicates that the scale was scored using two different methods: the midrange items were scored using the mode consensus method while non-midrange items were scored using the other methods.

3.2. Research question

As presented in Table 3, the mean item validity for raw consensus scoring was 0.013, but this validity increased to 0.023 when the 46 midrange items were excluded. Similarly, the mean item validity for the other methods was consistently lower when midrange items were included vs. when they were excluded. These results indicate that mean item validity is higher for non-midrange items than for midrange items, regardless of how the items were scored. Moreover, Fig. 2 shows that there was a U-shaped relationship

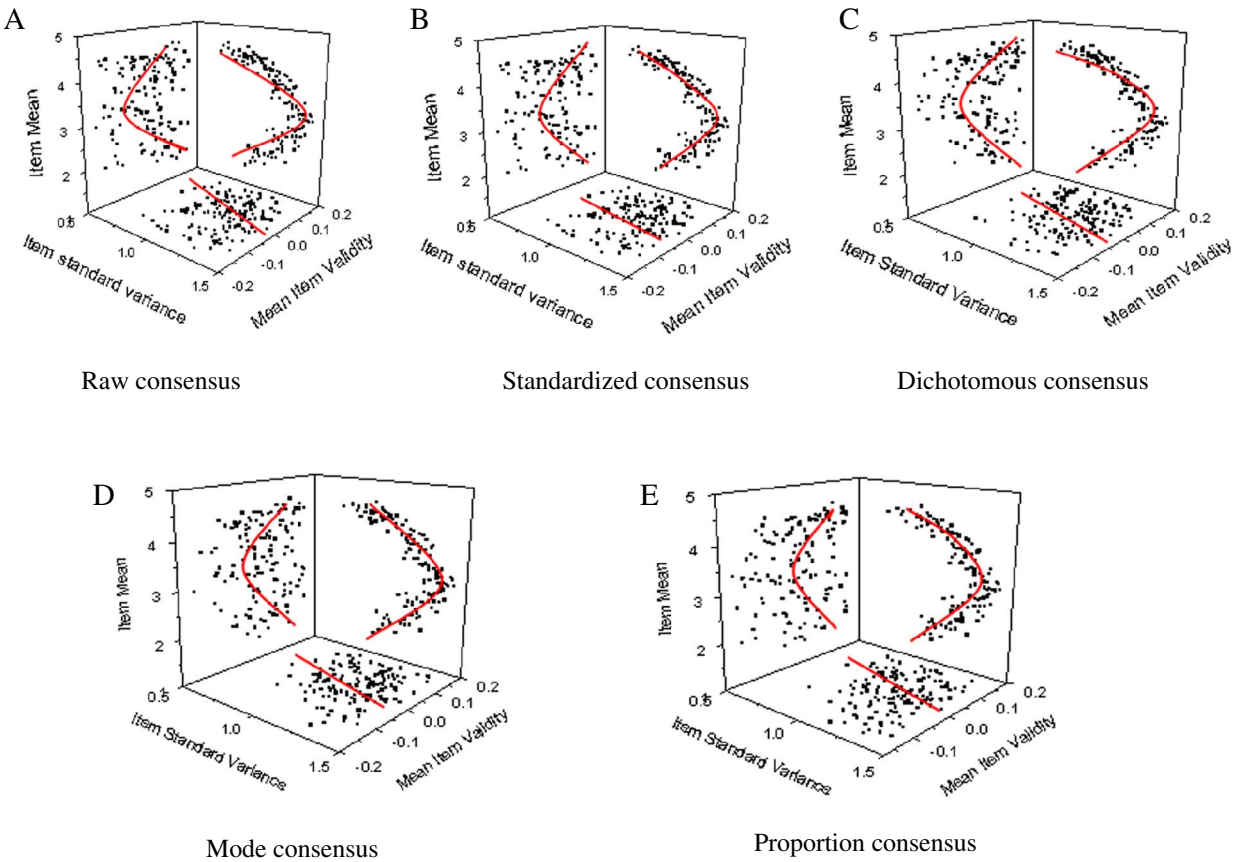

Fig. 2. The relationships between item means, standard variance, and criterion-related validity for the different consensus scoring methods.

Table 4
Mean item validities and scale validities for integrated scoring strategy.*

	Scale validity	Mean item validity
Integrated scoring method 1[a]	0.264**	0.052
Integrated scoring method 2[b]	0.253**	0.048

* $p < 0.05$.
** $p < 0.01$.
[a] For this integrated scoring method, the midrange items were scored using the mode consensus method and while the non-midrange items were scored using the proportion consensus method.
[b] For this integrated scoring method, the midrange items were scored using the mode consensus method while the non-midrange items were scored using the dichotomous consensus method.

between item means and item criterion-related validity and that there was a negative relationship between item standard deviation and item validity. The results of the binomial fitting and linear analysis also supported the U-shaped relationship between item means and item validities.

These results show that it makes sense to use integrated scoring methods. So, we tested whether the item validities could be improved by using integrated scoring methods—namely, using different consensus methods to score different types of items. Initially, we tested the effects of scoring midrange items using the mode consensus method while the remaining items were scored using the other methods. As shown in row 4 and row 6 of Table 3, after using the mode consensus method to score the midrange items, the mean item validities for the raw, standardized, dichotomous, and proportion consensus methods increased by 0.010, 0.008, 0.010, and 0.000 respectively.

As shown in Table 4, the mean item validity of the first integrated scoring method (i.e., scoring midrange items with mode consensus method and non-midrange items with the proportion consensus method) was 0.052, which was larger than any of the mean item validities obtained using a single scoring method. The second integrated scoring method (i.e., scoring midrange items with the mode consensus method, and scoring non-midrange items with the dichotomous consensus method) yielded a mean item validity of 0.048, which was higher than the mean item validity for both of the constituent scoring methods (e.g., dichotomous consensus, 0.028; mode consensus, 0.035). These results suggest that an integrated scoring method could yield higher mean item validities than any single scoring method could. Consistently, the scale validity of the first integrated scoring method was 0.264; for the second integrated scoring method it was 0.253, both of which were higher than using either single method (standardized, 0.224; dichotomous, 0.150) as recommended by McDaniel et al. (2011). Moreover, the scale validity of the first integrated scoring method (integrating mode and proportion consensus, 0.264) was higher than that using either the mode consensus (0.263) or the proportion consensus method (0.253). This suggests a slight advantage of using an integrated scoring approach to increase scale validity. Nonetheless, the results suggest that not all integrated methods will yield higher scale validities than single scoring methods, since the scale validity of the second integrated validity method (integrating by mode and dichotomous consensus, 0.253) was lower than that of SJTs scored by only mode consensus (0.263).

4. Discussion

SJTs are versatile instruments and are therefore increasingly popular for between-person (organizational decision making in the context of selection or promotion) and within-person assessments (self-assessment in the context of career counseling, guidance or development; Watson & McMahon, 2014). This study examined whether the SJT scoring method matters in processing the assessment outcomes and how scoring influences the criterion-related validity of SJT scores. Our study adds several important pieces of knowledge to the SJT literature and especially to the scoring on SJTs.

4.1. Main contributions

A key finding of our study is that alternative distance consensus scoring methods (i.e., standard, dichotomous, mode, and proportion consensus methods) are less prone to the extreme scoring tendency than the popular raw distance consensus method. Therefore, these alternative distance consensus scoring methods mitigate the negative effect of extreme response tendencies on SJT scores (i.e., exhibiting higher item validity and criterion-related validity). This is consistent with McDaniel et al. (2011), in that adjustments to the distance consensus scoring method for SJTs results in substantial improvements in item validities. Our study, however, also went beyond previous studies by examining which of these scoring methods best benefits the item validity of SJTs.

Extending McDaniel et al.'s (2011) results that adjustments to the raw consensus scoring method result in higher item validity, the present study shows that four different consensus scoring methods have higher validities than the raw consensus method. McDaniel et al. (2011) noted that response tendencies are best viewed as a source of systematic error in SJTs and therefore it is important to deal with them in the scoring procedure. Our results demonstrate that adjustments to consensus scoring methods result in improved criterion-related validity. In accordance with McDaniel et al. (2011), we also found that both standard and dichotomous consensus scoring had higher validities than did raw consensus scoring. Even more interesting, the mode and proportion consensus methods—two scoring methods firstly introduced to the SJT literature by this study—had a higher scale validity and mean item validity, respectively, than did the other methods.

The present study also makes an important contribution to the research on how to deal with midrange items of SJTs. Generally, more systematic research is needed to determine precisely what item characteristics influence item validity (McDaniel & Nguyen, 2001). Our findings suggest a fine-grained strategy in that midrange items with high variances could be scored using the proportion consensus method, whereas those with low variances could be scored using a dichotomized consensus method such as the mode or dichotomous consensus methods. Moreover, although our results correspond to McDaniel et al.'s (2011) in that item validity increased when midrange items are removed from the analysis, we note that other strategies — namely, use of the mode consensus method — might better solve the problems associated with midrange items. An alternative to dropping midrange items is to apply the mode consensus method in order to improve both item and scale validity. The fact that the mode consensus method is best suited to the validity of midrange items is perhaps because it utilizes the score that the majority of participants selected. In other words, it would counteract the high variance (i.e., lack of consensus) associated with midrange items. In contrast, applying the raw consensus method to these midrange items results in lower SJT scores for respondents with extreme response tendency. As such, our results extend the research base related to the removal and scoring of midrange SJT items.

Interestingly, the proportion and dichotomous consensus methods were superior to other methods for non-midrange items with lower variances. When comparing the results for excluding the midrange items and including such items, we found that the item validities were higher when including the midrange items rather than deleting them for these two consensus methods. However, the scale validity was lower when including the midrange items than when excluding them. Accordingly, there was no consensus as to which strategy, excluding or including, achieved the best results for all types of validity.

Finally, this study provides interesting, albeit preliminary, findings regarding using an innovative integrated scoring approach. Furthermore, when scoring items with different characteristics on SJTs, it may be advisable to employ an integrated method to achieve higher item validity. Although of the two integrated methods, only one yielded higher criterion-related validity than that using a single scoring method, the integrated methods at least yielded higher validity than most of other methods (e.g. raw, standardized). Clealy, these results warrant further investigation. Furthermore, this integrated scoring strategy could take into account the practical issue that some subsamples of respondents likely make better judgments than other subsamples, or that some items achieve better consensus agreement than others. Our results suggest integrating suitable scoring methods according to item characteristics could reduce the weakness of any single method.

4.2. Practical implications

Any scoring method should aim to maximally rule out systematic error variance because an accurate self-evaluation is important for both organizations and individuals in making career decisions. This study suggests which scoring methods might be best used to this end. Therefore, our comparison of different SJT scoring methods provides important practical implications for individual assessment in the context of career guidance and selection.

First, career guidance centers might develop SJTs to help individuals self-assess themselves as to how they would perform in a set of representative work situations. Use of adequate scoring methods serves a useful purpose in providing more accurate information for career guidance. It helps career choosers identify aspects of their self that are stable and use those as anchors for re-charting their career paths and career management (Watson & McMahon, 2014; Weng & McElroy, 2010). In addition, the mere presentation of these situations to individuals also gives them a realistic preview of what to expect in a given job or occupation.

Second, for organizations that hire and promote people, our findings suggest that the scoring method matters for the validity of SJTs by bringing two new alternative methods (mode and proportion methods) into the spotlight. In addition, scoring methods that control for extreme response tendency effects will be less bias toward Black employees/applicants who tend to use more extreme responses, as well as permit organizations to minimize the coachability of SJTs. Cullen et al. (2006) found that people could increase their scores more than one standard deviation by simply avoiding endorsing extreme responses. In short, scoring strategies of SJTs provide valuable solutions to improve the validity of assessment, selection and career development.

4.3. Limitations and future directions

Some limitations should be acknowledged. First, though our study showed that the suggested four scoring methods and the integrated scoring methods yield higher criterion-related validity at both scale and item level, the differences we found were not significant. This is probably because of the small sample size we used in this study. Therefore, some caution is required in interpreting the results. Second, only one SJT was used in our study, which might limit the generalizability of our findings. Third, we used a 5-point Likert scale on the SJT and did not replicate the results with other types of Likert scales. As such, the validities of these new scoring methods must be examined in SJTs using an even-number Likert scale. Fourth, the data were collected in China, where the prevalence of extreme response tendency is lower because participants are more likely to use midpoints on the scales (see Chen, Lee, & Stevenson, 1995).

Our research raises several interesting questions that future research should address. First, the findings suggest that the using of standardized, dichotomous, mode, and proportion methods could decrease the influence of response tendencies on personal scores. Given that there exist racial/ethnic differences in extreme response tendency (Bachman, O'Malley, & Freedman-Doan, 2010), future research should examine which scoring method is most helpful to minimize the racial/ethnic differences of SJTs. Second, our study provides promising findings for the using of integrated scoring approach. In addition to the replication of different scoring methods in other SJTs, future research should consider more options for integrating different scoring methods. Third, our findings show that the mean item validly of the SJTs is relatively low. This suggests that the improvement of the item validity would be an important step to

improve the whole scale validity. For example, as the traditional SJTs provide multiple comparable responses for each scenario, participants would be motivated to choose the best behavioral choice among the options (should be) rather than the choice they tend to exhibit (would be), probably leading to low criterion-related item validity. Future research should consider solutions to control the comparability of the options of each scenario and to examine how such solutions would increase the criterion-related validity.

5. Conclusion

Our study introduces various alternatives to the SJT scoring literature, namely two scoring methods (i.e., mode and proportion consensus) and one integrated scoring approach. It also compares the criterion-related validity of the two introduced scoring methods with raw consensus, standardized, and dichotomous consensus scoring, and tests the performance of an integrated scoring strategy. We go beyond previous studies not only by demonstrating the mechanism as to how raw scoring method yields low criterion-related validity, but also by showing the advantages of both alternative scoring methods and an integrated scoring approach. Specifically, this study provides evidence suggesting that the scoring methods that fail to rule out the effect of extreme response tendencies lead to low criterion-related validity, which may lead to unfair and unintended discrimination effects against Blacks who tend to use more extreme responses in SJTs. Moreover, two newly introduced methods and the integrated scoring strategy yield higher scale validities than other typical SJT scoring methods. Thus, our findings offer meaningful theoretical and practical implications for understanding the effectiveness of scoring methods of SJTs as key assessment instruments in the context of career decision making and counseling. We hope that our results stimulate more interest in these important issues.

Acknowledgements

This work was supported by Natural Science Foundation of China (Project no. 71373251; no. 71422014).

Appendix A. Example scenarios and items of situational judgment tests

The following SJT items present respondents with work related situations and a list of plausible courses of action. You are asked to read each situation and evaluate the appropriateness of each of the actions on the 5 point scale (1 = extremely ineffective, 5 = extremely effective).

Some of the employees that you supervise spend much time gossiping and being critical of other employees.						
a.	Ignore the gossiping as long as the work is getting done.	1	2	3	4	5
b.	Have a group meeting to discuss the problems caused by gossip.	1	2	3	4	5
c.	Warn the employees to do their work and mind their own business.	1	2	3	4	5
d.	Give these employees more work to do so they will be too busy to gossip.	1	2	3	4	5
e.	Speak to each employee in private and tell them this behavior will not be tolerated.	1	2	3	4	5
f.	Circulate a memo warning of future punishment for gossiping.	1	2	3	4	5

References

Arthur, W., Jr., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology, 99*, 535–545. http://dx.doi.org/10.1037/A0035788.

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 (Retrieved from) http://www.monitoringthefuture.org/.

Barchard, K. A., Hensley, S., & Anderson, E. (2013). When proportion consensus scoring works. *Personality and Individual Differences, 55*, 14–18. http://dx.doi.org/10.1016/j.paid.2013.01.017.

Barchard, K. A., & Russell, J. A. (2006). Bias in consensus scoring, with examples from ability emotional intelligence tests. *Psicothema, 18*, 49–54.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14*, 223–235. http://dx.doi.org/10.1111/j.14682389.2006.00345.x.

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33*, 335–352. http://dx.doi.org/10.1177/0146621608329891.

Campion, M. C., Ployhart, R. E., & MacKenzie, W. I., Jr. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance, 27*, 283–310. http://dx.doi.org/10.1080/08959285.2014.929693.

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment, 20*, 333–346. http://dx.doi.org/10.1111/j.1468-2389.2012.00604.x.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143–159. http://dx.doi.org/10.1037//00219010.82.1.143.

Chen, C., Lee, S. Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science, 6*, 170–175. http://dx.doi.org/10.1111/j.1467-9280.1995.tb00327.x.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment test: Construct assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83–117. http://dx.doi.org/10.1111/j.1744-6570.2009.01163.x.

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

Crook, A. E., Beier, M. E., Cox, C. B., Kell, H. J., Hanks, A. R., & Motowidlo, S. J. (2011). Measuring relationships between personality, knowledge, and performance

using single-response situational judgment tests. *International Journal of Selection and Assessment, 19*, 363–373. http://dx.doi.org/10.1111/j.1468-2389.2011.00565.x.

Cullen, M. J., Sackett, P. R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14*, 142–155. http://dx.doi.org/10.1111/j.1468-2389.2006.00340.x.

Geher, G., Warner, R. M., & Brown, A. S. (2001). Predictive validity of the emotional accuracy research scale. *Intelligence, 29*, 373–388. http://dx.doi.org/10.1016/S0160-2896(00)00045-3.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70.

James, L. R., Demarree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*, 85–98. http://dx.doi.org/10.1037/0021-9010.69.1.85.

Jansen, P. G. W., & Vinkenburg, C. J. (2006). Predicting management career success from assessment center data: A longitudinal study. *Journal of Vocational Behavior, 68*(2), 253–266. http://dx.doi.org/10.1016/j.jvb.2005.07.004.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a Likert-based testing procedure. *Intelligence, 21*, 247–266. http://dx.doi.org/10.1016/01602896(95)90016-0.

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. R. (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.

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior, 45*(1), 79–122. http://dx.doi.org/10.1006/jvbe.1994.1027.

Lievens, F. (2013). Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests. *Medical Education, 47*(2), 182–189. http://dx.doi.org/10.1111/medu.12089.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90*, 442–452. http://dx.doi.org/10.1037/0021-9010.90.3442.

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology, 97*(2), 460–468. http://dx.doi.org/10.1037/a0025741.

MacCann, C., Roberts, R. D., Matthews, G., & Zeidner, M. (2004). Consensus scoring and empirical option weighting of performance-based emotional intelligence (EI) tests. *Personality and Individual Differences, 36*, 645–662. http://dx.doi.org/10.1016/S01918869(03)00123-5.

Mayer, J. D., Caruso, D. R., & Salovey, P. (2000). Selecting a measure of emotional intelligence: The case for ability scales. In R. Bar-On, & J. D. A. Parker (Eds.). *The handbook of emotional intelligence: Theory, development, assessment, and application at home, school, and in the workplace* (pp. 320–342). San Francisco, CA, US: Jossey-Bass.

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3*, 97–105. http://dx.doi.org/10.1037/1528-3542.3.1.97.

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. http://dx.doi.org/10.1037//0021-9010.86.4.730.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*, 103–113. http://dx.doi.org/10.1111/1468-2389.00167.

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. http://dx.doi.org/10.1037/A0021983.

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

Motowidlo, S. J., Martin, M. P., & Crook, A. E. (2013). Relations between personality, knowledge, and behavior in professional service encounters. *Journal of Applied Social Psychology, 43*, 1851–1861. http://dx.doi.org/10.1111/jasp.12137.

Putka, J. D., & Waugh, G. W. (2007). 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*.

Sacco, J., Schmidt, D., & Rogg, K. (2000). Using readability statistics and reading comprehension scores to predict situational judgment test performance, black-white differences, and validity. *Paper presented at the Annual Society for Industrial and Organizational Psychology Conference, New Orleans, LA*.

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.

Smith, K., & McDaniel, M. (1998). Criterion and construct validity evidence for a situational judgment measure. *Paper presented at the 13th annual conference of the Society for Industrial and Organizational Psychology*. Dallas: TX.

Tokar, D. M., & Fischer, A. R. (1998). More on RIASEC and the five factor model of personality: Direct assessment of Prediger's (1982) and Hogan's (1983) dimensions. *Journal of Vocational Behavior, 52*(2), 246–259. http://dx.doi.org/10.1006/jvbe.1997.1585.

Volodina, A., Nagy, G., & Köller, O. (2015). Success in the first phase of the vocational career: The role of cognitive and scholastic abilities, personality factors, and vocational interests. *Journal of Vocational Behavior, 91*, 11–22. http://dx.doi.org/10.1016/j.jvb.2015.08.009.

Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology, 52*, 1236–1247. http://dx.doi.org/10.1037//0022-3514.52.6.1236.

Watson, M., & McMahon, M. (2014). Making meaning of quantitative assessment in career counseling through a storytelling approach. In G. Arulmani, A. J. Bakshi, F. T. L. Leong, & A. G. Watts (Eds.). *Handbook of career development* (pp. 631–644). New York: Springer.

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

Weng, Q., & McElroy, J. C. (2010). Vocational self-concept crystallization as a mediator of the relationship between career self-management and job decision effectiveness. *Journal of Vocational Behavior, 76*(2), 234–243. http://dx.doi.org/10.1016/j.jvb.2009.10.012.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*, 188–202. http://dx.doi.org/10.1016/j.hrmr.2009.03.007.

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management, 17*, 601–617. http://dx.doi.org/10.1177/014920639101700305.