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.