Personnel Assessment and Decisions

Volume 6 | Issue 1 Article 1

2020

Situational Judgment Tests: An Overview of Development Practices and Psychometric Characteristics

Deborah L. Whetzel
Human Resources Research Organization (HumRRO)

Taylor S. Sullivan
Human Resources Research Organization (HumRRO)

Rodney A. McCloy
Human Resources Research Organization (HumRRO)

Situational Judgment Tests: An Overview of Development Practices and Psychometric Characteristics

Deborah L. Whetzel[1], Taylor S. Sullivan[1], and Rodney A. McCloy[1]

1. Human Resources Research Organization (HumRRO)

KEYWORDS

situational judgment test, selection method, best practices, psychometric characteristics

ABSTRACT

Situational judgment tests (SJTs) are popular assessment methods often used for personnel selection and promotion. SJTs present problem scenarios to examinees, who then evaluate each response option for addressing the issue described in the scenario. As guidance for practitioners and researchers alike, this paper provides experience- and evidence-based best practices for developing SJTs: writing scenarios and response options, creating response instructions, and selecting a response format. This review describes scoring options, including key stretching and within-person standardization. The authors also describe research on psychometric issues that affect SJTs, including reliability, validity, group differences, presentation modes, faking, and coaching.

Situational judgment tests (SJTs) assess individual judgment by presenting examinees with problem scenarios and a list of plausible response options. Examinees then evaluate each response option for addressing the problem described in the scenario. An example SJT item is shown in Figure 1.

SJTs have been used in employment testing for nearly a century (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). The first widely used SJT was the George Washington Social Intelligence test in which several solutions to each situation were offered in a multiple-choice format, only one of which was judged correct (Moss, 1926). During World War II, Army psychologists developed measures to assess soldiers' judgment. These assessments provided scenarios and alternative responses to each scenario (Northrop, 1989). In the late 1940s, several SJTs were developed to measure supervisory skills, including the *Supervisory Practices Test* (Bruce & Learner, 1958). In the 1950s and 1960s, large organizations used SJTs to predict managerial performance (Campbell, Dunnette, Lawler, & Weick, 1970).

A major resurgence in SJT research and use occurred when they were described as low-fidelity simulations by Motowidlo, Dunnette, and Carter (1990). Their seminal article described the process for developing SJTs—including analysis of critical incidents, generation of response options, and creation of a scoring key—and provided estimates of reliability, validity, and group differences. Reasons for the continued popularity of SJTs are that they (a) address job-related competencies that cannot be easily measured with traditional multiple-choice tests, (b) have useful levels of criterion-related validity (McDaniel, Hartman, Whetzel, & Grubb, 2007), (c) have construct validity as many SJTs assess leadership and interpersonal skills (Christian, Edwards, & Bradley, 2010), (d) have incremental validity over cognitive ability measures (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001), (e) have small to moderate group differences (Hough, Oswald, & Ployhart, 2001; Whetzel, McDaniel, & Nguyen, 2008), and (f) can be presented in a variety of media formats, including text-based, avatar-based, and video-based (Chan & Schmitt, 1997; Lievens, Buyse, & Sackett, 2005a; 2005b; Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000).

The literature also contains a number of narrative reviews of SJTs. Lievens, Peeters, and Schollaert (2008) discussed the psychometric characteristics of SJTs and briefly described a three-step process for how they are developed, including (a) collecting critical incidents of work situa-

Corresponding author:
Deborah L. Whetzel
Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 700 Alexandria, VA 22314-1591
Email: dwhetzel@humrro.org

FIGURE 1.
Sample SJT Item

You and a colleague from another department are jointly responsible for coordinating a project involving both departments. Your colleague is not completing an appropriate portion of the work. What should you do?

A. Ask your supervisor to discuss the problem with your colleague's supervisor.
B. Remind your colleague that the project will not be completed effectively without effort from both of you.
C. Tell your colleague that you will discuss the problem with your colleague's supervisor if your colleague refuses to work on the project.
D. Tell your colleague that nonparticipation creates more work for you and makes it harder to finish the project.
E. Ask someone else from your colleague's department to help with the project.

tions, (b) asking SMEs to generate response options for each situation, and (c) developing a scoring key. Whetzel and McDaniel (2009) reviewed the psychometric characteristics of SJTs and covered a variety of additional topics, including correlations with personality, incremental validity beyond cognitive ability, applicant reactions, coaching, and faking. Campion, Ployhart and MacKenzie (2014) reviewed the state of SJT research published since 1990. They created a taxonomy of SJT features and used that framework to summarize research using a content analysis. Corstjens, Lievens, and Krumm (2017) described the contextualization of SJTs and how that makes them similar to assessment centers and work samples but at a lower level of fidelity.

The important contributions of this review are that we (a) provide best-practice guidance (based on both empirical evidence and our experience) for developing SJTs and (b) update the review of the psychometric characteristics of SJTs. We base our guidance on the scientific literature, research we have conducted, and experience we have gained developing and implementing SJTs. We have presented both empirical research findings (Sullivan & Hughes, 2018) and best-practice suggestions (Sullivan, 2018a; 2018b; 2019; Sullivan & Woolever, 2019) at national conferences. We have applied the recommendations we describe in this article to a variety of high-stakes testing situations, working with both public- and private-sector clients. We acknowledge there may be differences of opinion regarding which development practices are most effective but aim to provide a useful starting point for those embarking on SJT research and development.

Components of Situational Judgment Tests

An SJT is a test format that is well suited for measuring constructs related to making judgments in challenging situations. An SJT item comprises two elements: a scenario that describes the situation and a set of plausible options for responding to the scenario. Other features of SJTs include response instructions and response format, both of which can affect constructs measured and scores obtained by examinees. Below are some best-practice guidelines to consider when designing various parts of an SJT.

Scenarios

Scenarios are often built using critical incidents[1] (Flanagan, 1954) of job performance collected from subject matter experts (SMEs), such as job incumbents, supervisors, and customers (Motowidlo, Dunnette, & Carter, 1990; Motowidlo, Hanson, & Crafts, 1997; Weekley, Ployhart & Holtz, 2006). Incidents typically describe the (a) situation or problem, (b) action taken to resolve the situation, and (c) outcome or results of the action. Once incidents are collected, the developer selects incidents to be edited to create item scenarios. The use of such incidents enables contextualization of scenarios so that fidelity is enhanced.

A second approach is to develop SJTs to reflect an underlying model or dimension. As an example, Stevens and Campion (1994, 1999) developed a Teamwork KSA test based on knowledge, skills, and abilities needed for teamwork as identified in an extensive literature review.

Regardless of whether incidents are used, developers need to consider the following characteristics of scenarios: specificity, brevity, sensitivity, complexity, and use of specific terminology. Regarding specificity, research has shown that more specific items had greater validity than relatively general items (Reynolds, Sydell, Scott, & Winter, 2000). We offer that more specific scenarios require fewer assumptions on the part of examinees regarding the meaning of the scenario, which leads to higher levels of validity.

Brevity is another concern when developing scenarios. Brief scenarios help reduce reading or cognitive load, which may serve to reduce group differences (Sacco, Scheu, Ryan,

1 Although they are typically called "critical incidents," the "critical" designation often leads developers to believe they have to be of extreme importance or have an extreme consequence to be considered "critical." Hence, some developers prefer the term "performance incidents" (Sullivan, 2018a).

& Schmitt, 2000; Sacco, Schmidt, & Rogg, 2000). As such, avoiding verbiage that is unnecessary for identifying the correct responses may increase validity and reduce group differences.

Sensitive or potentially offensive issues and topics should also be avoided when constructing scenarios. Test materials should not contain language, roles, situations, or contexts that could be considered offensive or demeaning to any population group. A test form or pool of items should generally be balanced in multicultural and gender representation, or neutral. Strategies to accomplish this are to ensure inclusion of culturally diverse passages and/or to ensure all passages depict themes applicable to all groups.

Regarding scenario complexity, there is a fine line between too simplistic and too complex. If scenarios are too simplistic (i.e., there is only one reasonable response that is obvious to most examinees), then it will be difficult to create plausible alternative courses of action, and there will be little variance on the SJT item. On the other hand, long, complex scenarios may assess construct-irrelevant abilities such as reading comprehension and working memory (per the point regarding scenario brevity above). Although complexity is important for obtaining variance in responses, excessively lengthy scenarios may introduce increased demands on reading comprehension, which is not the intent of the SJT (McDaniel & Whetzel, 2007).

Finally, based on our experience, we also suggest avoiding organization-specific terminology, policies, or processes that would not be understood by or known to external candidates. Although scenarios may be job relevant to enhance face validity, all job candidates should be able to identify effective and ineffective responses. On the other hand, if the SJT is being used to make promotion decisions among internal candidates, it may be appropriate to include organization-specific scenarios.

Response Options

Once the scenarios are created, SMEs are often asked open-ended questions about what they would do given the situation described in each scenario (Motowidlo et al., 1997). These responses, often collected in an in-person workshop or virtual setting, are used to create response options. Having job incumbents and/or supervisors provide this information helps ensure the options are plausible and not so ridiculous that no one would ever perform the behavior described. The goal is to provide a range of possible responses that vary in effectiveness.

Construct-based response options have been used by test developers where the options are developed to represent indicators of various constructs. For example, Trippe and Foti (2003) developed an SJT to measure conscientiousness, agreeableness, and openness, and their response options represented different levels of each trait. Ultimately, they concluded that method factors accounted for a larger portion of variance in the SJT items than in traditional personality-type items. Motowidlo, Diesch, and Jackson (2003) wrote response options to represent both high and low levels of agreeableness, extraversion, and conscientiousness. Their results showed that individuals' levels of these traits were positively and negatively related to endorsing response options designed to express high and low levels of the traits, respectively. A plausible explanation is that the effectiveness of various behaviors representing the same trait in response to the same scenario may be transparent to the examinee. An example item intending to measure conscientiousness is shown in Figure 2. As one reviews the response options, it is clear that option D is the answer one would select to maximize one's conscientiousness score. Thus, if designing a high-stakes SJT targeted to a particular construct, care needs to be taken that behaviors chosen to represent various effectiveness levels are not so transparent that examinees can easily detect "correct" answers.

As with scenarios, specificity is an important consideration for developing response options (Weekley et al., 2006). Response options need to be clear and concise. To enable the examinee to respond to a single option, it is advisable to list only one action in each response option (i.e., avoid double- or triple-barreled phrasing). In some situations, there might be several things that should be done. However, each response option should state a single course of action regarding what should be done in general or what should be done first. It is problematic for the examinee if an option lists multiple actions and the examinee agrees with some actions but not others.

As a practical matter, we have found it useful to (a) distinguish between "active bad" (do something ineffective) and "passive bad" (ignore; do nothing), and (b) not use both in the same item (active bad is typically worse than passive bad). Although there are times when doing nothing is the best response, it is rarely selected by examinees. To our knowledge, this issue has not been addressed in the literature and is a topic for future research.

Finally, tone of the response option is an important consideration. Words that suggest tone (e.g., adjectives) need to be considered carefully. For example, an option that states, "Sternly tell the customer you cannot provide assistance" may provide an indication that this is not an effective response for measuring customer service. Similarly, an option that states "Pleasantly inform the customer of available services" may provide a tipoff to examinees who may not otherwise know the best response.

Response Instructions

An SJT item typically asks either what examinees should do or what examinees would do in a given situation. Should-do instructions assess examinees' ability to apply knowledge to challenging situations, whereas would-do instructions assess examinees' behavioral tendencies (Mc-

FIGURE 2.
Sample SJT Item Designed to Measure Conscientiousness

You have been asked to give a presentation to the Board of Trustees at your organization. You have sufficient time and resources to prepare the presentation. What should you do?

A. Start preparing the presentation one hour in advance since you work best under pressure.
B. Start preparing the presentation two or three days in advance and practice it a few minutes before the presentation.
C. Prepare the presentation well in advance, carefully checking it for accuracy and practice just before the presentation.
D. Prepare the presentation well in advance, carefully checking for accuracy and practice several times so that it feels natural to talk in front of the Board.

Daniel & Nguyen, 2001; Nguyen, Biderman, & McDaniel, 2005; Ployhart & Ehrhart, 2003). Ployhart and Ehrhart (2003) examined the reliability and validity of six types of SJT instructions. Their results indicated that asking what one would do showed somewhat higher levels of reliability and validity than asking what one should do. This is an important finding, especially because they tested SJTs with identical content using a within-subjects design, however, their sample sizes ranged from 21 (should do) to 30 (would do). This makes their conclusions somewhat tenuous.

On the other hand, a meta-analysis (McDaniel, Hartman, Whetzel, & Grubb, 2007) comparing knowledge instructions (k = 96; N = 22,050) with behavioral tendency instructions (k = 22; N = 2,706) showed no evidence of a response instruction moderator for criterion-related validity (both estimated population validities were .26). However, when the content was held constant, knowledge instructions (k = 3; N = 341) yielded higher corrected validities than behavioral tendency instructions (k = 3; N = 290) (.26 vs. .12). Similar to the concern regarding Ployhart and Ehrhart's (2003) results, these findings are based on few effect sizes and are subject to second-order sampling error.

McDaniel et al. (2007) noted that almost all research on SJTs has been conducted using concurrent research designs involving job incumbents. As Weekley and Jones (1999) noted, there is reason to suspect that these findings might not generalize to applicant settings. It would be unlikely that applicants in high-stakes testing situations, given behavioral tendency instructions, would select an option other than the one that they believe to be the best (thus, displaying their knowledge). This concern was addressed by Lievens, Sackett, and Buyse (2009), who studied the two types of response instructions while holding content constant in a high-stakes testing context. They found that, consistent with previous research, SJTs with knowledge instructions correlated more highly with cognitive ability instructions than SJTs with behavioral tendency instructions and that there were no differences in validity between the two instruction sets. This suggests that in high-stakes testing situations, examinees respond to behavioral tendency instructions as if they were knowledge instructions. Thus, we recommend using knowledge instructions for high-stakes testing situations.

Response Format

There are three common SJT response formats: rate, rank, and select most/least (or best/worst). The *rate* format instructs respondents to rate each response option—usually on a 1- to 5- or 1- to 7-point Likert scale—for its effectiveness in responding to the scenario. The *rank* response format instructs respondents to rank-order the response options from most effective to least effective. The *most/least* response format instructs test takers to identify the most and least effective options.

Research has shown that the design of the response format shapes respondents' mental processing and subsequent response behavior. Ployhart's (2006) predictor response process model suggests that respondents engage in four processes when responding to SJT items: comprehension, retrieval, judgment, and response. When examinees use the rate response format, they complete this process for each response option independently. However, when examinees use rank or most/least formats, they make comparative judgments. These comparative judgments may require multiple iterations before examinees identify their final response. Greater numbers of response options require greater numbers of comparisons, and examinees need to distinguish among all response options. After completing this series of processes, examinees not only must remember their tentative judgments for each response option but also must decide on the relative effectiveness of each option to rank them or remember which they deemed most and least effective. When some options seem similar, the task is even more difficult.

Taken together, the predictor response process model suggests that rank-order and most/least response formats require comparatively higher levels of information processing than the rate format. Research shows that the rate format

generally tends to outperform the rank-order or most/least formats with respect to internal consistency reliability, test–retest reliability, incremental validity over cognitive ability, group differences, respondent reactions, and examinee completion time (Arthur et al., 2014; Ployhart & Ehrhart, 2003).

Care should be taken when writing options for the most/least format, because the options for each item must vary in effectiveness (i.e., the best needs to be better than next best and worst must be worse than next worst). On the other hand, options using the rate format can be similar or different in level of effectiveness, which offers more flexibility. Perhaps the biggest advantage of the rate format is that it supplies the maximal amount of information about the response options, given that all options receive a score, and thereby supports the largest number of scoring options. The most/least format yields scores for only two of the response options for any item/scenario.

Thus, there are a variety of advantages for the rate format. Practical constraints, however, may limit its use. Scoring rate format SJT items is more complicated than scoring most/least SJT items, which tend to be scored dichotomously. Also, common rate format scoring algorithms may present challenges for pre-equating test forms because of the post-hoc calculations required.

Single-response SJTs present an alternative format to traditional SJTs (Crook et al., 2011; Motowidlo, Crook, Kell, & Naemi, 2009). In these SJTs, each scenario is presented with a single response, and the examinee rates the response for effectiveness. Advantages to this method are that each item can be classified into a performance dimension and scoring is simplified. These SJTs have been shown to measure personality (Crook et al., 2011) and procedural knowledge (Motowidlo et al., 2009). Further, they have been shown to have internal consistency reliability and validity estimates comparable to other SJTs (Crook et al., 2011; Martin & Motowidlo, 2010). A potential disadvantage to this format is that each scenario has a single response, which may increase the amount of reading for examinees. It also may be more effort for item writers to create new scenarios than to create new response options for a single scenario.

Below, we discuss scoring considerations for SJTs. Our focus is on the rational approach to scoring, as well as two methods for rescaling: key stretching and within-person standardization.

Scoring of SJTs

Two primary features distinguish SJTs from other assessments. First, SJTs may not have an unambiguously "correct" answer because the situations are often complex and have multiple contingencies. Second, SJT scoring must account for this ambiguity by having "more correct" and "less correct" answers, rather than "right" and "wrong" answers. There are three basic approaches for developing an SJT scoring key (Weekley et al., 2006), and they resemble those applied to biodata (see Cucina, Caputo, Thibodeaux, & MacLane, 2012): empirical, theoretical, and rational. The empirical approach involves creating a key based on the relations between the incumbents' responses and a criterion, such as their job performance ratings. This approach is feasible only if one has a large number of incumbents on whom to collect criterion data. The theoretical approach uses a key based on what a theory would suggest is the "best" answer or what the appropriate effectiveness rating should be. Similar to the transparency concern about construct-based SJTs described above, this approach may lead to obvious best answers, which may make the method unsuitable for use in selection. The rational approach involves creating the key based on SME judgments regarding the effectiveness of response options.

Comparing these methods has been an ongoing research area (e.g., Krokos, Meade, Cantwell, Pond, & Wilson, 2004; MacLane, Barton, Holloway-Lundy, & Nickles, 2001; Paullin & Hanson, 2001). For example, Krokos et al. (2004) compared five empirical keying methods with a rationally derived key and found that only one of the empirical keys held up on cross-validation. MacLane et al. (2001) compared an empirical key with a rational key developed using a large group of federal government experts. They found that the two keys had similar levels of validity and provided further support for the conclusion that empirical keying offered no real advantages over rationally developed keys (Paullin & Hanson, 2001; Weekley & Jones, 1999). Bergman, Drasgow, Donovan, Henning, and Juraska (2006) found that rational and empirical keys developed for a computer-based, leadership SJT were both related to performance and provided incremental validity over cognitive ability and personality. Because rational keys are used far more frequently than either empirical or theory-based keys (Campion et al., 2014), the remainder of this section focuses on our suggestions regarding rational scoring key development.

Developing a Rational Scoring Key Using Consensus-Based Scoring

Developing a rational scoring key for an SJT involves several steps. First, it is important to develop "over-length" forms that include more scenarios and response options than will ultimately be needed. If the final SJT is to include about 40 situational items, then at least 50 to 80 problem situations should be prepared (McDaniel & Whetzel, 2007). When seeking to develop operational items with four to five response options, it is advisable to develop between 7 to 10 draft response options reflecting various levels of effectiveness. When developing response options for the most/least rating format, we have found it useful to have the item writer provide a rationale for why the best option is better

than the next best and why the worst option is worse than the next worst. Ultimately, the SMEs' effectiveness ratings are typically used to determine the scoring key, but the item writers' rationales often provide insight.

To create the key, SMEs rate the response options for effectiveness and/or select best/worst options. At this stage, the SMEs may also provide additional ratings on the SJT items (e.g., degree to which items/scenarios measure a target competency, job relatedness, needed at entry, fairness/sensitivity).

Next, descriptive statistics (e.g., means and standard deviations) are computed on the effectiveness ratings and are used to inform decisions about which scenarios and response options to retain and which to drop. The means provide an indication of effectiveness of the response option, and the developer should choose options for each scenario that vary in effectiveness. The standard deviations index expert judgment agreement. Response options for which there is little agreement among experts on effectiveness (i.e., have a high standard deviation) should be dropped (McDaniel & Whetzel, 2007). It is also appropriate to set thresholds for competency and/or job relatedness and/or needed at entry (if the SJT is to be used for entry-level selection), retaining only items that exceed the thresholds.

For the most/least rating format, the keyed response is the option rated most/least effective by SMEs and/or most frequently selected as best/worst. Additional constraints, such as requiring nonoverlapping confidence intervals between the most effective option and the second-most-effective option, and between the least effective option and the second-least-effective option, may also be used. Most/least items are then scored dichotomously based on whether an examinee selects the keyed response.

For SJTs using the rate format, the most basic scoring scheme involves computing the distance between examinees' responses and the key (i.e., the mathematical difference between an examinee's indicated effectiveness rating and the mean or median SME effectiveness rating). Research has shown that rate scoring formats are susceptible to coaching, because SME-keyed responses tend to cluster near the middle of the scale (Cullen, Sackett, & Lievens, 2006). To counter this effect, we describe key stretching and within-person standardization below.

Key stretching. Each consensus-based key has a ceiling and a floor because it is the average of SMEs' effectiveness ratings. That is, an item rarely has a keyed score of 1 or 7, because those values represent the end points of the rating scale. Thus, an examinee could get a reasonably good score by rating every option as 4 (the middle of the 7-point rating scale, thus leading to a maximum deviation from the keyed response of three points) or by avoiding using ratings of 1 or 7. This issue can be corrected by stretching the scoring key away from the midpoint. After computing the initial key using the SME mean ratings, the following formula can be used to stretch the key (Waugh & Russell, 2006):

StretchedKeyValue = ScaleMidpoint + StretchingCoefficient * (SMEMean - ScaleMidpoint)

As an example, for a 7-point scale where the midpoint is 4, we typically use a stretching coefficient of 1.5. If the SME mean is 2.0, it gets stretched to 1.0, as shown below.

StretchedKeyValue = 4 + 1.5 * (2 - 4) = 4 + 1.5 * (-2) = 4 - 3 = 1

At the other end of the scale, if the SME mean is 6.0, it gets stretched to 7.0, as shown below.

StretchedKeyValue = 4 + 1.5 * (6 - 4) = 4 + 1.5 * (2) = 4 + 3 = 7

Using a stretched key, it is possible for a response option to be keyed outside the scale range. For example, an option with an SME mean of 1.60 would be rescaled to a value of 0.40 using a stretching coefficient of 1.50. In that case, the rescaled value must be moved within the scale range. Thus, a rescaled value of 0.40 should be moved to 1.00. If several key values get stretched outside the scale range, this indicates that the stretching coefficient is too large. It is important to use the same stretching coefficient for all response options.

Another practice is to round rescaled key values to the nearest whole number. Although it is not necessary to round the scoring key values, it is easier to interpret scores based on integers rather than decimals. In some cases, however, rounding will reduce the validity of the scores by a small amount. Another option is to round the total score; this approach would not result in reduced validity.

Within-person standardization. With respect to SJT response patterns, previous research has defined "elevation" as the mean of the items for each participant and "scatter" as the magnitude of a participant's score deviations from their own mean (Cronbach & Gleser, 1953). McDaniel, Psotka, Legree, Yost, and Weekley (2011) suggested that elevation and scatter can be used to identify extreme or mid-scale response styles that can introduce criterion-irrelevant noise when using the effectiveness rating SJT response format. Distance scoring, commonly used to score rate format SJT responses, examines the difference (or match) between an examinee's responses and the SME mean. This approach does not account for elevation or scatter. However, by standardizing item responses within each examinee, the within-person standardization scoring method eliminates the influence of such individual differences in response styles. This is especially important because of the Black–White mean difference in the preference for extreme responses on Likert scales (Bachman & O'Malley, 1984). Research

shows that, on average, Blacks tend to use extreme rating points (e.g., 1 and 7 on a 7-point scale) more frequently than Whites (Dayton, Zhan, Sangl, Darby, & Moy, 2006) and Asians (Wang, Hempton, Dugan, & Komives, 2008). These differences have been found in multiple large, nationally representative samples in the United States (Bachman & O'Malley, 1984). As such, this adjustment substantially reduces Black–White mean score differences.

McDaniel et al. (2011) noted that item validity is shown to have a U-shaped relation with item means. This holds both for SJTs with Likert score response formats and for SJTs where respondents identify the best and worst response options. Given the U-shaped relation, they suggest dropping items with midrange item means. As such, these adjustments tend to simultaneously increase validity and reduce mean group differences.

Characteristics of Raters

There are a number of issues to consider when identifying raters. These include source (e.g., SMEs, psychologists, other knowledgeable individuals) and diversity of perspective. Regarding source, we typically solicit ratings from SMEs (i.e., typically job incumbents with more than 6 months of experience on the target job or supervisors of the target job). Another source of potential raters is psychologists or job analysts. Research has shown high correlations between incumbent and analyst ratings (Tsacoumis & Van Iddekinge, 2006). Using psychologists as raters is a good approach when developing a construct-based SJT that requires knowledge of psychological theory or concepts.

When identifying SMEs, we try to obtain diversity of SME experience, perspective, and demographics. Beyond job incumbents, other sources of raters include high-level leaders, customers, and trainers (Weekley et al., 2006). Experience in different departments or units may be useful depending on the use of the SJT. Optimally, we suggest that at least 10–12 SMEs rate each item. If fewer raters are used, outliers may have excessive influence on mean ratings. As noted above, SJT design features and development approaches influence the psychometric properties of the assessment. Arguably two of the most important psychometric features of any assessment are its reliability and validity, as discussed below.

Reliability

Reliability refers to consistency of measurement (Guion, 2011). Put simply, internal consistency (i.e., the extent to which items measure homogeneous or unidimensional construct) can be estimated using split-half reliability or Cronbach's coefficient alpha (Cronbach, 1949, 1951); consistency over time can estimated using test–retest reliability; equivalence across tests purporting to measure the same construct can be estimated using parallel forms reliability.

Most SJTs, by definition, are multidimensional. In fact, even a single item with different response options can measure different constructs (Whetzel & McDaniel, 2009). The scale and item heterogeneity make Cronbach's coefficient alpha, a measure of internal consistency reliability, an inappropriate reliability estimate for most SJTs. Test–retest reliability is a more appropriate reliability estimate for SJTs, but it is rarely reported in research and practice because it involves multiple test administrations. Parallel form reliability also is rare, because it requires the use of different item content to measure the same constructs. Given the difficulty of identifying constructs assessed using SJTs, construct equivalence across forms can be difficult to assess. Due to these test development and data collection limitations, many researchers continue to provide internal consistency estimates, even though they underestimate the reliability of SJTs.

Campion et al. (2014) conducted a content analysis of SJT research and noted the contradiction between (a) researchers stating that internal consistency reliability is inappropriate given that SJTs are multidimensional, and (b) nearly every published study on SJTs still reporting internal consistency reliability. In the empirical studies that have been published since 1990, they noted that reports of coefficient alpha (88.4%) far exceed those of test–retest (5.5%), parallel form (3.4%) and split-half (2.7%) reliability. Average reliabilities (and number of samples) were .57 (n = 129) for coefficient alpha, .61 (n = 8) for test–retest reliability, .52 (n = 5) for parallel form reliability, and .78 (n = 4) for split-half reliability. Assuming that reliability is appropriately estimated, there are two primary concerns with low levels of reliability. First, scores cannot be more valid than they are reliable. Second, when used operationally to set minimum standards, low levels of reliability are difficult to defend.

That said, Campion et al. (2014) drew several conclusions regarding the reliability of SJTs. First, theory-based SJTs tended to have higher reliability than other SJTs, possibly because this approach focuses on a single construct. Second, video-based SJTs had lower reliability than written or text-based SJTs, possibly because the richer nonverbal information from video SJTs may contribute to greater item-specific variance than the information available from written SJTs. Third, the rate response format had higher reliability than selecting most/least, possibly because all response options are scored, and thus there are a greater number of items using this format. Fourth, concurrent designs yielded higher levels of reliability than predictive designs, possibly because there is more error variance among applicants with no experience with the organizational context used to develop SJTs.

In summary, most researchers agree that using coeffi-

cient alpha to assess the reliability of SJTs is inappropriate due to the multidimensional nature of SJTs. At best, alpha is a lower bound estimate of reliability. However, because it is easy to calculate (it is available in most statistical packages), many researchers report this statistic rather than collect data needed for reporting more appropriate indices (e.g., split-half, test–retest). We recommend split-half estimates (corrected using the Spearman-Brown prophecy formula), assuming content is balanced. They require only one test administration and provide a more realistic estimate of reliability than alpha, given that it includes all split halves (some of which could be quite dissimilar with regard to construct coverage). However, we recognize that the reliability of SJTs is an issue that requires more research (Sorrel et al., 2016).

Validity

The key consideration in evaluating a selection procedure is that evidence be accumulated to support an inference of job relatedness. The *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2018) embrace the *Standards for Educational and Psychological Testing* definition of validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). As such, validity is a unitarian concept that results from the accumulation of evidence from multiple sources (e.g., SJT item ratings, sound item development procedures, job analysis data) regarding inferences that can be drawn from test scores. Below, we provide evidence of validity from criterion-related studies as well as from construct validation methods.

Criterion-Related Validity

Criterion-related validity refers to inferences made based on the statistical relationship between test scores and an external criterion (e.g., job performance). Meta-analytic estimates of validity cumulate results across studies quantitatively, typically resulting in a sample-size-weighted mean and variance (Oswald & McCloy, 2003). A meta-analytic estimate of the validity of SJTs in predicting job performance across 118 coefficients (N = 24,756) yielded a corrected estimate of .26. As noted above, McDaniel et al. (2007) found that should-do (knowledge) and would-do (behavioral tendency) instructions yielded the same levels of criterion-related validity. This finding is consistent with those of Lievens et al. (2009) who showed that in high-stakes situations, there was no difference between the criterion-related validity of the SJTs under both response instruction sets, likely because in high-stakes settings both become knowledge instructions.

Incremental validity of SJTs over cognitive ability and personality also has been a focus of research. Using three samples, Clevenger, Pereira, Wiechmann, Schmitt, and Harvey (2001) found that the SJT was a valid predictor of performance for all three samples and provided incremental validity over cognitive ability, job experience, and conscientiousness in two of the samples. McDaniel et al. (2007) conducted a meta-analysis and found that incremental validity estimates of the SJT over the Big Five factors of personality, and a composite of cognitive ability and the Big Five, ranged from .01 to .02. The incremental validity of g over SJT ranged from .08 to .10, and the incremental validity of the Big Five over SJT ranged from .02 to .03. McDaniel et al. (2007) noted that, although these observed incremental values are small, few predictors offer incremental prediction over an optimally weighted composite of six variables (i.e., cognitive ability and the Big Five).

As mentioned above, most published research has focused on the validity of SJTs in the selection arena, commonly reporting results of concurrent validation studies in which incumbent performance on the SJT correlates with their performance on the job. However, SJTs have been used in the context of both selection and promotion in military settings. The U.S. Army has conducted several studies involving the application of SJTs to the selection of officers (e.g., Russell & Tremble, 2011) and noncommissioned officers (e.g., Knapp, McCloy, & Heffner, 2004). These applications are more in line with the notion of using an SJT to inform promotion decisions. In the Army officer sample, an SJT designed to measure "leadership judgment" accounted for incremental variance beyond the Armed Forces Qualification Test (AFQT) and overall performance during Officer Candidate School (OCS). In addition, a study of noncommissioned officers provides strong support for the use of an SJT in a promotion context. For noncommissioned officer samples, SJT performance correlated significantly with observed performance ratings, expected performance ratings, a rating of senior NCO potential, and overall effectiveness ratings (Waugh, 2004).

Construct Validity

There has been considerable debate regarding the construct validity of SJTs. Researchers have had difficulty empirically identifying factors measured by SJTs, perhaps due to the overlapping nature of constructs generally assessed using SJTs. After decades of research that did not psychometrically identify constructs assessed by the SJT, Christian et. al. (2010) classified construct domains assessed by SJTs and conducted a meta-analysis to determine the criterion-related validity of each domain. They found that SJTs most often assess leadership and interpersonal skills, and those that measure teamwork and leadership have relatively high validity when predicting job performance.

Some argue that SJTs measure a single factor (e.g.,

general judgment), whereas others assert that SJTs measure distinct constructs (e.g., competencies). Arguments for a single judgment factor include a study that identified a single general factor from 12 distinct rationally derived factors (Oswald, Friede, Schmitt, Kim, & Ramsay, 2005). The authors developed alternate forms using an approach that incorporated items described as "rationally heterogeneous yet empirically homogeneous" (p. 149). In other words, the SJT content suggested that different factors were assessed, but factor analysis did not reveal specific dimensions. Two studies examined the General Factor of Personality (GFP) using a video-based SJT that measured social knowledge and skills (Van der Linden, Oostrom, Born, Molen, & Serlie, 2014). The first study, using assessment center participants, revealed that high GFP individuals were better able to indicate the appropriate social behaviors in an SJT. High GFP participants were rated higher by others on leadership skills. The second study, using psychology students, showed that GFP was related to the display of actual social behavior in a situational webcam test (a series of 10 short videotaped vignettes in which the respondent played the role of a supervisor). The authors concluded that their findings supported the idea that the GFP was revealed in their SJT.

The presence or absence of scenarios as part of an SJT suggested that SJTs measure a general domain (context-independent) knowledge (Krumm et al., 2015). Using a team knowledge SJT, there were no significant difference in scores for between 46% and 71% of items whether the situation (i.e., scenario) was presented or not. This was replicated across domains, samples, and response instructions. However, the situations were more useful when the items measured job knowledge and when response options denoted context-specific rules of action (which may not be appropriate for entry-level selection). This suggests that a general knowledge of how to act in various situations is being measured in SJTs that assess interpersonal skills.

The argument that SJTs measure multiple factors has been made using correlational data and meta-analyses. McDaniel et al. (2007) assessed construct saturation by correlating SJTs with cognitive ability and the Big Five. They found that SJTs measures cognitive ability (M_ρ = .32), Agreeableness (M_ρ = .25), Conscientiousness (M_ρ = .27), Emotional Stability (M_ρ = .22), Extraversion (M_ρ = .14), and Openness (M_ρ = .13).[2] We note that these correlations were moderated based on response instructions. SJTs that ask knowledge questions were more highly correlated with cognitive ability and SJTs that ask behavioral tendency instructions were more highly correlated with personality (McDaniel et al. 2007).

As can be seen from above review, the debate continues regarding the number and content of constructs measured by the SJT. Although SJTs may be written with the intention of measuring one or multiple constructs, factor analyses have not yielded solutions that reflected authors' intent (Schmitt & Chan, 2006). Not surprisingly, more research is needed.

Group Differences

Research shows that although SJTs exhibit group differences, they are lower in magnitude than those exhibited by cognitive ability measures, making SJTs an important predictor of performance. Whetzel et al. (2008) found that, on average, White examinees perform better on SJTs than Black (d = 0.38), Hispanic (d = 0.24), and Asian (d = 0.29) examinees. Female examinees perform slightly better than male (d = -0.11) examinees. This effect is influenced by the lower levels of reliability typical of most SJTs. That is, the lower reliability of the SJT reduces the magnitude of subgroup differences relative to those observed for traditional cognitive tests.

Whetzel et al. (2008) also found that knowledge (i.e., should-do) response instructions result in greater race differences than behavioral tendency (i.e., would-do) instructions. The mean correlations show these differences are largely because of the knowledge instructions' greater association with cognitive ability. The data for this study were based primarily on incumbent data and would likely underestimate group differences due to range restriction.

Roth, Bobko, and Buster (2013) investigated Black–White differences by collecting scale-level data from incumbents in four jobs. The SJT in their research served as the first major hurdle in a selection system, thus minimizing range restriction in their data. Results indicated that two cognitively saturated (i.e., knowledge-based) scales were associated with Black-White d values of 0.56 and 0.76 (Whites scored higher than Blacks), whereas items from three scales measuring constructs related to interpersonal skills were associated with Black–White d values of 0.07, 0.20, and 0.50. This suggests that SJTs that are cognitively loaded result in greater group differences than those that are noncognitive in nature.

The extent of group differences using the three response formats (ranking, rating and selecting most/least) also has been a topic of research (Arthur et al., 2014). Using an integrity-based SJT administered to 31,194 job candidates, Arthur et al. found that having candidates rate all SJT options yielded lower correlations with cognitive ability (and, consequently, smaller group differences) than having candidates rank order the SJT options or select most/least effective responses. This is consistent with the idea that the rank and most/least response formats likely require higher levels of information processing than the rate format (Ployhart, 2006).

In summary, research has shown that to the extent that

SJTs are correlated with cognitive ability, group differences are larger and, to the extent that SJTs are correlated with personality or interpersonal skills, group differences are smaller. As such, this research does point to strategies for reducing group differences: (a) measuring interpersonal skills (rather than knowledge), (b) using rating formats (rather than ranking or selecting best/worst), and (c) using within-person standardization (McDaniel et al., 2011), as described above.

Item Presentation Methods

There are numerous methods for presenting SJT items. These include text-based, video-based, and avatar-based methods. Research has shed light on differences among presentation modes. Chan and Schmitt (1997) conducted a laboratory experiment comparing text- and video-based SJTs, and found that a video-based SJT had significantly less adverse impact than a text-based SJT (perhaps due to reduced reading load) and that students perceived the video-based SJT to have more face validity than the text-based SJT. Similarly, Richman-Hirsch et al. (2000) found that students reacted more favorably to a multimedia format of an SJT measuring conflict resolution skills than to a written version of the same test. However, some have argued that video-based SJTs might insert irrelevant contextual information and unintentionally bring more error into SJTs (Weekley & Jones, 1997).

Lievens and Sackett (2006) studied the predictive validity of video- and text-based SJTs of the same content (interpersonal and communication skills) in a high-stakes testing environment (N = 1,159 took the video-based SJT; N = 1,750 took the text-based SJT). They found that the video-based SJT correlated less with cognitive ability (r = .11) than did the text-based version (r = .18). For predicting interpersonally oriented criteria, the video-based SJT had higher validity (r = .34) than the written version (r = .08).

Lievens, Buyse, and Sackett (2005a; 2005b) investigated the incremental validity of a video-based SJT over cognitive ability for making college admission decisions (N = 7,197). They found that when the criterion included both cognitive and interpersonal domains, the video-based SJT showed incremental validity over cognitively oriented measures for curricula that included interpersonal courses but not for other curricula.

Another presentation mode involves avatar-based SJTs. These are similar to video-based SJTs except that rather than having actors portray roles, computer-generated avatars interact with examinees. The use of avatar-based SJTs may be less costly than video-based SJTs because they are easier to edit (e.g., not requiring one to reshoot an entire video when changes are needed). Avatars can be two-dimensional (they may appear as cartoons) or three-dimensional (more human like). When creating three-dimensional avatars, one needs to consider the "uncanny valley" phenomenon. The uncanny valley occurs when computer-generated figures bear a close, but not exact, resemblance to human beings (MacDorman & Chattopadhyay, 2016). This elicits uncanny (or strangely familiar) feelings of eeriness or revulsion for the viewer. The term "valley" denotes a dip in the human observer's affinity for the replica. For this reason, many developers have opted to use two-dimensional avatars for SJTs.

In summary, video-based SJTs are likely the more costly alternative, but they also tend to have lower group differences than text-based SJTs. The costs may be reduced by using avatars, but at the time of this writing, we know of no research comparing avatar-based with video-based SJTs.

Faking

Faking on high-stakes selection measures has been defined as an individual's deliberate distortion of responses to achieve a higher score (McFarland & Ryan, 2000). Although there is some debate as to the effect of faking on validity (e.g., Ellingson, Smith, & Sackett, 2001; Schmidt & Ryan, 1992), most agree that faking affects the rank order of applicants and ultimately who is hired (Rosse, Stechner, Levin, & Miller, 1998).

Response instructions provided to examinees affect the extent to which SJTs are fakable. Nguyen et al. (2005) conducted a study in which 203 student participants indicated both the best and worst responses (i.e., knowledge), and the most likely and least likely responses (i.e., behavioral tendency) to each situation. Nguyen et al. also varied whether the students were asked to "fake good" first or respond honestly first. Using a within-subjects design, Nguyen et al. found that the faking effect size for the SJT behavioral tendency response format was 0.34 when participants responded first under honest instructions and 0.15 when they responded first under faking instructions. The knowledge response format results were inconsistent, probably because it is difficult to "fake" knowledge (i.e., either one knows the answer or one does not). They also found that knowledge SJT scores from the honest condition correlated more highly with cognitive ability (r = .56) than did behavioral tendency SJT scores (r = .38).

Peeters and Lievens (2005) studied the fakability of an SJT using college students. Their SJT comprised 23 items related to student issues (e.g., teamwork studying for exams, organizing, accomplishing assignments). Students were asked how they would respond (behavioral tendency instructions). Their results showed that students in the fake condition had significantly higher SJT scores than students

2 M_ρ is the estimated mean population correlation

in the honest condition. To assess whether the faking effect was practically significant, they computed the effect size, which was about one standard deviation (d = 0.89); women (d = 0.94) were better able to fake than men (d = 0.76). They also identified how many "fakers" were in the highest quartile to simulate the effect of a selection ratio of .25. They found that the highest quartile consisted of 76% fakers and 24% honest respondents. In contrast, the lowest quartile consisted of 31% fakers and 69% honest respondents. This shows that faking on an SJT has substantial effects on who would be selected when using behavioral tendency instructions in a low-stakes testing environment—a result likely to be magnified in a high-stakes environment.

In summary, when people fake, they probably do so in a selection context. SJTs with behavioral tendency instructions likely have limited validity, because job applicants are likely to respond as if knowledge instructions were provided. One possible remedy for faking is to use knowledge instructions rather than behavioral tendency instructions. Otherwise, the current literature has not pointed to a clear relation between SJTs and faking, although they appear to be less vulnerable than traditional personality measures (Hooper, Cullen, & Sackett, 2006).

Coaching

In high-stakes testing, examinees may seek the assistance of a coaching intervention, especially if the examinee obtained a fairly low score on a first attempt. Such coaching interventions range from practice on sample items to intensive instruction as part of commercial test coaching programs (Messick & Jungeblut, 1981). Cullen et al. (2006) tested two SJTs with different response formats: one using the best/worst format (Situational Judgment Inventory [SJI]) and one using the rate format (College Student Questionnaire [CSQ]). After coaching on response strategies (e.g., being organized, never taking the easy way out, avoiding aggressive displays in interpersonal disputes) using a video-based training program, results showed that the coaching program for the SJI was ineffective at raising SJI scores, but the coaching program for the CSQ was somewhat effective at raising CSQ scores. For the CSQ, Cullen et al. also tested a "scale" effect where they simulated scores by eliminating extreme responses. Results showed that if training had encouraged participants to use midpoints on the scale, their scores would have increased substantially (up to 1.57 standard deviations).

Lievens, Buyse, Sackett, and Connelly (2012) assessed the effects of commercial coaching on SJT scores as part of a high-stakes selection system for admission to medical school in Belgium. Researchers examined individuals who took the SJT and, having failed, took it again one month later. A subset of these individuals received commercial coaching. Results suggested that attending a commercial coaching program improved SJT scores greatly (d = 0.59) between the first and second examinations. The authors interpreted this as a large effect, as all "uncoached" candidates did use one or more self-preparatory activities. So, this difference can be considered the incremental effect of a formal coaching program over and above self-preparation strategies.

Stemming, Sackett, and Lievens (2015) examined the effect of coaching for medical school admissions. One surprising result was that the use of paid tutoring had a negative effect on SJT scores (d = -0.19). Attending information sessions at the university (d = 0.51) and completing the exercises in the official test brochure (d = 0.39) produced significant positive effects. The validity of the SJT in predicting GPA in interpersonal skills courses (r = .17) was slightly reduced (r = .15) in a model that controlled for the SJT coaching activities. Thus, the criterion-related validity of the SJT was not degraded by the availability of coaching.

To summarize, organizationally endorsed coaching (provided by information guides) may be more likely to result in increased SJT scores than coaching provided by test preparation organizations. However, if such coaching is taken by examinees who scored poorly on first taking an SJT, their scores may be improved with or without coaching, simply due to regression to the mean. Concerns about the potential unfairness of coaching can be countered by making effective coaching available to all examinees in the form of organizationally endorsed coaching. Scoring adjustments including key stretching (Waugh & Russell, 2006) and standardizing scores within person (McDaniel et al., 2011), as discussed earlier, can help mitigate these concerns but may not be appropriate under certain testing conditions (e.g., when pre-equating forms is required).

Summary

As with any selection method (e.g., job knowledge tests, assessment centers, interviews), there is a clear recognition that SJT quality is influenced by decisions regarding its design, development, and scoring. The research outlined above is intended to help assessment developers make these decisions. It is clear from both psychometric properties and examinee response behavior that not all SJT designs are the same, and not all designs may be appropriate for all intended uses and assessment goals.

SJT research continues apace, and there is doubtless more to learn regarding issues such as construct validity and the relative effectiveness of video-based and avatar-based SJTs. Our review of the literature, along with our experience researching and implementing SJTs for multiple clients in multiple contexts, suggests several guidelines and best practices (see Figure 3).

FIGURE 3.
Review of Best-Practice Guidelines

Scenarios	• Critical incidents enhance realism of scenarios. • Specific scenarios tend to yield higher levels of validity, because they require fewer assumptions on the part of the examinee. • Brief scenarios reduce candidate reading load, which may reduce group differences. • Avoid sensitive topics and balance diversity of characters. • Avoid overly simplistic scenarios that yield only one plausible response. • Avoid overly complex scenarios that provide more information than is needed to respond to the question.
Response options	• Ask SMEs for what they would do to ensure viability of response options. • Create response options that have a range of effectiveness levels for each scenario. • If developing a construct-based SJT, be careful about transparency of options. • List only one action in each response option (avoid double-barreled responses). • Distinguish between active bad (do something wrong) and passive bad (do nothing). • Check for tone (use of loaded words can give clues as to effectiveness).
Response instructions	• Use knowledge-based ("should do") instructions for high-stakes settings (candidates will respond to this question regardless of instruction). • Use behavioral tendency ("would do") instructions if assessing non-cognitive constructs (e.g., personality).
Response format	• Use the rate format where examines rate each option, as this method (a) provides the most information for a given scenario, (b) yields higher reliability, and (c) elicits the most favorable candidate reactions. • Single-response SJTs are easily classified into dimensions and have reliability and validity comparable to other SJTs, but they can have higher reading load given each scenario is associated with a single response.
Scoring	• Empirical and rational keys have similar levels of reliability and validity. • Rational keys based on SME input are used most often. • Develop "overlength" forms (more scenarios and options per scenario than you will need). • Use 10–12 raters with a diversity of perspective. Outliers may skew results if fewer raters are used. • Use means and standard deviations to select options (means will provide effectiveness levels; standard deviation will provide level of SME agreement).
Reliability	• Coefficient alpha (internal consistency) is not appropriate for multidimensional SJTs. • Use split-half, with Spearman-Brown correction, assuming content is balanced.
Validity	• Knowledge and behavioral tendency instructions have similar levels of validity. • SJTs have small incremental validity over cognitive ability and personality. • SJTs have been used in military settings for selection and promotion. • SJTs likely measure a general personality factor. • SJTs correlate with other constructs, such as cognitive ability and personality.
Group differences	• SJTs have smaller group differences than cognitive ability tests. • Women perform slightly better than men on SJTs. • Behavioral tendency instructions have smaller group differences than knowledge instructions. • Rate format has lower group differences than rank or select best/worst.
Presentation methods	• Avatar- and video-based SJTs have several advantages in terms of higher face and criterion-related validity, but they may have lower reliability. • Using avatars may be less costly, but developers should consider the uncanny valley effect when using three-dimensional human images.
Faking	• Faking does affect rank ordering of candidates and who is hired. • Faking is more of a problem with behavioral tendency (would-do) response instructions, especially in high-stakes situations. • SJTs generally appear less vulnerable to faking than traditional personality measures.
Coaching	• Examinees can be coached on how to maximize SJT responses. • Scoring adjustments (e.g., key stretching, within-person standardization) can reduce this effect.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Arthur Jr., W., 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. doi: 10.1037/a0035788

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black–White differences in response styles. Public Opinion Quarterly, 48, 491–509. doi:10.1086/268845

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

Bruce, M. M. & Learner, D. B. (1958). A supervisory practices test. Personnel Psychology, 11, 207-216. http://dx.doi.org/10.1111/j.1744-6570.1958.tb00015.x

Campbell, J. P., Dunnette, M. D., Lawler, E.E., & Weick, I. E. (1970). Managerial behavior, performance, and effectiveness. New York, NY; McGraw Hill.

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

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. doi: 0021-9010/97

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

Clevenger, J., Pereira, G. M., Wiechmann, D, Schmitt, N., & Harvey, V. (2001). Incremental validity of situational judgment tests. Journal of Applied Psychology, 86, 410–417. http://dx.doi.org/10.1037/0021-9010.86.3.410

Corstjens, J., Lievens, F., & Krumm, E. (2017). Situational judgment tests for selection. In H.W. Goldstein, E.D. Pulakos, J. Passmore, & C. Semedo (Eds.), The Wiley Blackwell handbook of the psychology of recruitment, selection, and employee retention (pp. 226–246). Oxford, UK: Wiley. https://doi.org/10.1002/9781118972472.ch11

Cronbach, L.J. (1949). Statistical methods applied to Rorschach scores: A review. Psychological Bulletin, 46, 393–429. http://dx.doi.org/10.1037/h0059467

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L.J., & Gleser, G.C. (1953). Assessing similarity 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. doi: 10.1111/j.1468-2389.2011.00565.x

Cucina, J.M., Caputo, P.M., Thibodeaux, H.F., & MacLane, C.N. (2012). Unlocking the key to biodata scoring: A comparison of empirical, rational, and hybrid approaches at different sample sizes. Personnel Psychology, 65, 385-428. https://doi.org/10.1111/j.1744-6570.2012.01244.x

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

Dayton, E., Zhan, C., Sangl, J., Darby, C., & Moy, E. (2006). Racial and ethnic differences in patient assessments of interactions with providers: Disparities or measurement biases? American Journal of Medical Quality, 21, 109–14, doi. org/10.1177/1062860605285164

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. Journal of Applied Psychology, 86, 122-133.

Guion, R.M. (2011). Assessment, measurement, and prediction for personnel decisions. New York, NY: Routledge, Taylor & Francis Group.

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358. http://dx.doi.org/10.1037/h0061470

Hooper, A. C., Cullen, M. J., & Sackett, P. R. (2006). Operational threats to the use of SJTs: faking, coaching, and retesting issues. In J. A. Weekley & R. E. Ployhart (Eds), Situational judgment tests: Theory, measurement, and application (pp. 205-232). Mahwah, NJ: Erlbaum.

Hough, L. M., Oswald, F. L. & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. International Journal of Selection and Assessment, 9, 152-194. doi: 10.1111/1468-2389.00171

Knapp, D. J., McCloy, R. A., & Heffner, T. S. (Eds.). (2004). Validation of measures designed to maximize 21st-century Army NCO performance (Technical Report 1145). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Krokos, K. J., Meade, A. W., Cantwell, A. R., Pond, S. B. & Wilson, M. A. (2004, April). Empirical keying of situational judgment tests: Rationale and some examples. Paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL

Krumm, S., Lievens, F., Huffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? Journal of Applied Psychology, 100, 399-416. doi: 10.1037/a0037674

Lievens, F., Buyse, T., & Sackett, P. R. (2005a). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. Jour-

nal of Applied Psychology, 90, 442-452. doi: 10.1037/0021-9010.90.3.442

Lievens, F., Buyse, T., & Sackett, P. R. (2005b). Retest effects in operational selection settings: Development and test of a framework. Personnel Psychology, 58, 981-1007. http://dx.doi.org/10.1111/j.1744-6570.2005.00713.x

Lievens, F., Buyse, T., Sackett, P. R., & Connelly, B. S. (2012). The effects of coaching on situational judgment tests in high-stakes selection. International Journal of Selection and Assessment, 20, 272-282. https://doi.org/10.1111/j.1468-2389.2012.00599.x

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. Personnel Review, 37, 426-441. https://doi.org/10.1108/00483480810877598

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. Journal of Applied Psychology, 91, 1181-8. doi: 10.1037/0021-9010.91.5.1181

Lievens, F., Sackett, P. R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. Journal of Applied Psychology, 94, 1095–1101. doi: 10.1037/a0014628

MacDorman, K. F., & Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. Cognition, 146, 190–205. doi:10.1016/j.cognition.2015.09.019

MacLane, C. N., Barton, M. G., Holloway-Lundy, A. E., & Nickles, B. J. (2001, April). Keeping score: Expert weights on situational judgment responses. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Martin, M. P., & Motowidlo, S. J. (2010, April). A single-response SJT for measuring procedural knowledge for human factors professionals. Poster session presented at the 25th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. Personnel Psychology, 60, 63-91. doi: 10.1111/j.1744-6570.2007.00065.x

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

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E.P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. Journal of Applied Psychology, 86, 730-40. doi: 10.1037//0021-9010.86.4.730

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. doi: 10.1037/a0021983

McDaniel, M. A., & Whetzel, D. L. (2007). Situational judgment tests. In D. L. Whetzel & G. R. Wheaton (Eds.), Applied measurement: Industrial psychology in human resources management (pp. 235-257), Mahwah, NJ: Erlbaum.

McFarland, L.A., & Ryan, A.M. (2000). Variance in faking across noncognitive measures. Journal of Applied Psychology, 85, 812-821. http://dx.doi.org/10.1037/0021-9010.85.5.812

Messick, S. & Jungeblut, A. (1981). Time and method in coaching for the SAT. Psychological Bulletin, 89, 191–216. http://dx.doi.org/10.1037/0033-2909.89.2.191

Moss, F. A. (1926). Do you know how to get along with people? Why some people get ahead in the world while others do not. Scientific American, 135, 26–27. doi: 10.1038/scientificamerican0726-26

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. Journal of Business Psychology, 24, 281–288. doi:10.1007/s10869-009-9106-4

Motowidlo, S. J., Diesch, A. C., & Jackson, H. L. (2003, April). Using the situational judgment format to measure personality characteristics. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

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., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D.L. Whetzel & G.R. Wheaton (Eds.). Applied measurement methods in industrial psychology (pp. 241-260). Palo Alto, CA: Davies-Black.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. International Journal of Selection and Assessment, 13, 250–260. doi: 10.1111/j.1468-2389.2005.00322.x

Northrop, L. C. (1989). The psychometric history of selected ability constructs. Washington, DC: U.S. Office of Personnel Management.

Oswald, F. L., Friede, A .J., Schmitt, N., Kim, B. K., & Ramsay, L. J. (2005). Extending a practical method for developing alternate test forms using independent sets of items. Organizational Research Methods, 8, 149-164. doi: 10.1177/1094428105275365

Oswald, F. L., & McCloy, R. A. (2003). Meta-analysis and the art of the average. In K. R. Murphy (Ed.), Validity generalization: A critical review (pp. 311-338). Mahwah, NJ: Erlbaum.

Paullin, C., & Hanson, M. A. (2001, April). Comparing the validity of rationally derived and empirically derived scoring keys for a situational judgment inventory. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. Educational and Psychological Measurement, 65, 70–89. doi: 10.1177/0013164404268672

Ployhart, R. E. (2006). The predictor response process model. In J. A. Weekley & R. E. Ployhart (Eds.), Situational judgment tests: Theory, measurement, and application (pp. 83–105). Mahwah, NJ: Erlbaum.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. International Journal of Selection and Assessment, 11, 1-16. doi: 10.1111/1468-2389.00222

Reynolds, D. H., Sydell, E. J., Scott, D. R., & Winter, J. L. (2000,

April). Factors affecting situational judgment test characteristics. Paper presented at the 15th Annual Conference of the Society of Industrial and Organizational Psychology, New Orleans, LA

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. Journal of Applied Psychology, 85, 880-887. doi: 10.1037/0021-9010.85.6.880

Rosse, J. G., Stechner, M. D., Levin, R. A., & Miller, J. L. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. Journal of Applied Psychology, 83, 634.644. http://dx.doi.org/10.1037/0021-9010.83.4.634

Roth, P. L., Bobko, P., & Buster, M. A. (2013). Situational judgment tests: The influence and importance of applicant status and target constructs on estimates of black-white subgroup differences. Journal of Occupational and Organizational Psychology, 86, 394-409. doi:10.1111/joop.12013

Russell, T. L., & Tremble, T. R. (2011). Development and validation of measures for selecting soldiers for the Officer Candidate School (Study Note 2011-02). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Sacco, J. M., Scheu, C., Ryan, A. M., & Schmitt, N. W. (2000, April). Understanding race differences on situational judgment tests using readability statistics. Paper presented at the 15th Annual Conference of the Society of Industrial and Organizational Psychology, New Orleans, LA.

Sacco, J.M., Schmidt, D., & Rogg, K.L. (2000, April). Using readability statistics and reading comprehension scores to predict situational judgment test performance, black-white differences, and validity. Paper presented at the 15th Annual Conference of the Society of Industrial and Organizational Psychology, New Orleans, LA.

Schmidt, M. J. & Ryan, A. M. (1992). Test taking dispositions: A missing link? Journal of Applied Psychology, 77, 629-637. http://dx.doi.org/10.1037/0021-9010.77.5.629

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

Society for Industrial and Organizational Psychology. (2018). Principles for the validation and use of personnel selection procedures. Bowling Green, OH: Author.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgment test scores: A new approach based on cognitive diagnosis models. Organizational Research Methods, 19, 506-532. doi: 10.1177/1094428116630065

Stemming, M. S., Sackett, P. R., & Lievens, F. (2015). Effects of organizationally endorsed coaching on performance and validity of situational judgment tests. International Journal of Selection and Assessment, 23, 174-181. doi: 10.1111/ijsa.12105

Stevens, M. J. & Campion, M. A. (1994). The knowledge, skill, and ability requirements for teamwork: Implication for human resource management. Journal of Management, 20, 503-530. doi: 10.1177/014920639402000210

Stevens, M. J. & Campion, M. A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. Journal of Management, 25, 207-228. doi: 10.1016/S0149-2063(99)80010-5

Sullivan, T. S. (2018a, April). What is the future of the critical incident technique? Panel discussion conducted at the 33rd Annual Conference of the Society for Industrial Organizational Psychology, Chicago, IL.

Sullivan, T. S. (2018b, July). So you want to learn how to develop a situational judgment test (SJT)? Tutorial conducted at the International Personnel Assessment Council Conference, Alexandria, VA.

Sullivan, T. S. (2019, April). In the trenches: Use of SJTs in high-stakes, high-volume testing programs. Panel discussion conducted at the 34th Annual C onference of the Society for Industrial and Organizational Psychology, National Harbor, MD.

Sullivan, T. S., & Hughes, M. G. (2018, April). Situational judgment tests: Debating response formats. In A. M. Harris, and M. G. Hughes, (Chairs), Of situations and responses: Unpacking the elements of situational judgement tests. Symposium presented at the Society for Industrial Organizational Psychology Conference, Chicago, IL.

Sullivan, T. S., & Woolever, N. (2019, March). Developing situational judgment tests: Not as frightening as you think! Presented at the Association of Test Publishers (ATP) Innovations in Testing Conference, Orlando FL.

Trippe, M. D. & Foti, R. J. (2003, April). An evaluation of the construct validity of situational judgment tests. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Tsacoumis, T. & Van Iddekinge, C. H. (2006). A comparison of incumbent and analyst ratings of O*NET skills (FR05-66). Alexandria, VA: Human Resources Research Organization.

Van der Linden, D., Oostrom, J. K., Born, M. Ph., van der Molen, H. T., & Serlie, A. W. (2014). Knowing what to do in social situations: The general factor of personality and performance on situational judgment tests. Journal of Personnel Psychology, 13, 107-115. doi: 10.1027/1866-5888/a000113

Wang, R., Hempton, B., Dugan, J. P., & Komives, S. R. (2008). Cultural differences: Why do Asians avoid extreme responses. Survey Practice, 1. doi:10.29115/SP-2008-0011; https://www.surveypractice.org/article/2913-cultural-differences-why-do-asians-avoid-extreme-responses

Waugh, G. W. (2004). Situational judgment test. In D. J. Knapp, R. A. McCloy, & T. S. Heffner (Eds.). Validation of measures designed to maximize 21st-century Army NCO performance (Technical Report 1145). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Waugh, G. W., & Russell, T. L. (2006, May). The effects of content and empirical parameters on the predictive validity of a situational judgment test. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. Personnel Psychology, 50, 25-49. doi: 10.1111/j.1744-6570.1997.tb00899.x

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. Personnel Psychology, 52, 679–700. https://doi.org/10.1111/j.1744-6570.1999.tb00176.x

Weekley, J. A., Ployhart, R. E. & Holtz, B. C. (2006). On the de-

velopment of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), Situational judgment tests: Theory, measurement, and application (pp. 157–182). Mahwah, NJ: Erlbaum.

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

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. Human Performance, 21, 291-309. doi: 10.1080/08959280802137820

RECEIVED 05/06/19 ACCEPTED 09/16/19