AFRL-RH-WP-TR-2019-0044

LITERATURE SUMMARY: BEST PRACTICES IN SITUATIONAL JUDGMENT TEST (SJT) DEVELOPMENT

Taylor S. Sullivan
Deborah L. Whetzel
Rodney A. McCloy

Human Resources Research Organization (HumRRO)

May 2019
Interim Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2019-0044 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//
THOMAS R. CARRETTA
Work Unit Manager
Supervisory Control and Cognition Branch
Warfighter Interface Division

//signature//
MAJ. JOSEPH C. PRICE
Chief, Supervisory Control and Cognition Branch
Warfighter Interface Division

//signature//
LOUISE A. CARTER
Chief, Warfighter Interface Division
Airman Systems Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
19-05-19	Interim	02 JAN 18 – 19 APR 19

4. TITLE AND SUBTITLE
Literature Summary: Best Practices in Situational Judgment Test (SJT) Development

5a. CONTRACT NUMBER
FA8650-14-D-6500, Task Order 0007

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Taylor S. Sullivan, Deborah L. Whetzel, Rodney A. McCloy

5d. PROJECT NUMBER
5329

5e. TASK NUMBER
09

5f. WORK UNIT NUMBER
H0SA (532909TC)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 700
Alexandria, VA 22314-1578

8. PERFORMING ORGANIZATION REPORT NUMBER
2018 No. 023

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate,
Warfighter Interface Division,
Supervisory Control & Cognition Branch
Wright-Patterson AFB, OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHCI

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2019-0044

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for public release

13. SUPPLEMENTARY NOTES
88ABW-2019-3573, cleared 5 August 2019

14. ABSTRACT

The U.S. Air Force is exploring the use of a situational judgment test (SJT) as a potential augmentation to the Weighted Airman Promotion System. SJTs are frequently used to assess more complex relational skills, such as interpersonal skills and leadership. SJTs present respondents with problem scenarios and a set of possible response options. Respondents then evaluate the effectiveness of the responses for addressing the problem described in the scenario. AFPC/DSYX developed a prototype SJT for use with E-7s and meant to assess People/Team Competencies listed in Air Force Doctrine Annex 1-1, Force Development. A variety of approaches have been advanced for developing and scoring SJT items, and there is ongoing debate in the academic literature and in practice as to which are most effective. This report summarizes research on various scoring approaches and propounds what the Human Resources Research Organization (HumRRO) considers to be best practice in SJT development. Included is a discussion of effective methods for developing test questions and scoring keys, as well as the process for reviewing and vetting items (both before and after initial administration).

15. SUBJECT TERMS
Competencies, interpersonal skills, leadership, scoring models, situational judgment test, SJT

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** (Monitor)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Thomas R. Carretta
Unclassified	Unclassified	Unclassified	SAR	25	**19b. TELEPHONE NUMBER** *(Include Area Code)* (937) 713-7143

TABLE OF CONTENTS

1.0 OVERVIEW

The Weighted Airman Promotion System (WAPS) determines promotions to the ranks of E-5 to E-9 within the U.S. Air Force (AF). The WAPS comprises a formula for weighting various components characterizing a person's readiness for promotion. Two standardized tests serve as WAPS components: (a) a Specialty Knowledge Test (SKT) – a measure of technical knowledge pertaining to the Air Force specialty (AFS) to which the individual belongs, and (b) the Promotion Fitness Exam (PFE) – a measure of general Air Force knowledge covering topics such as history, customs, resource management, dress and appearance, and security. SKTs are specific to each AFS, but the PFE is given to all members of a given rank, regardless of AFS. The AF is exploring the use of a situational judgment test (SJT) as a potential augmentation to the WAPS. SJTs are frequently used to assess more complex relational skills, such as interpersonal skills and leadership (e.g., Christian, Edwards, & Bradley, 2010). AFPC/DSYX developed a prototype SJT for use with E-7s and meant to assess People/Team Competencies listed in Air Force Doctrine Annex 1-1, Force Development.

SJTs present respondents with problem scenarios and a set of possible response options. Respondents then evaluate the effectiveness of the responses for addressing the problem described in the scenario. SJTs have been used in employment testing for almost a century (McDaniel, Morgeson, Finnegan, Campion & Braverman, 2001; Moss, 1926), and their popularity is on the rise. Reasons for this rising popularity are that they (a) address job-related competencies that cannot be easily measured with traditional multiple-choice test formats, (b) yield useful criterion-related validity, (c) have incremental validity over cognitive ability measures (McDaniel et al., 2001), and (d) have small to moderate subgroup differences (Hough, Oswald, & Ployhart, 2001; Whetzel, McDaniel, & Nguyen, 2008).

A variety of approaches have been advanced for developing and scoring SJT items, and there is ongoing debate in the academic literature and in practice as to which are most effective. The goal of this report is to summarize research on various scoring approaches and propound what the Human Resources Research Organization (HumRRO) considers to be best practice in SJT development. Included is a discussion of effective methods for developing test questions and scoring keys, as well as the process for reviewing and vetting items (both before and after initial administration). HumRRO has successfully followed the guidelines presented below for numerous clients (e.g., the Society for Human Resources Management [SHRM] and the Department of State) who work with us to develop SJTs for use in high-stakes testing situations.

2.0 WHAT IS A SITUATIONAL JUDGMENT TEST

A situational judgment test (SJT) is a test format that is best suited for measuring constructs related to making judgments in challenging situations. An SJT is arguably both a measure of a specific construct *and* a testing format. SJTs have been used to assess both knowledge and personality. The more an SJT measures knowledge, the more it will correlate with general intelligence; the more that it measures personality, the less it will correlate with intelligence. An SJT item comprises two elements: a *scenario*, which describes the situation, and several possible *actions*. These actions are also called the *response options*—or just *options*, for short. Like any assessment method, there is a clear recognition that SJT design influences SJT quality. Below are some evidence-based guidelines to consider when developing SJT scenarios and response options.

2.1 Guidelines for Developing Scenarios

- Focus on specific, challenging situations that require judgment. The situations should be faced by many incumbents (i.e., they should be representative).
- Ensure scenario description is realistic, clear, and concise.
- Avoid situations and terminology that cannot be understood by all incumbents.
- Keep most scenario descriptions fairly short: 4–6 lines in length or about 100 words. Tests with long scenarios can have only a few items. Long scenarios also tend to tap several dimensions and might tap abilities such as reading and working memory.
- Write the scenario to lend itself to numerous possible actions or responses that vary in effectiveness.
- Avoid scenarios where "get more information" is the best/most effective action. It is too obvious as a good action to take. It is acceptable to include "get more information" if it is a poor action to take—which is possible if there is not enough time to get more information.
- Be consistent with the tense and the actor (second person [you] vs. third person [they]).

2.2 Guidelines for Developing Response Options

- Ensure that your response options are clear and concise (one sentence is usually sufficient).
- List only <u>one</u> 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. To be concise, state what should be done in general or what should be done first. It is a dilemma for the examinee—and bad measurement practice—if an option lists multiple actions and the examinee agrees with some actions but not others.
- Include an appropriate amount of detail so examinees can evaluate effectiveness.
- Avoid options that are clearly not the best way to react.
- Avoid options that are tantamount to "get more information" or "do nothing."

- Distinguish between active bad (do something ineffective) and passive bad (ignore; do nothing), and do not use both in same item (active bad is typically worse than passive bad).
- Consolidate response options thoughtfully. Eliminate redundancies and overlap across response options. Determine the independent concepts in the set. Write one response for each concept. Drop the redundant responses.

3.0 RESPONSE INSTRUCTIONS

An SJT item typically asks either what examinees *should* do or what examinees *would* do in the situation. Should-do items assess examinees' ability to apply knowledge to challenging situations, whereas would-do items assess examinees' behavioral tendencies.

McDaniel, Hartman, Whetzel and Grubb (2007) showed that should-do (i.e., knowledge) instructions correlated more highly with cognitive ability and that would-do (i.e., behavioral tendency) instructions correlated more highly with personality. Important to note is that would-do SJTs can be faked just like other tests that measure personality constructs. Therefore, many organizations do not use would-do SJTs for selection or in other situations where there is motivation to fake. Faking is less of an issue with should-do SJTs, because examinees' responses measure whether they *know* what to do in a situation rather than what they would ostensibly *do*. If an examinee has no knowledge or experience related to the situation; however, the examinee might still respond based on personality.

McDaniel et al. (2007) found that should-do instructions had higher levels of criterion-related validity than would-do instructions. However, Lievens, Sackett, and Buyse (2009) 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.

4.0 RESPONSE FORMAT

Three common SJT response formats appear in the literature: rate, rank, and most/least. The ***rate*** format instructs respondents to rate each response option—usually on a 1- to 5- or 1- to 7-point Likert scale—in terms of its effectiveness as a response to the situation presented in the scenario (i.e., item stem). The ***most/least*** response format instructs test takers to identify the most and least effective options (sometimes presented as best/worst). The ***rank*** response format instructs respondents to rank-order the response options from most effective to least effective. Based on HumRRO's experience, the rate and most/least (i.e., best/worst) formats are more frequently used than the rank response format.

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 is relevant here. According to the model, respondents engage in four processes when responding to SJT items—comprehension, retrieval, judgment, and response—all of which are influenced by individual differences and SJT design features.

When examinees complete the rate response format, they complete the response process for each response option independently. However, when examinees complete the rank- or most/least format, they must make comparative judgments. These comparative judgments may require multiple iterations before examinees arrive at or generate a final response. After completing this series of processes, the examinee must not only remember their tentative judgments for each response option, but also decide on the relative effectiveness of each option to rank them or remember which they deemed most and least effective.

The complexity of the comparison process is further magnified by the number of response options an item has. In addition, because the rank and most/least response formats do not permit ties, they require examinees to distinguish between all response options. When some options seem similar, this requires even further consideration by the respondent.

Taken together, the predictor response process model suggests that rank and most/least response formats ultimately require comparatively higher levels of information processing than the rate format. Indeed, research confirms differences in how these items perform given the differential processes respondents are using when they respond. The rate format generally tends to outperform the rank or most/least with respect to internal consistency reliability, test-retest reliability, incremental validity over cognitive ability, subgroup differences, respondent reactions, and examinee completion time; however, the rate format is more vulnerable to response distortion (Arthur et al., 2014; Ployhart & Ehrhart, 2002; Waugh & Russell, 2005).

The reliability of rate items tends to be higher because there is one response (i.e., data point) per option, whereas in the most/least format there are only one or two responses per scenario. It is also harder to write options for the most/least format, because they must vary considerably in effectiveness within each item. Under the rate format, options can be similar (i.e., there can be ties) or different in effectiveness, which offers more flexibility. In addition, the rate format supplies the maximal amount of information about the response options, given that all options receive a score. The most/least format yields scores for only two of the response options for any item/scenario.

Thus, there are a variety of psychometric and practical advantages for the rate format. Practical constraints, however, may limit its use. For example, scoring rate format SJT items is more complicated than scoring most/least SJT items, which tend to be scored dichotomously. In addition, the scores on rate format SJTs tend to be less intuitive unless they are rescaled.

5.0 SCORING

Two primary features make SJTs unique from other assessments. First, SJTs may not have an unambiguously "correct" answer because the situations being presented 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. Therefore, we have to derive a way to determine the "keyed" or correct response to SJT items. Because of the inherent ambiguity in SJTs, there is unlikely to be perfect agreement regarding the optimal effectiveness rating or the most or least effective response options. Research has presented various scoring strategies that attempt to account for these complexities.
There are three basic approaches for developing an SJT scoring key (i.e., set of correct responses):

1. **Empirical:** The key is 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.
2. **Theoretical:** The key is based on what a theory would say is the "best" answer or what the appropriate effectiveness rating should be. This approach is rare for at least two reasons: (a) few SJTs have an underlying theory, and (b) this approach often leads to obvious best answers, which makes the method unsuitable for use in selection.
3. **Rational**: The key is defined by subject matter experts' (SME) judgments. This method is the most common and well-researched key-development strategy. Research suggests that rationally keyed SJTs perform at least as well as other scoring methods (see, for example, Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). Given these findings, the remainder of this section focuses on rational scoring key development.

5.1 Developing a Rational Scoring Key Using Consensus-Based Scoring

There are several steps to developing a rational scoring key for an SJT. First, it is important to develop "overlength" forms that include more scenarios and response options than ultimately needed. When seeking to develop operational items with 4 to 5 response options, we often develop between 7 to 10 draft response options reflecting various levels of effectiveness.

These overlength forms are then administered to SMEs. HumRRO often has SMEs rate all response options for effectiveness and select best/worst options. Although this might appear redundant, our data have shown that the "best" response is not always the option with the highest mean effectiveness rating, and likewise the "worst" response is not always the option with the lowest mean effectiveness rating. When developing items for the most/least (best/worst) rating format, HumRRO often asks SMEs to provide a rationale for why the best is better than then next-best and why the worst is worse than the next-worst. Ultimately, the SMEs' effectiveness ratings will be used to determine the "rightness" and "wrongness" of examinee ratings. At this stage, the SMEs are often asked to complete auxiliary ratings on the SJT items (e.g., degree to which items/scenarios measure a target competency, job relatedness).

Next, diagnostic statistics (e.g., mean, standard deviation, item-total correlations) are computed and used to inform decisions about which scenarios and response options to retain and which to

drop. It is also appropriate to set a threshold for competency- and/or job-relatedness, retaining only those items exceeding this threshold.

For the most/least rating format, the keyed response is the option rated most/least effective by SMEs and/or most frequently selected best/worst. HumRRO often imposes additional constraints, such as requiring non-overlapping confidence intervals between the "most" option and the second most effective option, and between the "least" option and the second least effective option. Most/least items are then scored dichotomously based on whether an examinee successfully 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 mean or median SME effectiveness rating). Research has shown that rate scoring formats are more susceptible to coaching, because SME-keyed responses tend to cluster near the middle of the scale (Cullen et al., 2006; see section on **Coaching** below).

5.2 Characteristics of Raters

Using SMEs is the most common way to develop the rational scoring key. There has not been much research on ideal characteristics of SMEs to include in the scoring key-development process. HumRRO applies several rules of thumb:

- At least 15 SMEs should rate each item, but more is better.
- Diversity in perspectives is good.
- SMEs should be individuals having operational experience in the SJT content and familiarity with the target population.

Rater sources other than SMEs have also been suggested in the literature:

- **Job incumbents/novices:** This is a good option when SMEs are not available (Legree, 1995). In HumRRO's experience, keys developed with incumbents tend to be very similar to those developed using SMEs. If using novice incumbents versus expert incumbents, it is advisable to have a larger number participating.
- **Psychologists:** This is a good approach when developing a construct-based SJT that requires knowledge of psychological theory or concepts.
- **Others:** Other types of raters have been suggested but not explicitly researched (to our knowledge). These include high-level leaders, customers, and trainers.

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. Below, we discuss research on these features with regard to SJTs.

6.0 SJT RELIABILITY

A meta-analysis conducted by McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001) found that internal consistency coefficients of SJT measures varied between .43 and .94. Ployhart and Ehrhart (2003) found that the type of response instructions influenced internal consistency. Asking examinees "to rate the effectiveness of each response" led to the highest internal consistency (.73). Asking examinees to choose two response alternatives ("Pick the best and worst response") led to somewhat lower internal consistency (.60), whereas response instructions wherein examinees had to choose only one response (e.g. "What is the most effective response?") had the lowest internal consistency (.24).

These findings notwithstanding, estimating the reliability of SJTs is difficult largely because SJTs typically assess multiple constructs and are often construct-heterogeneous at the item level. The scale and item heterogeneity make Cronbach's alpha an inappropriate reliability estimate (Cronbach, 1949, 1951). Test-retest reliability is a more appropriate reliability estimate for SJTs, but it is rarely reported in research and practice. Parallel form reliability also is rare, because it requires the use of different item content to measure the same constructs. Because it is difficult to identify particular constructs assessed using SJTs, construct equivalence across forms can be difficult to attain.

Due to these test development and data collection limitations, many researchers continue to provide internal consistency estimates with or without acknowledging that they underestimate the reliability of SJTs. Indeed, Campion, Ployhart, and MacKenzie (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%) 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. There are two primary concerns with these relatively 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.

In summary, most researchers agree that using coefficient 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., test-retest).

7.0 CRITERION-RELATED VALIDITY

McDaniel et al. (2007) provided a meta-analytic estimate of the validity of SJTs in predicting job performance. Across 118 coefficients (N = 24,756), their estimate was .26. They noted that the validity results were almost entirely based on concurrent validation studies in which the respondents were incumbents. To understand the possible effects of applicant variables (e.g., faking), predictive studies need to be conducted using applicants as subjects. (See section on **Faking** below.)

Clevenger, Pereira, Weichmann, Schmitt, and Harvey (2001) found that an SJT was a valid predictor of job performance in three independent samples, and incrementally so over job knowledge, cognitive ability, and conscientiousness in two of the three samples. These samples all used concurrent validation designs. Relative to the other predictors, the SJT's partial correlation with performance, controlling for the other predictors, was higher in most comparisons. McDaniel et al. (2007) estimated the incremental validity of SJTs over cognitive ability, the Big 5 factors of personality, and a composite of cognitive ability and the Big 5. SJTs offer incremental validity over a composite of cognitive ability and the Big 5 with incremental values ranging from .01 to .02. The specific response instructions (e.g., should-do versus would-do) did not appear to meaningfully moderate SJT's incremental validity. McDaniel et al. 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 5).

SJTs have been used in the context of both selection and promotion, and in particular, in military settings. Most published research has focused on their use and validity in the selection arena, commonly reporting results of concurrent validation studies in which incumbent performance on the SJT is correlated with their performance on the job (e.g., McDaniel et al., 2001; McDaniel et al., 2007). However, the U.S. Army has conducted several studies involving the application of SJTs to the selection of Officers (e.g., Russell and Tremble, 2011) and non-commissioned officers (e.g., Knapp, McCloy, and Heffner, 2004). These applications are more in line with the notion of using an SJT to inform promotion decisions.

In the Army officer (i.e., Select OCS) sample, an SJT designed to measure "leadership judgment" accounted for incremental variance beyond the Armed Forces Qualification Test (AFQT) in academic performance as well as in overall performance during Officer Candidate School (OCS). In addition, the non-commissioned officer (NCO21) study provides strong support for the use of an SJT in a promotion context. For both E5 and E6 soldiers, the SJT performance was significantly correlated with observed performance ratings, expected performance ratings, senior NCO potential rating, and overall effectiveness ratings.

8.0 CONSTRUCT VALIDITY

There has been considerable debate about determining the construct validity of SJTs. Researchers have had difficulty empirically identifying factors measured by SJTs (e.g., distinct competencies), perhaps due to the overlapping nature of constructs generally assessed using SJTs. Some argue that SJTs measure a single factor (e.g., general judgment), whereas others assert that SJTs are capable of measuring distinct constructs (e.g., competencies).

8.1 Single Factor

Oswald, Freide, Schmitt, Kim, and Ramsay (2005) identified a single general factor from 12 distinct rationally derived factors. They developed alternate forms using an approach that incorporated items that were "rationally heterogeneous yet empirically homogeneous" (p. 149). Van der Linden, Oostrom, Born, Molen, and Serlie (2014) conducted two studies that examined the General Factor of Personality (GFP) using a video-based SJT measuring social knowledge and skills. They found that high GFP individuals (N = 180 examinees of an assessment center) were better able to indicate the appropriate social behaviors in an SJT. High GFP participants were rated higher by others on leadership skills.

Krumm et al. (2015) suggested that SJTs measure a general domain (context-independent) knowledge. For between 43% and 71% of items, it did not matter whether the situation (i.e., stem) 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 would not be appropriate at the entry level). This suggests that a general knowledge of how to act in various situations is being measured in most SJTs.

8.2 Multiple Factors

McDaniel et al. (2007) assessed construct saturation by correlating SJTs with cognitive ability and the Big 5. They found that SJTs measure cognitive ability (M_ρ = .33–.46), Agreeableness (M_ρ = .27–.31), Conscientiousness (M_ρ = .25–.31), Emotional Stability (M_ρ = .26–.30), Extraversion (M_ρ = .30), and Openness (M_ρ = .13).[1]

Christian, Edwards, and Bradley (2010) found that to predict contextual performance, SJTs measuring interpersonal, teamwork, or leadership skills were more valid than SJTs that included heterogeneous composites. When predicting managerial performance, SJTs measuring interpersonal or leadership skills were more valid than SJTs that included heterogeneous composites. Although they noted that several meta-analyses had relatively few studies, they concluded that matching the criterion measured (contextual performance) to the SJT construct led to higher criterion-related validities than treating the SJT as a measurement method that assesses a wide variety of constructs.

[1] M_ρ is the estimated mean population correlation.

9.0 SUBGROUP DIFFERENCES

Whetzel, McDaniel, and Nguyen (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. In addition, research has shown 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 that these differences are largely because of the knowledge instructions' greater association with cognitive ability.

Roth, Bobko, and Buster (2013) collected scale-level data from four jobs in which SJTs were part of the first major hurdle of selection, thus providing an analysis of how constructs might relate to standardized White-Black group differences when range restriction concerns are minimized. Results indicated that cognitively saturated (i.e., knowledge-based) scales were associated with White-Black d values of 0.56 and 0.76 (Whites scored higher than Blacks), whereas items measuring constructs related to interpersonal skills were associated with White-Black d values of 0.07, 0.20, and 0.50.

Thus, findings have been somewhat mixed in terms of subgroup differences in SJT scores. HumRRO uses several strategies to mitigate potential subgroup differences. First, we conduct a bias and sensitivity review of all item content. During this activity, reviewers focus principally on three major issues:

- Test materials should not contain any language, roles, situations, or contexts that could reasonably be considered offensive or demeaning to any population group.
- A total 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 within each form and/or to ensure all passages depict universal themes applicable to all groups.
- No test material should contain elements extraneous to the job-related content and skills being assessed as part of the test specifications. Such extraneous material could provide an unfair advantage or disadvantage to population groups.

Second, we try to ensure that all scenarios and options are written clearly and concisely, and that they do not include advanced vocabulary. Reducing the reading load and reading difficulty helps to mitigate racial subgroup differences. In addition, because we are typically not attempting to measure vocabulary knowledge or reading skills with an SJT, this provides a purer measure of the constructs the SJT is targeting.

10.0 VIDEO VS. TEXT-BASED PRESENTATION

SJT presentation mode can vary (e.g., text, verbal, video). Research has shed light on some differences across presentation modes. Chan and Schmitt (1997) conducted a laboratory experiment comparing text- and video-based SJTs, finding 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, Olson-Buchanan, and Drasgow (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, Buyse and Sackett (2005b) examined 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.

Lievens and Sackett (2006a) also 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).

11.0 FAKING

Nguyen, Biderman, and McDaniel (2005) suggested that the response instructions provided to examinees affect the extent to which SJTs are fakable. In their study, 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. They also varied whether people were asked to "fake good" first or respond honestly first. Using a within-subjects design, they 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, interpersonal skills, social responsibility, perseverance, and integrity). Students were asked how they would respond (behavioral tendency instructions). Their results showed that students in the fake condition (N = 153) had significantly higher SJT scores than students in the honest condition (N = 138). To assess whether the faking effect was practically significant, they computed the effect size which was about one standard deviation (d = 0.89), with women (d = 0.94) being 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 76% of fakers and 24% of honest respondents were in the highest quartile. The lowest quartile consisted of 69% honest respondents and 31% fakers. This shows that faking on an SJT has substantial effects on who would be selected.

In summary, when people fake, and they probably do in a selection context, SJTs with behavioral tendency instructions likely have limited validity, because job examinees are likely to respond as if knowledge instructions were provided. One possible remedy for faking is to use knowledge instructions rather than behavioral tendency questions. 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).

12.0 COACHING

Cullen, Sackett and Lievens (2006) and Ramsey et al. (2003) 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), 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 mid-points 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 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. They computed Cohen's *d* for [(posttest coached – posttest control) – (pretest coached – pretest control)]/σ. 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 initial 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 (.17) was slightly reduced (.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. 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. Several adjustments can help account for this. The first involves "stretching" the scoring key (Waugh & Russell, 2005; discussed below). The second involves standardizing scores within person, which also reduces subgroup differences (McDaniel et al., 2011). Third, there is some evidence that using "should-do" instructions instead of "would-do" instructions helps reduce the effects of coaching.

12.1 Key Stretching

As noted above, the rate format SJT is particularly susceptible to coaching. The variability of an examinee's scores correlates highly (in a negative direction) with SME judgment scores. 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 a 4 (the middle of the rating scale) or by avoiding using ratings of “1” or “7” (Cullen, Sackett, & Lievens, 2004). For this reason, some researchers correct for this compression towards the scale midpoint 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, 2005):

StretchedKeyValue = ScaleMidpoint + StretchingCoefficient * (SmeMean - ScaleMidpoint)

It is possible for the stretched key for an option to be outside the scale range. For example, an original value of 1.60 is rescaled to 0.40 using a stretching coefficient of 1.50. In that case, move the rescaled value within the scale range. So, a rescaled value of 0.40 is moved to 1.00. If several key values get stretched outside the scale range, this is an indication that this practice is stretching the key too much. In that case, one should use a smaller stretching coefficient. The key is to use the same stretching coefficient for all response options.

Another typical 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 compared to decimals. In some cases, rounding will reduce the validity of the scores by a small amount.

12.2 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 his/her own mean (Cronbach & Gleser, 1953). McDaniel et al. (2011) suggested elevation and scatter reflect extreme or midscale response styles that can introduce criterion-irrelevant noise into 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 (i.e., creating a within-person *z* score for each examinee) and matching the aggregated within-person *z* scores with standardized mean SME ratings, the within-person standardization scoring method eliminates the influence of such individual differences in response styles.

13.0 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 made regarding its design, development, and scoring. The research outlined above has paved the way in helping assessment developers make these decisions, and it is clear from both psychometric properties and test-taker response behavior that not all SJT designs are the same, and not all designs may be appropriate for the intended use and assessment goals. The approach to SJT development and scoring ultimately depends on a variety of factors, including the assessment goals and end-user preferences, which is a testament to the extremely versatile, informative nature of SJT-based assessment.

Our review of the literature suggested the following guidelines and best practices that will not be appropriate for every SJT but that provide a good starting point for developers seeking to employ an SJT in their selection system.

- *Response Instructions*: Use should-do questions unless seeking to assess personality traits or other behavioral tendencies, where would-do questions are better suited.
- *Response Format*: Consider the ***rate*** format. This provides data (e.g., effectiveness ratings) for all response options rather than just, say, two (as in the ***most/least*** format). It also thereby permits the largest range of potential scoring options.
- *Scoring*: Rational scoring is the most feasible approach.
- *SJT Reliability*: Test-retest or alternate forms (if the situation permits it) reliability estimates are preferred to internal consistency estimates.
- *Method of Presentation*: Video-based SJTs have several advantages in terms of higher face and criterion-related validity.
- *Faking*: Prevarication is more of a problem with SJTs requesting would-do responses, but currently the measures do seem less vulnerable overall than traditional personality measures.
- *Coaching*: There is some evidence that responding to SJTs can be coached, although some researchers believe some scoring methods are likely less coachable (e.g., within-person standardization). More research on this topic is needed.

14.0 REFERENCES

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.

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.

Blair, C. A., Hoffman, B. J., & Ladd, R. T. (2016). Assessment centers vs. situational judgment tests: Longitudinal predictors of success. *Leadership & Organization Development Journal, 37,* 899-911.

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.

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.

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.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. Psychological Bulletin, 50(6), 456-473. http://dx.doi.org/10.1037/h0057173.

Clevenger, J., Pereira, G. M., Weichmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment test. *Journal of Applied Psychology, 86,* 410-417.

Cronbach, L. J. (1949). Statistical methods applied to Rorschach scores: A review. *Psychological Bulletin, 46,* 393−429.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297−334.

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.

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.

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.

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.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a Likert-based testing procedure. *Intelligence, 21,* 247-266. doi:10.1016/0160-2896(95)90016-0

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. *Journal of Applied Psychology, 90,* 442-452.

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.

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.

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advance-level high-stakes selection. *Journal of Applied Psychology, 96,* 927-940.

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

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.

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.

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.

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.

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.

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.

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.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89*, 187-208.

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.

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.

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.

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.

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.

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.

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

Waugh, G. W. & Russell, T. L. (2003). Predictor situational judgment test (PSJT). In D. J. Knapp (Ed.), *Select21 measure development progress report* (IR-03-74). Alexandria, VA: Human Resources Research Organization.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*, 25-49.

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