International Journal of Education & Management Studies, 2023, 13(4), 423-427
Peer-reviewed and Indexed Journal
https://iahrw.org/our-services/journals/international-journal-of-education-and-management-studies/

ISSN-p-2231-5632-e-2321-3671
NAAS Rating 4.58

# Situational Judgment Test: A Brief Overview

Deepesh Rathore
*Department of Psychology, Lakshmibai College, University of Delhi, Delhi*

Situational Judgment Tests (SJT) are used to assess a participant's response to a particular situation. The aim of this paper is to review the SJT as a method of measurement of work-related behavior. It examines the theoretical basis of SJT and psychometric properties in terms of reliability and validity. The steps involved in constructing SJTs have also been highlighted along with different response instructions that can be employed. Furthermore, its utility in analyzing the sub-group differences compared to cognitive and personality tests is also covered.

*Keywords:* situational judgment test, psychometric properties, work-related behaviors

Situational Judgment Tests (SJTs) are mainly used in work settings (Weekley & Polyhart, 2006) and have become popular among industrial and organizational (I/O) psychologists for the past two decades. One of the earliest known examples of SJTs is of United States Civil Services Exams dating as far back as 1873 (Dubois, 1970). In these tests' participants are given a situation and they are required to respond in an open-ended manner regarding their response in that particular situation.

One of the most widely used situational judgment tests used in the early part of the 20th is the George Washington University Social Intelligence Test (Hunt, 1928) in which the participants have to choose from close-ended responses. SJT use continued throughout World War 2 where the tests were used to assess the judgment of the soldiers (Northrop, 1989). Its usefulness was also explored in other areas to gauge the level of supervisory potential (Cardall, 1942; File, 1945) and later on the level of managerial potential of employees (Bruce & Learner, 1958; Kirkpatrick & Planty, 1960). For example, the early identification of the management potential used by the standard oil company (Campbell, Dunnette, Lawler, & Weick, 1970).

However, these tests were also criticized because they correlate too highly with intelligence tests which suggest that all they are measuring is the cognitive ability of participants (File & Remmers, 1971). One more criticism that was leveled against the SJTs was that their factor structure was too complex (Northrop, 1989). But in the 1990s interest regarding SJTs as a selection instrument increased after the publication of low-fidelity simulation tests (Motowidlo, Dunnette, & Carter, 1990). This test was developed for the selection of entry-level managers in the telecom industry, in which an employee had to select the most likely response as well as the least likely response for a given work situation among the five possible alternatives.

It was also established that SJTs have a high level of validity in predicting job performance (McDaniel, Hartman, Whetzel, & Grubb, 2007). The research has also shown that the SJTs are less biased when compared with other cognitive ability measures towards women and minorities (Whetzel, McDaniel, & Nguyen, 2008) and as the SJTs involve work-related situations, these tests have high face and content validity (Salgado et al., 2001).

## Underlying Theory

Two theoretical frameworks underlie SJTs, the first is that of *behavioral consistency* (Motowidlo, Dunnette, & Carter, 1990; Motowidlo, Hooper, & Jackson, 2006) which means that the past behavior acts as a guide for any behavior in the future. Second, is the *implicit trait theory* (Motowidlo et al., 2006). It focuses on the link between personality traits and behavioral effectiveness and states that the effectiveness of behavior is dependent upon individual differences in the personality traits that are exhibited in the behavior. For instance, if in a situation, a response that involves a high level of conscientiousness is the most effective response, then the person who has a high level of conscientiousness will opt for that response in greater numbers as compared to a person having a low level of conscientiousness.

## Subgroup Differences

The subgroup differences, both in terms of race and gender are important concerns when it comes to using a test as a selection tool. It has been found that the SJTs have lower subgroup differences as compared to cognitive ability measures. The difference in performance is expressed as standardized mean difference (d). The difference of one indicates that the difference in the score of a group is one standard deviation above the mean score of another group. The difference was obtained through SJTs between White/Black ($d$= 0.40), White/Hispanics ($d$= 0.37), and White/ Asian ($d$= 0.47). The difference although high is only half as much as obtained through cognitive ability measure between White/Black (d=0.99), White/Hispanics ($d$= 0.58), and White/Asian ($d$= -0.20), where the difference between Whites and Asians are in favor of Asian as they outperform the White participant in cognitive ability assessment. Interestingly, the difference obtained through video-based SJTs between White/Black participants is comparatively lower at $d$= 0.10, this difference can be attributed to the low loading of verbal ability in video-based SJTs (Ployhart & Holtz, 2008; Ployhart, Van Iddekinge, & MacKenzie, 2011).

Author Note

Dr. Deepesh Rathore, Assistant Professor, Department of Psychology Lakshmibai College, University of Delhi, Delhi
I have no known conflict of interest to disclose.
Correspondence concerning this article should be addressed to
Dr. Deepesh Rathore, Assistant Professor, Department of Psychology Lakshmibai College, University of Delhi, Delhi
E-mail: drathore55@gmail.com

On the other hand, the racial difference obtained through personality measures is almost half as that obtained through SJTs, White/Black (*d*= -0.04 to 0.20), White/Hispanic (*d*= -0.01 to 0.10), and White/Asian (*d*= 0.01 to 0.18) (Ployhart, Van Iddekinge, & MacKenzie, 2011).

As far as gender differences are concerned, female participants perform better than male participants, where the difference between Males/Females (*d*= -0.11) is in favor of females (Whetzel & McDaniel, 2009).

## *Steps Involved in the Development of SJTs*

As with other tests, item writing under the situational judgment test method involves focusing on three key elements, creating situations, creating responses, and scoring the responses.

*Step 1 Creating Situations:* The first step while developing an SJT is creating situations or scenarios to which the participants would respond. Situations for SJTs can be formed through critical incidents, Job analysis, and using a theoretical framework. First, let us look at the critical incident technique, according to Flanagan (1954), an incident is critical when it makes a significant contribution either positively or negatively to the activity and it should be capable of being analyzed or studied. Therefore, to write items a list of critical incidents or situations is pooled together, the critical incidents can be directly observed in the work environment or it can be written down by interviewing a supervisor or a senior manager (Weekley & Jones, 1999). Yet another way to note down critical incidents is by accessing the archival database. Job analysis data can also be used to construct scenarios, but is less frequently used (Aggarwal & Thakur, 2013).

Another approach is through theories, it is possible to identify the relevant job-related constructs. It is possible to capture the psychological demands, issues, challenges, and outcomes which can act as a basis to write SJTs items (Flanagan, 1954).

Stemler and Sternberg (2006) have proposed a broader framework to measure practical intelligence via SJTs, they have categorized the areas through which various situations can be formed. The areas include dealing with self, dealing with others, and dealing with tasks. Dealing with others has been further classified into dealing with supervisors, dealing with peers, and dealing with subordinates.

An area of concern with SJTs is the degree of complexity of the item stems (McDaniel & Nguyen, 2001). Items that are relatively simple, less lengthy, and do not contain multiple parts are preferred as they are easy to comprehend and also require less time to complete. Another related issue is that of the fidelity of the test item. Fidelity means how realistically an item captures the actual work-related situations, or any real-life situations (Hanson & Borman, 1989).

*Step 2 Creating Responses:* The next step is the generation of response options for a particular situation. The responses contain both effective as well as ineffective choices and are behavioral in nature. The responses can be developed by either the SMEs or SJT developers, studies have shown that responses generated by SMEs tend to be more effective and broader in range (Wagner & Stenberg, 1991). The responses to a particular situation can range from 3 to 12 responses.

*Step 3 Scoring the Responses:* The last step in the development of an SJT is to select an appropriate scoring mechanism. In general, there are two basic approaches to scoring the responses, one is the forced choice and the other is the Likert-type scale.

In the forced choice method, there is only one correct response and the rest of the distractors are incorrect. It is one of the simplest ways to score an SJT and has been applied across many studies (Stevens & Campion, 1999; Sinar et al., 2002; Hunter, 2003). One improvement over this scoring method was done by Mototwidlo et al. (1990) where the researcher opted for best/worst or most likely/least/likely response options for each situation. These methods are more effective as they give scores to options other than the best possible course of action, which in turn increases the variance associated with an item score (Weekley, Harding, Creglow, & Ployhart, 2004) which in turn has the potential to increase the validity of the SJT.

Another extension of this method allows the participant to rate all the responses on a continuum and then compare their raking with that of the given by the SMEs and then find the correlation using Pearson's rank order correlations (Weekley et al., 2004).

Yet another variation possible in this case is allowing the participants to rate the effectiveness of each response option on a scale of 1 to 6 and then comparing their response with that of SMEs and assigning values of 1, 2, and 3 depending upon the percentage of agreement (Chan & Schmitt, 2002). This scoring method has been found to explain incremental variance over both the cognitive ability construct as well as personality constructs (Chan & Schmitt, 2002).

In the case of Likert scale scoring, rating the effectiveness of each response option is one of the most widely used, SJTs using such response instructions are proposed to have incremental validity over cognitive ability, experience, and the big five personality traits (Chan & Schmitt, 2002). Also, since each response is rated more data points offer potential benefits in terms of reliability and validity.

## *Response Instructions*

Two response instructions are used in SJTs, knowledge and behavioral tendency. In the knowledge instruction, the participant is asked to rate the effectiveness of the responses given for a particular situation. It has been observed that knowledge response instruction correlates with cognitive ability tests (Lievens et al., 2009). Behavioral tendency instruction on the other hand asks the respondents to indicate what they would likely do in that situation. Studies have shown that SJTs having this type of response instruction correlates highly with personality tests (Olaru et al., 2019).

In other words, we can state that the knowledge response instruction acts as a measure of maximal performance, while the behavioral tendency response instruction can be said to be a measure of typical performance. Maximal performance measures have been used to assess ability while the typical performance measures have traditionally been to assess personality and interest (Cronbach, 1984). Job knowledge tests and work sample tests are examples of maximal performance measures as they ask the participants for the best possible or most effective response in a given situation, which tests the knowledge of the participants. A typical performance measure asks the participants how would they typically behave in that situation, which also leads to the flaws related to impression management and self-deception, while these two errors are not found in the maximal performance measure (Palhaus, 1984).

**Table 1**

*Characteristic Response Instruction for Both the Knowledge and Behavioral Tendency Response Format*

| Knowledge response instructions | Behavioral tendency instructions |
|---|---|
| Should do (Hanson, 1994) | Would do (Bruce, 1953) |
| Best response (Greenberg, 1963) | Most & least likely (Pulakos & Schmitt, 1996) |
| Best & worst response (Clevenger & Haaland, 2000) | Rate and rank what you would most likely do (Jagmin, 1985) |
| Rate effectiveness (Chan & Schmitt, 1997) | Rate the tendency to perform on a Likert scale (Doherty, 2005) |

*Note.* Adapted from McDaniel et al. (2007)

Although the different response format correlates highly with either cognitive or personality tests, the SJTs have shown incremental validity over both cognitive and personality constructs as measured by performing hierarchical linear regression (McDaniel, Hartman, Whetzel, &Grubb, 2007).

## *Reliability*

The *internal consistency coefficient* for SJTs ranges from 0.43 to 0.94 (McDaniel et al., 2001). It is argued that coefficient alpha is not a good measure of the reliability of multidimensional tests (Cronbach, 1951). Since SJTs are measurement methods that measure multiple factors it is not a good idea to rely on the internal consistency data to assess the reliability of SJTs.

Therefore, researchers agree that test-retest and parallel form are the more appropriate methods through which reliability should be calculated. The SJTs have a high *test-retest reliability* of 0.84 (Ployhart et al., 2004) other studies conducted on the test-retest reliability of SJTs have similarly found a significantly high-reliability coefficient ranging from 0.77 to 0.89 (Bruce & Leaner, 1958; Richardson, Bellows, Henry, & Co., 1981).

*Parallel form reliability* for SJTs is 0.76 (Chan & Schmitt, 2002). Parallel form reliability is rarely cited in research studies as it is a very difficult and time-consuming process to create parallel forms for the same constructs. Lievens and Sackett (2007) have proposed three different approaches to create parallel forms. First is the random assignment approach in which a large pool of items is created for every domain that tests measures and then randomly assign items to the alternate forms. Second is the incident isomorphism approach, in which a pair of items are created from the same critical incident which are then assigned to alternate forms. Third is the item isomorphism approach in which the same incident, context, and responses are used to form items but the items differ in wording and grammar.

Response instructions act as a moderating factor that influences the reliability coefficient of SJTs (Ployhart & Ehrhart, 2003). The reliability coefficient was highest for the response instruction "Rate the effectiveness of each response" at 0.73. the response instruction where the participants had to "choose the best and worst response" the reliability coefficient was 0.60. the reliability was lowest (0.24) in the case of response where the participant had to "choose the most effective response".

## *Validity*

Validity can be defined as “the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

*Construct Validity:* In a meta-analytic study conducted by McDaniel et al. (2007) the construct-related validity evidence for SJTs was analyzed, the result across many studies suggests that the response instructions act as a moderating factor. The knowledge response instructions which measure maximal performance are correlated more with cognitive ability measures and behavioral tendency instructions which measure typical performance tend to correlate more with the personality measure (Big 5). The correlation with

**Table 2**

*The Correlation Coefficient of Response Instructions with Cognitive Ability and Big 5 Personality Traits*

| Construct | Response instructions | Number of studies | Correlation coefficient (ρ) |
|---|---|---|---|
| Cognitive ability | Knowledge instruction | 69 | .35 |
| | Behavioral tendency instruction | 26 | .19 |
| Openness | Knowledge instruction | 11 | .14 |
| | Behavioral tendency instruction | 8 | .11 |
| Conscientiousness | Knowledge instruction | 38 | .24 |
| | Behavioral tendency instruction | 15 | .34 |
| Extroversion | Knowledge instruction | 14 | .15 |
| | Behavioral tendency instruction | 11 | .08 |
| Agreeableness | Knowledge instruction | 34 | .19 |
| | Behavioral tendency instruction | 17 | .37 |
| Neuroticism | Knowledge instruction | 33 | .12 |
| | Behavioral tendency instruction | 16 | .35 |

*Note.* Adapted from McDaniel et al. (2007)

cognitive ability construct for knowledge response instruction is 0.35 and for behavioral tendency instructions is 0.19. Similarly, the correlation with personality construct is more with behavioral tendency response instructions as compared to knowledge response instructions, with the exceptions for extroversion (0.15 vs 0.08) and openness to experience (0.14 vs 0.11). These results are expected as both these factors tend to correlate highly with cognitive ability.

The cognitive ability construct on the whole showed a correlation of 0.32 with SJTs and the Personality construct showed a correlation of 0.25 for agreeableness, 0.37 for conscientiousness, 0.22 for neuroticism, 0.14 for extroversion, and 0.13 for openness (McDaniel et al., 2007; Whetzel & McDaniel, 2009).

*Criterion-related Validity:* Criterion validity as measured by job performance indicators across 95 studies was measured to be 0.34 (McDaniel et al., 2001). In studies where job analysis was done for creating SJTs, the correlation coefficient was higher when compared with those where job analysis was not carried out (0.38 vs 0.29).

SJTs have also been found to have good validity in educational settings where they are used in the admission process (Oswald et al., 2004; Lievens et al., 2005).

*Incremental Validity:* The result of the meta-analysis conducted by McDaniel et al. (2007) showed that SJTs provide incremental validity over cognitive ability in the range of 0.03 to 0.05 (3% to 5%), while in the case of personality, the range is 0.06 to 0.07 (6% to 7%). Also, incremental validity over both the cognitive ability and personality ranges between 0.01 to 0.02 (1% to 2%).

## Discussion

The development and use of SJTs have a long history in the field of industrial and organizational psychology (Ployhart & MacKenzie, 2011). They have been used extensively as a selection method especially post-1990s (Campion, Ployhart, & Mackenzie, 2014). The SJTs have also been used to measure a wide variety of constructs, including employee integrity (Becker, 2005), leadership (Peus, Barun, & Frey, 2013), emotional intelligence (Austin, 2010; Chadha & Singh, 2006; Sharma et al., 2013); team role knowledge (Mumford et al., 2008); strategic knowledge of classroom management in elementary school (Gold & Holodynski, 2015); procedural knowledge (Motowidlo et al., 2009); personality (McDaniel & Wheztel, 2016); general aviation pilot judgment (Hunter, 2009); personal initiative (Bledow & Frese, 2009); medical education and training (Patterson, Zibarras, & Ashworth, 2016) to name a few.

SJTs have a high level of validity in predicting job performance (McDaniel, Hartman, Whetzel, & Grubb, 2007) and are less biased when compared with other cognitive ability measures towards women and minorities (Whetzel, McDaniel, & Nguyen, 2008). The reason why the use of SJT as a selection tool has increased is that SJTs offer incremental validity over cognitive ability as well as personality tests (Clevenger et al., 2001). Through incremental validity, we look for additional variance in the criterion, above and beyond that is explained by an already existing measure (Schmidt & Hunter, 1998). For example, SJTs help in predicting job performance beyond cognitive ability and personality (Chan & Schmitt, 2002; McDaniel et al., 2001; McDaniel et al., 2007).

Another area in which SJTs are now being used is to measure emotional intelligence (Austin, 2010) several tests are created based on situation-response format, these tests as in earlier ones give a situation and ask the participants to respond by selecting one of the multiple responses. The situations are usually in the form of interaction between employees, friends, and family members and involve different types of emotions and feelings (Chadha & Singh, 2006; Sharma, Gangopadhyay, Austin, & Mandal, 2013).

## Limitations of the Study

SJTs are paper pencil based, which makes it difficult to visualize some of the encounters described in the situations, although SJTs are known to have high fidelity or a high degree of realism compared to other tests, but they still lack in the representations of emotions present in the facial expressions, body language, tone of voice, etc. All such features are possible in a video-based SJT otherwise known as multi-media SJTs.

The test items describe scenarios and exchanges taking place between two or more employees, but as the technology is changing so is the workplace as well as the nature of work performed by employees. So, as a scale constructor, it means frequent revisions of the item stems to capture the reality of the situations at the workplace.

## Future Research

In order to better represent the situations described in the item stems and also to include elements of body language, gestures, and facial expressions, a storyboard with pictures can be created. The scale, therefore, would include images like the Thematic Apperception Test (TAT; Morgan & Murray, 1935). It would require more resources as compared to a simple Likert-type scale.

Next level could be a video involving real-life situations, including the elements of visual and audio, these types of tests are known as the Multi-Media Situational Judgement Test (Olson-Buchanan & Drasgow, 2006). These types of tests combine the elements of the pictures-based test described above, along with emotions embedded in tone of voice can also be brought forward, to aid the participants to make more informed choice regarding how to handle a particular situation described in item stem. It represents the maximum a researcher can hope to achieve via SJTs in order to make a situation clear to the participants.

## References

Aggarwal, A., & Thakur, G. S. M. (2013). Techniques of performance appraisal-a review. *International Journal of Engineering and Advanced Technology (IJEAT), 2*(3), 617-621.

Bruce, M. M., & Learner, D. B. (1958). A supervisory practice tests. *Personnel Psychology, 11,* 207-216.

Bruce, M.M. (1953). The prediction of effectiveness as a factory foreman. *Psychological Monographs: General and Applied, 67,* 1-17.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. (1970). *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill.

Cardall, A.J. (1942). *Preliminary manual for the test of practical judgment.* Chicago: Science Research.

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.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15,* 233−254.

Clevenger, J. P., & Haaland, D. E. (2000, April). *The relationship between job knowledge and situational judgment test performance.* Paper presented at the 15th Annual Conference of the Society for Industrial and Organization Psychology, Inc. New Orleans, LA.

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

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York, NY: Harper & Row.

DuBois, P. H. (1970). *A history of psychological testing.* Boston: Allyn & Bacon.

File, Q. W. (1945). The measurement of supervisory quality in industry. *Journal of Applied Psychology, 29,* 381-387.

File, Q. W., &. Remmers, H. H. (1971). *How supervise?* Manual 1971 revision. Cleveland, OH: Psychological Corporation.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-358.

Greenberg, S. H. (1963). *Supervisory judgment test manual (Technical Series No. 35).* Washington, DC: Personnel Measurement Research and Development Center, U.S. Civil Service Commission, Bureau of Programs and Standards, Standards Division.

Hanson, M. A. (1994). *Development and construct validation of a situational judgment test of supervisory effectiveness for first-line supervisors in the U.S. Army.* Unpublished doctoral dissertation, University of Minnesota.

Hanson, M.A., & Borman, W. C. (1989). *Development and construct validation of a situational judgment test of supervisory effectiveness for first-line supervisors in the U.S. Army.* Paper presented at the 4th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, Georgia.

Hunt, T. (1928). The measurement of social intelligence. *Journal of Applied Psychology, 12,* 317-334.

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *The International Journal of Aviation Psychology, 13,* 373-386.

Kirkpatrick, D. L., &. Planty, E. (1960). *Supervisory inventory on human relations.* Chicago: Science Research Associates.

Lievens, F., & Sackett, P.R. (2007). Situational judgment tests in high stakes settings: issues and strategies with generating alternate forms. *Journal of Applied Psychology, 92*(4), 1043-1055.

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

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. *The Journal of Applied Psychology, 94*(4), 1095-1101. https://doi.org/10.1037/a0014628

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.

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.

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.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91,* 749-761.

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

Olaru, G., Burrus, J., MacCann, C., Zaromb, F. M., Wilhelm, O., & Roberts, R. D. (2019). Situational Judgment Tests as a method for measuring personality: Development and validity evidence for a test of Dependability. *PloS One, 14*(2), e0211884. https://doi.org/10.1371/journal.pone.0211884

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.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598-609.

Ployhart, R. E., &: Holtz, B. c. (2008). The diversity validity dilemma: Strategies for reducing racio-ethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61,* 153-172.

Ployhart, R. E., Van Iddekinge, C. H., & MacKenzie, W. I., Jr. (2011). Acquiring and developing human capital in service contexts: The interconnectedness of human capital resources. *Academy of Management Journal, 54*(2), 353-368. http://dx.doi.org/10.5465/AMJ.2011.60263097

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), 1-16.

Ployhart, R.E., Porr, W., & Ryan, A.M. (2004). *New development in SJTs: Scoring, coaching and incremental validity.* Paper presented at the 19th annual convention of the Society for Industrial and Organizational Psychology, Chicago, IL, April.

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9,* 241-258.

Richardson, Bellows, Henry & Co. (1981). *Technical Reports Supervisory Profile Record, Richardson, Bellows.* Henry & Co., Washington, DC.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In N. R. Anderson, D. S. Ones, H. K. Sinangil, and C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology* (Vol. 1., pp. 165−199) London: Sage.

Sinar, E. F., Paquet, S. L., & Scott D. R. (2002). *Internet versus paper selection tests: Exploring comparability issues.* Paper presented at the 17th annual conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario, Canada.

Stemler, S. E., & Sternberg, R. J. (2006). Using situational judgment tests to measure practical intelligence. In J. A. Weekley and R. E. Ployhart (Eds.), *SIOP organizational series. Situational judgment tests: Theory, measurement, and application* (pp. 107-131). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

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.

Wagner, R. K., & Sternberg, R. J. (1991). *Tacit knowledge inventory for managers: User manual.* San Antonio, TX: The Psychological Corporation.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52,* 679−700.

Weekley, J. A., Harding, R, Creglow, A., & Ployhart, R. E. (2004). *Scoring situational judgment tests: Does the middle matter?* Paper presented at the 19th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Weekley, J.A., & Ployhart, R.E. (2006). *Situational judgment tests: Theory, measurement and application.* Jossey-Bass, San Francisco, CA.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19,* 188-202.

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. http://dx.doi.org/10.1037/0022-3514.46.3.598

Received November 8, 2023
Revision received December 2, 2023
Accepted December 6, 2023