

Situational Judgment Tests in Personnel Selection

Michael A. McDaniel, Deborah L. Whetzel¹ and Nhung T. Nguyen²

Virginia Commonwealth University & Work Skills First, Inc.

¹Human Resources Research Organization (HumRRO)

²Towson University

ABSTRACT

Situational judgment tests are designed to assess an applicant's judgment regarding a situation encountered in the work place. These tests are becoming increasingly popular due to their face validity and accumulated validity evidence. This chapter described current research and practice on situational judgment tests. Specifically, we focus on the needs of applied practitioners and researchers who are considering the use of these tests. The chapter begins with a description of the characteristics of situational judgment tests that will enable test developers to make informed decisions. Some of the characteristics affect test validity and the magnitude of demographic differences. These characteristics include test fidelity, stem length, complexity and comprehensibility, nested items, nature of responses, response instructions, and item heterogeneity. We then consider the psychometric characteristics of situational judgment tests including their construct validity, criterion-related validity, and demographic subgroup differences of SJTs. Finally we provide a process for building a situational judgment test. These steps include setting the boundaries for the content of the test, collecting and sorting critical incidents, creating items stems and item responses, selecting items response instructions and developing scoring keys.

INTRODUCTION

Situational judgment tests (SJTs) are designed to assess an applicant's judgment regarding a situation encountered in the work place (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). The above definition suggests that SJTs are similar to at least three other types of tests in which an applicant's work judgment is assessed. They are: job knowledge tests (Dye, Reck, & McDaniel, 1993), tacit knowledge tests (Sternberg, Forsythe, Hedlund, Horvath, Wagner, Williams, Snook, & Grigorenko, 2000),

and work sample tests (Callinan & Robertson, 2000). SJT items present respondents with work-related scenarios and a list of possible courses of action. Respondents are asked to evaluate the possible courses of action for either the likelihood they would perform the action or the effectiveness of the action. The assumption underlying SJTs is that how an individual performs on a job simulation predicts future job performance (Motowidlo, Dunnette, & Carter, 1990). An illustrative situational judgment item is presented in Figure 1.

<p>Everyone in your office has received a new computer except you. No one has said anything to you about this situation.</p>
<p>A. Assume it is a mistake and talk to your boss about the situation.</p>
<p>B. Take a new computer from a co-worker's desk</p>
<p>C. Confront your supervisor and ask why you are being treated unfairly.</p>
<p>D. Quit the job.</p>

Figure 1: Illustrative Situational Judgment Item

Research on SJTs indicates that these tests are useful and are becoming popular selection tools both in the U.S. and Europe (McDaniel et al., 2001; Salgado, Viswesvaran, & Ones, 2001). There are at least three reasons for the increasing popularity of SJTs. First, SJTs have been shown to have substantial validity as predictors of job performance (McDaniel et al., 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007). Second, SJTs have been demonstrated to have less race-based adverse impact than cognitive measures (Chan & Schmitt, 1997; Motowidlo & Tippins, 1993; Motowidlo et al., 1990; Whetzel, McDaniel, & Nguyen, 2008; Weekley & Jones, 1997, 1999). Third, because SJTs describe work-related situations, these measures are often viewed as having both face and content validity (Motowidlo, Hanson, & Crafts, 1997; Salgado et al., 2001). This chapter compliments previous reviews of SJTs (e.g., McDaniel & Nguyen, 2001; McDaniel et al., 2001; McDaniel et al., 2007) by describing eight characteristics along which SJTs vary and then offering guidance on how to develop SJTs.

Characteristics of Situational Judgment Tests

SJTs vary widely in their format. We believe it is useful to understand SJT characteristics for two reasons. First, knowing the range of formats for developing SJTs can enable developers to make informed decisions concerning how to build their SJT. Second, some of the SJT characteristics have validity and racial differences implications. The

distinguishing characteristics of SJTs that are discussed below are: test fidelity, stem length, stem complexity, stem comprehensibility, nested items, nature of response, response instructions, and item heterogeneity.

Test fidelity. Test fidelity concerns the extent to which the format of the stem is consistent with how the situation would be encountered in a work setting. As an example, a high fidelity test for a pilot would be a computerized simulation of flying a plane. A high fidelity version of SJTs may involve presenting the situations using a short video, whereas a low fidelity version would involve presenting the situations in a written format (paper and pencil or computer presentation of text). The distinction between video vs. written is a rough operationalization of fidelity. There are levels of fidelity within types of presentation. For example, a SJT may have more fidelity if the situation is described in some detail using vocabulary common to the job; a video SJT may have less fidelity if the video departs from aspects of the actual work situation. For example, if the actors in the video are college students, the video may not reflect a work setting in which incumbents are of varying ages.

It is likely that video-based SJTs reduce the reading and other cognitive demands of a SJT. Consequently, video-based SJTs will likely reduce SJT's correlations with cognitive ability as well as reduce mean subgroup differences on SJT performance compared to written-based SJTs. Nguyen, McDaniel, and Whetzel (2005) conducted a systematic review of the literature and found that video-based SJTs are less related to cognitive ability and have lower mean subgroup differences than written-based SJTs. Recently, Lievens, Buyse and Sackett (2005) examined the incremental validity of a video-based SJT over cognitive ability for making college admission decisions. 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. Further, Lievens and Sackett (2006) also studied the predictive validity of video-based and written SJTs of the same content (interpersonal and communication skills) in a high-stakes testing environment. They found that the video-based SJT had a lower correlation ($r = .11$) with cognitive ability than the written version ($r = .18$). For predicting interpersonally oriented criteria, the video-based SJT had higher validity ($r = .34$) than the written version ($r = .08$).

In sum, video-based SJTs show a high degree of promise, both in terms of face validity and incremental validity over cognitive ability for predicting performance in high-stakes settings, thus providing additional support for their use. Of course, one must weigh the cost of their development in the decision to use such tests. The cost of actors, videographers, studios, etc. may make this expense fairly prohibitive compared to traditional pencil and paper based SJTs.

Finally, although lower mean subgroup differences in video-based SJTs are an advantage, their reduced correlation with cognitive ability may result in lower validity.

Stem length. Stem length is another distinguishing feature of SJTs. Some stems are very short (e.g., *Everyone receives a new computer but you*). Other stems present very detailed descriptions of situations. For example the Tacit Knowledge Inventory (Wagner & Sternberg, 1991) contains relatively long stems. SJT items with short stems are useful in that one can administer more items in a fixed amount of time than SJT items with long stems. On the other hand, items with longer stems may incorporate more detailed information and may better serve some assessment goals such as the assessment of finding solutions to difficult and complex situations. Preliminary research comparing the criterion-related validity coefficients between short and long item stems showed no significant differences between the two (Friede, Imus, & Oswald, 2005), suggesting the advantage of short SJT stems.

Stem complexity. Stems also vary in the complexity of the situation presented. This characteristic is related to stem length described above. A stem of low complexity may be stated in a few words. Consider the example stem: *You have difficulty with a new assignment and need instructions*. This stem describes a relatively simple situation with clear possible responses. For example, the employee could seek assistance from a supervisor, a knowledgeable co-worker, or the employee could gain knowledge of the assignment from reading. In contrast, an example of a high complexity stem would be: *You have multiple supervisors who are not cooperating with each other, and who are providing conflicting instructions concerning which assignment has the highest priority*. This stem describes a complex situation in which potential responses also may be complex. It is important in SJT development not to confuse substantive stem complexity as described above with artificial stem complexity due to word redundancy.

Stem comprehensibility. Stem comprehensibility is another distinguishing feature of SJTs. It is more difficult to understand the meaning and import of some situations than others. Sacco, Scheu, Schmitt, Schmidt, and Rogg (2000) examined the comprehensibility of stems using a reading formula and then investigated the effect of reading level on subgroup differences and validity. In two studies, they found significant positive relationships between subgroup differences and the reading level of the situation. In one study, they found that reading level was positively associated with validity as well as subgroup differences. In both studies, the SJTs contained long, detailed situations followed by sub-situations. In another study, the SJT was constructed such that the items were less verbally complex. In this study, reading level was neither related to subgroup differences nor validity. This series of studies suggest that SJT format, especially presenting the situations with low levels of verbal complexity may alter the relationships between reading level and subgroup differences and validity.

McDaniel and Nguyen (2001) noted that the last three features (length, complexity and comprehensibility) of the situations are interrelated and probably drive the cognitive loading of the items. If a SJT characteristic affects the cognitive loading of the item, it is likely to have implications for both the item's mean racial differences and validities. These results suggest that there is a trade off between minimizing subgroup differences and maximizing validities. Although using highly valid SJTs is likely to result in larger subgroup differences, using SJTs with minimal subgroup differences is likely to result in lower validities. We will discuss this diversity-validity dilemma in detail in later paragraphs.

Nested items. Some SJTs present an opening paragraph describing an event within a company within which SJT items are embedded. For example, the opening paragraph might describe the need for a large training program to accompany the implementation of a new computer system. A follow-up SJT item might address challenges in finding trainers or challenges in scheduling the training. McDaniel and Nguyen (2001) called these items "nested" because the SJT items are nested under an opening paragraph. Aon Consulting used this format in some of their SJTs (Clevenger & Halland, 2000; Parker, Golden, Russell, & Redmond, 2000).

Nature of responses. The nature of the responses is another SJT characteristic. Unlike item stems that vary widely in format, item responses are usually presented in a written format and are relatively short. Even SJTs that use video to present the situation often present the responses in written form, sometimes accompanied by an audio presentation (e.g., Olson-Buchanan, Drasgow, Moberg, Mead, Keenan, & Donovan, 1998).

Response instructions. When presenting a SJT item, there are a variety of instructions that one can provide to the respondent. A two-dimensional taxonomy of common response instructions is shown in Figure 2. The first dimension has two categories: Behavioral Tendency and Knowledge. Behavioral Tendency instructions ask respondents to report how they typically respond. Knowledge instructions ask respondents to display their knowledge of the effectiveness of behavioral responses. The second dimension reflects the number of scoreable responses the item yields. Some response instructions yield one dichotomous response per item (e.g., *What would you mostly likely do?*). This permits only one scoreable response per item. Some response instructions yield two dichotomous responses per item (e.g., *What would you most likely do and what would you least likely do?*). This yields two scoreable responses per item. Some response instructions yield as many scoreable responses as there are response options (e.g. *Rate the response for effectiveness*).

	Number of Scoreable Responses		
	One scoreable response	Two scoreable responses	As many scoreable responses as response options
Behavioral Tendency	What would you most likely do?	What would you most likely do? What would you least likely do?	Rate each response for the likelihood that you would perform the response. Rank order the responses from the most likely to the least likely.
Knowledge	Pick the best answer.	Pick the best answer and then pick the worst answer.	Rate each response for effectiveness. Rank order the response from the best to the worst.

Figure 2: A Taxonomy of Response Instructions in Situational Judgment Tests

The choice of whether to use behavioral tendency response instructions or knowledge response instructions is an important one that likely affects:

- Applicant faking
- The magnitude of cognitive and non-cognitive correlates
- Criterion-related validity
- Magnitude of mean subgroup differences

McDaniel and Nguyen (2001) speculated that SJTs with knowledge instructions may be less fakeable than SJTs with behavioral tendency instructions. They argued that SJTs with knowledge instructions are like knowledge tests. Consider the question: *What is the cube root of 27?* One either knows the answer or one does not. A respondent can guess the answer but it is not fakeable.

In SJTs using knowledge instructions, respondents are asked to identify the correct answer (i.e., *What is the best action to take in this situation?*). As with a knowledge item, if the respondent does not know the answer, the respondent could guess the answer, but the respondent cannot fake the response. In contrast, SJTs with behavioral tendency instructions are similar to personality items in that they solicit self-reports of typical behavior. Consider the personality item: *How dependable are you?* An undependable person could respond honestly and report s/he is not dependable or s/he could lie and state that he is dependable. Thus, personality items can be easily faked. If a SJT item

asks how would one likely respond to a situation where project files are in disarray, a disorganized person might respond honestly and argue that she would leave the files in disarray, or the person might lie and argue that she would organize the files. Thus, like personality items, SJT items with behavioral tendency instructions should be relatively easy to fake. To our knowledge, only one study has addressed this issue empirically. Nguyen et al. (2005) found that respondents can fake a SJT with behavioral tendency instructions, but cannot meaningfully improve their scores through faking on the same SJT with knowledge instructions. Although one study cannot settle an issue, it appears reasonable that SJTs with knowledge instructions are faking resistant and SJTs with behavioral tendency instructions are more readily faked.

In addition to their potential impact on faking, the choice of response instructions affects the magnitude of cognitive and non-cognitive correlates of SJTs. McDaniel et al. (2007) found that SJTs with knowledge instructions are moderately correlated with cognitive ability and have lower correlations with personality. The opposite is true for SJTs with behavioral tendency instructions. Specifically, SJTs with behavioral tendency instructions tend to be moderately correlated with personality and have lower correlations with cognitive ability.

McDaniel et al. (2007) found no criterion-related validity differences between knowledge instruction SJTs and SJTs with behavioral tendency instructions. The estimated corrected mean validity for SJTs for both instructions was .26. Further, the non-zero lower 90th percentile values for both types of SJTs reported in that study indicate that the validity of SJTs generalize.

Whetzel et al. (2008) also found mean racial differences varied by response instructions. SJTs with knowledge response instructions yielded larger mean racial differences than SJTs with behavioral tendency instructions. They argued that the amount of mean racial differences in SJTs was controlled by the extent to which the SJT correlated with cognitive ability.

Thus, the decision to use knowledge instructions or behavioral tendency instructions is an important one. Relative to SJTs with behavioral tendency instructions, SJTs with knowledge instructions are likely to be faking resistant, more correlated with cognitive ability, less correlated with personality, have larger racial differences with the same criterion-related validity. One should consider these issues carefully and choose wisely.

Item heterogeneity. Item heterogeneity refers to the extent to which an item measures multiple constructs. SJT items tend to be construct heterogeneous at the item level and they likely vary in the extent of their heterogeneity. McDaniel & Nguyen (2001) and McDaniel et al. (2007) have found that SJTs are typically correlated with cognitive ability, agreeableness, conscientiousness, and emotional stability. For example,

the response to the item “*Everyone received a new computer but you*” could be correlated with cognitive ability or agreeableness or conscientiousness or emotional stability, or all four. Because it is difficult to define the construct measured by a particular item, it is probably best to think of SJTs as a measurement method in which multiple constructs are measured.

Our position concerning SJTs as a measurement method differs from that of Sternberg and colleagues (Sternberg et al, 2000). McDaniel et al. (2001) demonstrated that Sternberg and colleagues’ practical intelligence items are best classified as situational judgment tests. Sternberg (Sternberg et al. 2000) asserted that practical intelligence tests form a common factor, that is, they tend to measure one construct, unrelated to *g*. McDaniel and Whetzel (2005) and Gottfredson (2003) have reviewed the evidence on this matter and concluded that Sternberg’s assertion was not supported. Thus, we believe that it is best to view SJTs as measurement methods that can and do measure multiple constructs, including cognitive ability.

PSYCHOMETRIC CHARACTERISTICS OF SJTS

We have described factors that may be associated with the construct validity, criterion-related validity, and subgroup differences of SJTs. In the following paragraphs, we present a summary of empirical evidence addressing the psychometric characteristics of SJTs.

Reliability

The construct domains tapped by SJTs are multidimensional. Therefore, internal consistency reliability typically is not the appropriate estimate of reliability and it likely provides a downwardly biased estimate of true SJT reliability. Ployhart and Ehrhart (2003) attempted to construct a SJT in which the construct domain was relatively homogeneous. They obtained reliabilities ranging from .65 to .73 for SJT forms involving making ratings, .30 to .65 for SJT forms with two choices, and .24 to .65 for SJT forms with only one choice. Regarding test-retest reliability, estimates ranged from .20 to .92, likely due to small samples used in the study (*Ns* ranged from 21 to 30). McDaniel et al. (2001) reported a cumulative mean reliability of .77 across studies. This value is considered an underestimate due to the heterogeneity of the SJT items and because most of the reliability coefficients were based on internal consistency estimates. More appropriate estimates of reliability should be obtained through test-retest or alternate forms.

Two studies have sought to identify methods for constructing alternative forms of SJTs (Lievens & Sackett, 2007; Oswald, Friede, Schmitt, Kim & Ramsay, 2005). Lievens and Sackett (2007) examined three approaches for developing SJT items using item generation theory (Irvine & Kyllonen, 2002). The Random Assignment Strategy is used when a large pool of SJT items is developed for a particular construct and the

items are randomly assigned to alternative forms. The Incident Isomorphism Strategy is used when pairs of items are developed from the same critical incident (e.g., a physician dealing with a patient who refuses medication) and one item is assigned to each form. The Item Isomorphism strategy is used when multiple items reflect the same content domain, the same incident, and the same context of the item stem and responses. The only differences in items across forms are in wording and grammar. These strategies form a continuum of item similarity across forms with random assignment at the low end and item isomorphism at the high end.

The effects of these approaches on alternate forms reliability were studied in a high-stakes context (i.e., admission to medical college). Since only students who did not pass the test the first time participated in the retest, they corrected the observed correlations for indirect range restriction (i.e., candidates were selected on the basis of a third variable). For the domain of interpersonal/communication skills, corrected correlations were .34 for the random assignment strategy, .56 for the incident isomorphic strategy, and .68 for the item isomorphic strategy. These reliability estimates seem low compared to those achieved with cognitive ability measures in which the expected reliability estimates are .80 and above. However, because several assumptions of test-retest reliability were violated due to the high-stakes nature of the test (e.g., the fact that only those who failed the first time took the second test), these numbers are not considered test-retest reliability indices. Thus, as benchmark, they computed coefficients based only on those who took the test the second time after failing it the first time for general mental ability (GMA). None of the coefficients for GMA fell above .70. Only SJTs developed using the item isomorphic approach yielded values (.68) that were nearly equal to those of the GMA (.67), which seems appropriate given the nature of the high-stakes testing context.

Oswald et al. (2007) took a different approach. They combined items selected randomly from each of 12 content domains (e.g., the Random Assignment strategy described by Lievens & Sackett, 2007) to create large numbers of parallel forms. Parallel forms had to pass several criteria (e.g., the means had to be similar within $|d| \leq .05$, alpha reliabilities were to be at or above .70, and criterion-related validity with GPA was to be at or above .15). Further, they trimmed the outlying 20% of standard deviation values (10% of each tail) out of the distribution. Of the 10,000 alternative forms tested, 144 forms remained. This is an empirical approach assuming a large number of items are available.

In sum, since SJTs are largely measurement methods, it is very challenging to identify methods for developing alternate forms of SJTs. Given the multidimensional nature of SJTs, these methods of creating different kinds of items and alternate forms show great promise.

Construct Validity Evidence

Several primary studies have been conducted documenting the validity of SJTs (e.g., Chan & Schmitt, 1997; Motowidlo et al., 1990; Olson-Buchanan et al., 1998) as well as several meta-analyses (McDaniel et al., 2001; McDaniel et al., 2007; McDaniel & Nguyen, 2001). McDaniel et al. (2007) found that SJTs measure cognitive ability and the Big Five personality traits to varying degrees and the magnitude of the relationships is moderated by the SJT response instructions. They showed that SJTs with behavioral tendency instructions were more correlated with personality than SJTs with knowledge instructions (Agreeableness .37 vs. .19; Emotional Stability .35 vs. .12; Conscientiousness .34 vs. .24). In contrast, SJTs with knowledge instructions were more highly correlated with cognitive ability than SJTs with behavioral tendency instructions (.35 vs. .19).

In sum, the primary correlates with situational judgment tests are cognitive ability, agreeableness, conscientiousness, and emotional stability. These findings suggest that it may be possible to change the construct validity of a SJT by altering the response instructions. As mentioned above, the finding that SJTs have moderate correlates with personality and cognitive ability suggests that they are best viewed as measurement methods.

Criterion-Related Validity Evidence

McDaniel et al. (2001) conducted a meta-analysis to determine the criterion-related validity of SJTs. McDaniel et al. (2007) updated and reanalyzed the 2001 data. They showed that when the content of the SJT was not held constant, the criterion-related validity estimates were the same for SJTs with knowledge and behavioral tendency response instructions (.26). However, when the SJT was held content, knowledge SJTs emerged as having larger criterion-related validity (.26) compared to behavioral tendency SJTs (.12). The authors cautioned any affirmative conclusions from this finding due to the small number of samples ($k = 3$) on which the above meta-analytic validities were computed.

Incremental Validity Evidence

Several researchers have examined the incremental validity of situational judgment tests over measures of cognitive ability (Clevenger et al., 2001; Chan & Schmitt, 2002; O'Connell, Hartman, McDaniel, Grubb, & Lawrence, 2007; Weekley & Jones, 1997; 1999). Two meta-analyses of this topic also have been conducted (McDaniel et al., 2001; McDaniel et al., 2007). The research shows that SJTs provide incremental validity over cognitive ability. Because SJTs are measurement methods and can measure different constructs to varying degrees, the incremental validity of SJTs over cognitive ability will likely vary with the cognitive saturation of the SJT. SJTs that are highly correlated with cognitive ability may not have much incremental validity over cognitive ability.

Likewise, SJTs that measure non-cognitive job-related constructs might have useful levels of incremental validity over cognitive ability.

Little data exist on the incremental validity of SJTs over both cognitive ability and personality. O'Connell et al. (2007) noted the incremental validity of the SJT over cognitive validity but reported very little incremental validity over both cognitive validity and personality. However, Weekley and Ployhart (2005) reported that a SJT provided incremental validity beyond cognitive ability, personality, and experience. Clearly, more research is needed to have a definitive answer to this question.

Subgroup Differences

Whetzel et al. (2008) conducted the most comprehensive meta-analysis of subgroup differences in SJT performance. They found that race differences in SJT performance were largely explained by the cognitive loading of the SJT such that the larger the cognitive loading, the larger the mean race differences. Race differences in SJT performance was found to be a function of response instructions with knowledge instruction SJTs producing more mean race differences, and thus more adverse impact, than behavioral tendency SJTs. They also reported a small gender difference in SJT performance favoring women. Interestingly, they found little discernible effect of cognitive loading on male-female SJT performance differences. However, conscientiousness and agreeableness moderated the mean sex differences in SJT performance such that the higher the saturation of the above constructs, the larger the sex differences favoring women. Chan and Schmitt (1997) showed that video-based situational judgment tests produced less Black-White difference than the traditional paper-and-pencil format.

These studies suggest that there are characteristics of SJTs that can affect subgroup differences (e.g., video vs. paper-and-pencil administration and response instructions). These variables should be further investigated to determine causes for those differences (e.g., cognitive saturation) and how SJTs may be developed to reduce such differences.

Balancing Diversity – Validity using SJTs within the Legal Selection System in India

According to the Indian Constitution, employment discrimination is prohibited in public organizations per Article 16 (Premarajan, Thorton, & Padhi, 2008). Currently, no laws exist to protect Indian citizens from employment discrimination by private employers although there is a movement to pass legislations with support from the government to change this status quo. Whereas it is equally important to select a talented and diverse workforce, SJTs tend to produce subgroup differences that put minority applicants at a disadvantage when it comes to gaining employment opportunities. To decrease the magnitude of subgroup differences in SJT performance, one can reduce the cognitive loading of the test, however, the validities are lowered accordingly. Ployhart and Holtz

(2008) recently proposed sixteen strategies to address this diversity – validity dilemma (Pyburn, Ployhart, and Kravitz, 2008). The most effective strategy was to use alternative measurement methods such as SJTs because compared to traditional cognitive ability tests such as the Wonderlic Personnel Test, SJTs measure a variety of knowledge, skills, and abilities (KSAs), have higher face validity, and lower reading requirement (Ployhart & Holtz, 2008).

Although research on SJTs conducted in India is almost non-existent, there was some evidence of cognitive test performance differences in people of backward castes, scheduled castes, and minority groups (Kulkarni & Puhan, 1988 as cited in Premarajan et al., 2008). It is possible that SJTs will produce subgroup differences in India to the same extent such differences were reported in the U.S. We hope that this review will stimulate such research in India given the promising potential of SJTs as a valid selection tool.

HOW TO BUILD A SITUATIONAL JUDGMENT TEST

In this section of the paper, we describe frequently used methods for developing SJTs. There may be other useful methods for building SJTs, but these are the techniques that we have used successfully. The seven steps shown in Figure 3 are described below.

Step 1	Identify the job(s) for which SJT is being developed and set boundaries for content.
Step 2	Collect critical incidents
Step 3	Sort critical incidents
Step 4	Create item stems from critical incidents
Step 5	Generate item responses
Step 6	Select item response instructions
Step 7	Develop scoring key

Figure 3: Steps in Developing a Situational Judgment Test

Step 1. Identify the job(s) for which SJT is being developed and set boundaries for content. One needs to consider the kinds of jobs for which the SJT is being developed. For example, if the SJT is to be used for a class of jobs that contains both supervisors and non-supervisors, one needs to determine if there will be a separate test or supplemental items used only with applicants for supervisory positions. If the test is to be used for only for supervisors, will it be used for supervisors across content specialties (e.g., human resources, accounting, finance, information technology, etc.)?

Decisions need to be made regarding the level of technical content to be used in SJT items. One needs to decide if it is appropriate to include technical knowledge in SJT items. Using technical knowledge may limit the life span of items when there are changes in technology. For example, if the job involves information technology, and there are software features described in a SJT item, if the software changes, the item(s) will need to be updated. Likewise, a test developer who does not share the expertise with the subject matter expert (SME) will find it difficult to edit the item stem. However, SJTs can be a cost-effective method for assessing knowledge of complex technical issues compared to work samples.

Step 2: Collect critical incidents. We draw from Anderson and Wilson's (1997) work in presenting our discussion of critical incidents (Flanagan, 1954). A sample critical incident form is shown in Figure 4.

A critical incident includes three important pieces of information: 1) a description of the situation that led to the incident, 2) the actions or behaviors of the focal person in the incident, and 3) the results, or outcome of those actions. Given these three pieces of information, an interpretation as to the effectiveness of the actions can be made. The description of the situation is important because it helps the SJT developer understand the circumstances, anticipate certain actions, and understand why certain actions were or were not taken. It may include information such as the type of industry, type of job, specific tasks performed, environmental conditions, and relationship among others in the situation. Descriptions of the action are important because they describe the behavior of the focal person. Finally, descriptions of the outcome are important because they provide the basis for inferences as to the effectiveness of the behavior and the skills needed to enact the behavior. The form that SMEs use to write incidents should include prompts for the situation, the behavior, the outcome, the Knowledge, Skill, and Ability (KSA) or competency for which the incident is written, and a rating of the behavior's effectiveness.

Critical incidents often are collected in a workshop setting in which SMEs are asked to describe actual behaviors they have exhibited or observed others exhibit on the job. The remainder of the discussion in this section described procedures for conducting a critical incident workshop (Anderson & Wilson, 1997).

When conducting the workshop, one should provide the participants with plenty of room and privacy. When possible and accurate, the participants should be told that the critical incidents will be anonymous. These practices are recommended because critical incidents are often embarrassing to someone (e.g., "My supervisor make a bad decision..."). Privacy and anonymity permits such critical incidents to be disclosed.

Critical Incident	Participant # _____
1. What was the situation leading up to the behavior?	
2. What did the person do?	
3. What was the outcome or result of the person's action?	
4. Circle the number corresponding to the KSA or competency described in this incident:	
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15	
5. Circle the number below that best reflects the level of performance that the behavior exemplifies.	
<div style="display: flex; justify-content: space-around; font-size: 1.2em;"> 1 2 3 4 5 6 7 </div> <div style="display: flex; justify-content: space-around; margin-top: 10px;"> Highly Ineffective Moderately Effective Highly Effective </div>	

Figure 4: Sample Critical Incident Form

The leader of the workshop should also raise the comfort level of the participants. For example, many are embarrassed by the inaccuracy of their spelling or the quality of their writing. Thus, SMEs should be told that these issues are unimportant. Since many people are more comfortable typing on a computer than writing with a pen, it may be helpful to make computers available.

The first part of the workshop (about 30 minutes or so) should be used to train the SMEs on how to write critical incidents. During this training, the individual conducting the workshop should review the goals of the workshop, explain the format of critical incidents, explain some tips for writing a useable critical incident, and provide examples of both useable and unusable critical incidents. During this review and background discussion of critical incidents, participants should be encouraged to ask questions. When providing examples of incidents, it may be best to use an example that is not part of the job being analyzed because a job-relevant example may unduly narrow participants' focus. In other words, if an example incident for the job of Customer Service

Representative involved responding to phone calls, it is likely that a disproportionate number of incidents written in that workshop would involve responding, or not responding, to phone calls.

Some subject matter experts may have difficulty writing critical incidents and may need to be prompted or coached. A variety of prompts that can be used to coach SMEs into writing usable critical incidents are shown in Figure 5.

When multiple workshops are conducted, it may be useful to determine the competencies (i.e., KSAs) tapped by the incidents after the first workshop. The critical incident form provided in Figure 5 permits the respondents to identify the competencies relevant to the critical incident. After the initial workshop, one can tally the competencies addressed in the critical incidents and direct participants in future critical incident workshops to target the competencies that have been underrepresented by critical incidents collected at the prior workshop.

In the workshop, it should be emphasized that incidents should describe actions SMEs have seen a person perform, not what the SMEs inferred from the action about the skills or personal characteristics of the person. For example, rather than write that an individual “*displayed loyalty*,” the reports should describe what the individual did that displayed loyalty (e.g., *worked all night to finish a job, or defended the supervisor’s position to a group of subordinates*).

When SMEs start writing incidents, the SJT developer should encourage and reinforce them. The purpose is to shape their behavior so that they write productively. The incidents should be reviewed during the workshop and as they are being handed in to ensure compliance with instructions. If an incident does not contain important information (i.e., describes an individual knowledge, skill, or ability, rather than the behavior that occurred), one should probe the writer quietly for more detail about the behavior that occurred. If a really good incident is written at the beginning of the workshop, the workshop leader may ask the writer if the incident can be read aloud to provide other SMEs with an example of a well written incident. It is important to ensure the privacy of the item writer, especially if the incident describes ineffective behavior—the person about whom the incident is written may be in the workshop. Also, well written incidents can describe both effective and ineffective behavior. Since many individuals hesitate to write, especially in a group setting, small editorial changes should be ignored in the workshop. These changes can be made after the workshop. Although the number of incidents written by each SME will vary, it is reasonable to expect that an average of 5-10 critical incident reports can be generated by each SME in a two-hour workshop.

- Think about a time when someone did a really good job
- Think about a time when someone could have done something differently.
- Think of a recent work challenge you faced and how you handled it
- Think of something you did in the past that you were proud of.
- Think of a time when you learned something the hard way. What did you do and what was the outcome?
- Think of a person whom you admire on the job. Can you recall an incident that convinced you that the person was an outstanding performer?
- Think of a time when you realized too late that you should have done something differently. What did you do and what was the outcome?
- Think about the last six months. Can you recall a day when you were particularly effective? What did you do that made you effective?
- Think of a time when you saw someone do something in a situation and you thought to yourself, "If I were in that same situation, I would handle it differently." What was the scenario you saw?
- Think about mistakes you have seen workers make when they are new at the job.
- Think about actions taken by more experienced workers that help them to avoid making mistakes.

Figure 5: Prompts to Encourage the Writing of Critical Incidents (adapted from Anderson & Wilson, 1997)

Step 3. Sort critical incidents. After incidents are collected and edited, they need to be reviewed and categorized into groups by SJT developers who are knowledgeable about the job(s) for which the SJT is being developed. It is useful to have multiple people sort the critical incidents into piles. Some sorts are more appealing than others. The content of the critical incidents dictates the piles. Typical content piles are shown in Figure 6.

The goal of the sorting is to two-fold:

- Identify duplicate or near duplicate critical incidents
- Identify areas in which item stems will be written.

When sorting incidents, several of them will be near duplicates. For example, there will be many incidents in which SMEs have too much work to do with inadequate resources or the boss is a hindrance, rather than helpful, or co-workers are difficult. There is a tradeoff because duplicate incidents do not add new information to a test, but

stems with similar content may allow the developers to get a better understanding of the content area and increase the reliability of the SJT.

When reviewing the sorted piles of critical incidents, one may identify content areas that would be inappropriate to share with job applicants. For example, most employers would not want SJT items on a test that would cover topics such as employment discrimination, workplace violence, or topics that are a source of conflict within the organization (e.g., lack of promotional opportunities, unpopular new policies). The sorted critical incidents should be reviewed by decision makers. Some content areas that seem acceptable to the SJT developer may not be acceptable for test content by higher level decision makers within the organization.

However, if a job analysis has been conducted in which work behavior or duties are identified, it is possible to have SMEs create item stems (Step 4) bypassing Steps 2 and 3. This shortcut assumes that the job analysis is well documented and that SMEs are adequately trained to develop situations likely to occur on the job. If incidents can be created directly from the job analysis during the first half of a one-day workshop, and responses to the situations can be developed during the second half of the workshop, a first draft of a SJT can be developed using only a single day of SME time (Brull, personal communication, June 20, 2005). For the remainder of this chapter, we assume that critical incidents have been collected and sorted.

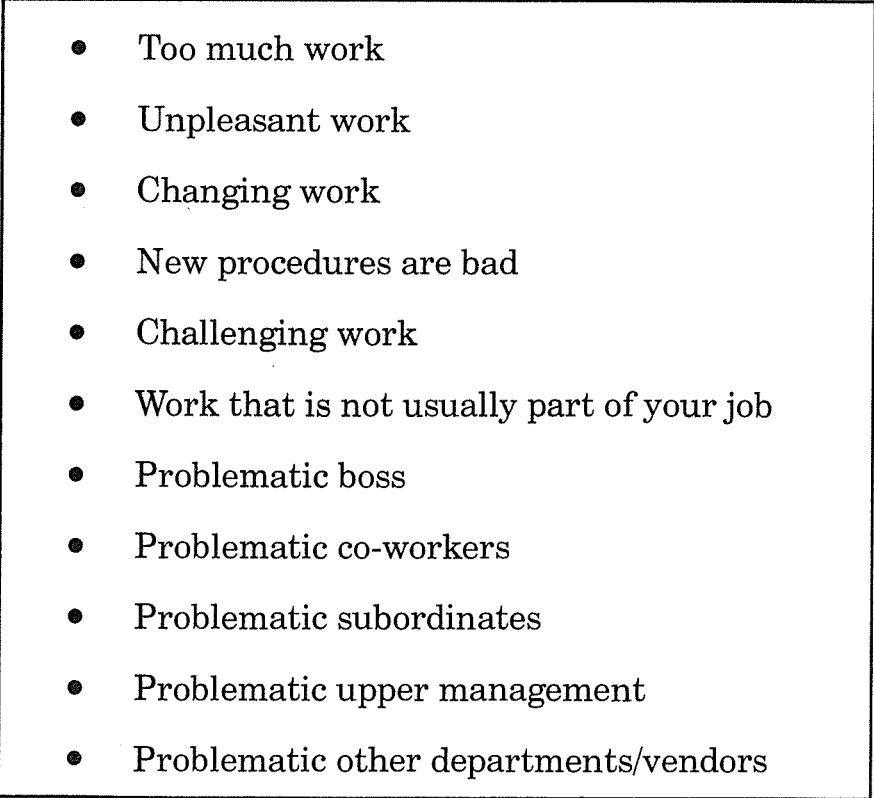
- 
- Too much work
 - Unpleasant work
 - Changing work
 - New procedures are bad
 - Challenging work
 - Work that is not usually part of your job
 - Problematic boss
 - Problematic co-workers
 - Problematic subordinates
 - Problematic upper management
 - Problematic other departments/vendors

Figure 6: Descriptions of Content Areas Typically Found in Critical Incidents

Step 4. Create item stems from critical incidents. The critical incident statements need to be rewritten to produce SJT item stems. When writing stems, one needs to consider potential redundancy of item stems. The same item does not need to be written twice but one needs to decide how redundant the items are permitted to be. For example, how many problematic co-workers items are needed? Consider the following possibilities for topics in item stems:

- Good co-worker gone bad
- Co-worker breaks rules
- Co-worker is rude
- Co-worker is lazy
- Co-worker needs training
- Co-worker has poor personal hygiene.

One needs to consider the number of items covering a topic that seem reasonable, given the content areas to be covered and the length of the test.

Stems need to be written at an appropriate level of specificity. The critical incident is probably job relevant for the SME who wrote the item but it may not be job related for all jobs to be covered by the SJT. Consider a critical incident concerning the difficulty in learning a new software package for inventory control. If all jobs do not require the use of this software, the stem could be written to refer to difficulty in learning new software in general. If all the jobs do not require any kind of software, the stem could be written as difficulty in learning a new work procedure.

The stems need to be edited for clarity and brevity. Stems with ambiguous meanings will result in disagreement concerning the effectiveness of the responses. One should also standardize the use of common vocabulary (e.g., boss vs. supervisor, co-worker vs. team member). Making these stylistic decisions on vocabulary early in the process will reduce editing time.

Step 5. Develop item responses. Response alternatives should represent different strategies for handling each situation. The alternatives should all seem reasonable but some should be more “correct” than others for the situation. The more correct alternatives should be more attractive to applicants with the best potential for success on the job.

To collect item responses, a survey of item stems should be assembled with space available for respondents to write potential responses to the stem. The critical incident from which the stem was developed probably should contain one response to the situation. If there are more stems than a SME can respond to in the given amount of time, the

survey can be split into several sections and only one section is administered to any given SME. Multiple SMEs should write additional responses for each stem. Prompts that can be used to coach SMEs on writing responses are shown in Figure 7.

- What would you do?
- What is the best thing to do?
- What is an ineffective response that you think many people would do?
- What would an ineffective employee do?
- Think of a really good employee that you know well. What would that employee do in this situation?
- Think of a poor employee that you know well. What would that employee do in this situation?

Figure 7: Prompts for Writing Item Responses

A given SME expert will often only be able to generate two or three non-redundant responses. To maximize the number of non-redundant responses, multiple SMEs should work independently. There will be variability in the number of responses written per stem. A pool of SMEs working independently can usually generate between five and twelve non-redundant responses.

After the critical incident workshops, the employer may realize the labor demands of this project and it may be difficult to obtain as many SMEs as needed to generate item responses. To be responsive to the labor pressure, the test developer might generate some item responses to reduce the number of additional SMEs needed. This is possible for item stems that lack specific technical content such as generating responses to a stem concerning a situation in which it is difficult to work with a co-worker. However, item stems that are highly technical (e.g., a complex cost accounting situation) will likely require SMEs to generate the item responses.

The item responses will require editing for redundancy, acceptability, clarity, and brevity. With multiple SMEs working independently, it is likely that one will have identical or nearly identical item responses. Although it is not recommended to include the same item response twice, some redundancy might be allowed in responses to convey a nuance. For a situation in Figure 1 concerning not receiving a new computer, consider the following responses:

- Confront your boss about
- Assume it was a mistake and speak with your boss...

Both responses address speaking with your boss, but the tone of the responses differ. Some responses will be unacceptable because they describe behavior that is so low in effectiveness that no applicant would judge the response to be effective. For a scenario concerning a conflict with your supervisor, no applicant would find the following response effective: *Punch the boss in the face*. Whereas all or almost all respondents will indicate that this is an effective response or a response they would be very unlikely to perform, there will be no variance in the evaluation of that response. Most responses will need edited for clarity and brevity.

Step 6. Select item response instructions. Earlier in the chapter, we noted distinctions between knowledge instructions and behavioral tendency instructions. We generally recommend using knowledge instructions for two reasons:

- SJTs with knowledge instructions show somewhat higher criterion-related validity than tests with behavioral tendency instructions after controlling for item content (McDaniel et al., 2007)
- SJTs with knowledge instructions are likely more resistant to faking than SJTs with behavioral tendency instructions (McDaniel & Nguyen, 2001; Nguyen et al., 2005)

However, one can expect greater mean racial differences with a knowledge instruction than with a behavioral tendency instruction (Whetzel et al., 2008). Also, if the SJT is to be used as part of a selection battery with a cognitive ability test, one might get more incremental prediction from an SJT with behavioral tendency instructions than one with knowledge instructions. This is due to the finding that SJTs with behavioral tendency instructions have relatively lower correlations with cognitive ability (McDaniel et al., 2007) and thus may have better incremental validity over and above a cognitive ability test.

Step 7. Develop scoring key. Typically, a scoring key is developed by collecting judgments from SMEs about the effectiveness of the alternative response options for handling each work-related situation. The SJT developer prepares a questionnaire in which SMEs are asked to evaluate the response alternatives for each job situation. Depending on the kind of scoring key developed, there are two different kinds of judgment. For one judgment, SMEs evaluate the effectiveness of each alternative response by rating each response on a scale ranging from very ineffective to very effective. The other kind of judgment involves having SMEs identify the best alternative and the worst alternative for each work situation.

When effectiveness ratings are collected, the SJT developer computes the mean rating of effectiveness given to each item and the standard deviation around that mean rating. If the standard deviation is high, that response alternative should not be used. When best/worst ratings are collected, the SJT developer should compute the proportion

of experts who endorsed each alternative as the most effective, the proportion who endorsed each alternative as least effective. If there is substantial disagreement about the best/worst ratings, the response option should not be used (Motowidlo et al., 1997). When there is disagreement about the best response, it is often due to some ambiguity in the item stem, usually something left unsaid. If the respondent makes one assumption a response is effective. If another assumption is made, the same response may be considered ineffective. Similarly, responses that do not have any variance (i.e., all SMEs agreed that a behavior is the best (or worst) or scored it high (or low) in effectiveness) should be dropped. If the instructions involve having examinees choose the best or worst response, the following simple scoring pattern is recommended:

- 1 Indicating that the keyed best response is the worst response
 Indicating that the keyed worst response is the best response
- +1 Indicating that the keyed best response is the best response
 Indicating that the keyed worst response is the worst response
- 0 Any other response

If a Likert scale is used to rate the effectiveness of each response option, we recommend using a similar keying strategy. Consider the following four-point rating scale for judging individual responses:

1	2	3	4
Very Ineffective	Ineffective	Effective	Very Ineffective

We recommend the following keying:

-1	Indicates that an effective behavior is ineffective or very ineffective Indicates that an ineffective behavior is effective or very effective
+ 1	Indicates that an effective behavior is effective or very effective Indicates that an ineffective behavior is ineffective or very ineffective.

This scoring strategy is recommended for several reasons. First, it requires those who are making the key to only agree on whether the response option is an effective behavior or an ineffective behavior. One can more easily get agreement on this dichotomous decision than if one requires the keying decision makers to distinguish between effective and very effective or between ineffective and very ineffective. Second,

there are individual differences in how respondents interpret relative statements (e.g., effective vs. very effective). Two respondents might consider a given response option to be at the same level of effectiveness but one respondent may describe it as “effective” while another may see it as “very effective.” This happens due to the respondents’ varying interpretations of the word “very.” For scoring key development, we recommend the dichotomous scoring strategy above. However, we also recommend that an even number Likert rating scale be used to collect the respondents’ ratings because respondents may feel constricted by two-point, dichotomous ratings scales.

A second scoring option is data-assisted rational keying that involves collecting effectiveness ratings and using the mean to determine the effectiveness of the responses. Response options that are clearly effective or ineffective based on the means can be scored using the dichotomous scoring procedure described above. Response options for which the mean ratings are near the middle of the scale (i.e., neither effective nor ineffective) should not be scored.

Another data-assisted rational keying approach involved deviation scoring from the mean (Legree, Psotka, Tremble, & Bourne, 2005). Here, the mean rating is determined to be the correct answer and ratings diverging from the mean receive lower scores. For example, if the mean SME effectiveness of a particular response is 2.5 and an examinee rates the response at 3, s/he loses $\frac{1}{2}$ point. Conversely, if the examinee scores the response at 2.0, s/he also loses $\frac{1}{2}$ point. If one does not transform the score (e.g., add 100 to all scores), the highest possible score is a zero and the lowest possible is some negative number. If one uses this scoring strategy we recommend adding some constant to the score to move all scores into a positive range. It is difficult to explain negative score to respondents.

A final scoring option is to use empirical scoring approaches similar to those used in developing scoring keys for biodata (e.g., Hogan, 1994). Krokos, Meade, Cantwell, Pond, and Wilson (2004) studied the use of empirical scoring. Their reasoning was that SJTs are often developed to measure complex, social or practical aspects of performance in employment situations, which are nearly the same as tests of tacit knowledge (McDaniel & Whetzel, 2005). When SJTs are based on a scoring algorithm developed based on the consensus judgment of SMEs, the chance that the correct answers will be the most obvious is increased. Thus, more transparent items are more likely to be kept when SMEs are asked to determine the correct answer. Contrary to Legree et al (2005), they state that if respondents’ answers are used to develop the scoring key, the “SMEs” consist of both high and low applicants/performers. Thus, low validity coefficients for SME-scored SJTs could be a result of differing perceptions of the construct among respondents.

In contrast, when empirical keying is used, the “SMEs” are the high performing respondents as measured by the criterion. Using empirical keying, the most transparent

option (the one that seems the best) may be endorsed by both high and low performing respondents. Thus, it will not differentiate between criterion groups and therefore, will not be weighted. A response option that is endorsed less frequently, but only by the high performing respondents, will be weighted more heavily with empirical scoring.

In our experience, when items are empirically keyed, item responses that describe negative behavior tend to have higher criterion-related validity than item responses that describe effective behavior. By selecting only the items with the highest criterion-related validity, one is likely to have an unbalanced key (i.e., one can obtain a passing score by saying that the majority of behaviors are ineffective). To the extent that the negatively biased key becomes public knowledge among applicants (e.g., in promotional exams where applicants may know each other), the test may become compromised.

Another issue to be considered when keying a SJT is developing a key using incumbent judgments and then using the SJT to test applicants. Using a call center example, based on company norms, incumbents may respond that getting customers off the phone quickly is an effective response. However, an applicant who has excellent customer service skills may answer that they should keep the customer on the phone until the customer is completely satisfied. Thus, incumbents may respond based on company norms or job knowledge that may be contrary to what applicants with good qualities would normally do. If it appears that an item is answered one way by most applicants and another way by most incumbents, one may wish to drop the item from the test.

A similar concern involves the use of test keys across organizations. While many keys are probably generalizable across organizations, there may be some exceptions. For example, some SJTs include content related to confronting one's boss about an issue with the boss's behavior or decisions. Though this may be acceptable in some organizations, it may be unacceptable in other organizations, especially those that are hierarchical in nature.

CONCLUSION

In this chapter, we have addressed SJTs from the perspective of a practitioner. We have described eight characteristics of SJTs that guide practitioners in making decisions about the format of their SJT. We have also provided a seven step procedure for developing a SJT. We recognize that there are other approaches to developing SJTs and encourage those using such procedures to document and share those procedures to encourage the development of high quality SJTs.

REFERENCES

- Anderson, L. & Wilson, S. (1997). Critical incident technique. In D.L. Whetzel, & G.R. Wheaton (Eds.) *Applied Measurement Methods in Industrial Psychology* (pp.89-114). Palo Alto: Davis Black.
- Brull, H. (June 20, 2005). Personal communication.
- Callinan, M., & Robertson, I.T. (2000). Work sample testing. *International Journal of Selection and Assessment*, 8, 248-260.
- Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in SJTs: 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). *Examining the relationship between job knowledge and situational judgment performance*. Paper presented at the 15th Annual Conference of the Society of Industrial and Organizational Psychology. New Orleans. April.
- Clevenger, J., Pereira G.M., & Weichmann, D. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, 86, 410-417.
- Dye, D.A., Reck, M., & McDaniel, M.A. (1993). Moderators of the validity of written job knowledge measures. *International Journal of Selection and Assessment*, 1, 153-157.
- Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin*, 51, 327-358.
- Added
- Friede, A., Imus, A., & Oswald, F.L. (2005). *Using shorter items in a situational judgment inventory*. Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology: Los Angeles, CA.
- Gottfredson, L.S. (2003). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence*, 31, 343-397.
- Hogan, J.B. (1994). Empirical keying of background data measures. In G.S. Stokes & M.D. Mumford (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 69-107). Palo Alto, CA: CPP Books.
- Irvine, S.H., & Kyllonen, P.C. (Eds.). (2002). *Item generation and test development*. Mahwah, NJ: Erlbaum.
- Krokos, K., Meade, A.W., Cantwell, A.R., Pond, S.B., Wilson, M.A. (2004). *Empirical keying of situational judgment tests*. Paper presented at the 16th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.
- Kulkarni, S.S., & Puhan, B.N. (1988). Psychological assessment: Its present and future trends. In J. Pandey (Ed.), *Psychology in India: The state of the art* (pp. 19-91). New Delhi, India: Sage.

- Legree, P.J., Psotka, J., Tremble, T., & Bourne, D. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R.D. Roberts (eds), *Emotional Intelligence: An International Handbook* (pp. 155-179). Berlin, Germany: Hogrefe & Huber.
- 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. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, 91, 1181-1188.
- 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, 1043-1055
- McDaniel, M.A., Hartman, N.S., Whetzel, D.L., & Grubb III, W.L. (2007). Situational judgment tests, knowledge, behavioral tendency, 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, 730-740.
- 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. & Whetzel, D.L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence*, 33, 515-525
- Motowidlo, S.J., & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology*, 66, 337-344.
- 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., 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: Davis 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.
- Nguyen, N.T., McDaniel, M.A., & Whetzel, D.L. (2005). *Subgroup differences in situational judgment test performance: A meta-analysis*. Paper presented at the 20th annual conference of the society for Industrial & Organizational Psychology, Los Angeles.

- O'Connell, M.S.; Hartman, N.S.; McDaniel, Michael A.; Grubb, W.L., & Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual performance. *International Journal of Selection and Assessment*, 15, 19-29.
- Olson-Buchanan, J.B., Drasgow, F.; Moberg, P.J., Mead, A.D., Keenan, P.A., & Donovan, M.A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology*, 51, 1-24.
- 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.
- Parker, C.W., Golden III, J.H., Russell, D.P., Redmond, M.R. (2000). *The development of a construct-related scoring key of a situational judgment inventory for enhancing criterion-related validity*. Paper presented at the 15th Annual Conference of the Society of Industrial and Organizational Psychology: New Orleans. April.
- Ployhart, R.E., & Holtz, B.C. (2008). The diversity-validity dilemma: Strategies for reducing race/ethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, 61, 153-172.
- 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.
- Pyburn, K.M., Ployhart, R.E., & Kravitz, D.A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology*, 61, 143-151.
- Premarajan, R.K., Thornton, G.C. III, & Padhi, P.K. (2008). Legal environment for selection in India. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 258-259
- Sacco, J.M., Scheu, C.R. Ryan, A.M., Schmitt, N., Schmidt, D.B. & Rogg, K.R. (2000). *Reading level and verbal test scores as predictors of subgroup differences and validities of situational judgment tests*. Unpublished manuscript.
- 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, & C. Viswesvaran (Eds.), *Handbook of Industrial, Work, & Organizational Psychology: Vol. 1* (pp. 165-199). London and New York: Sage.
- Sternberg, R.J., Forsythe, G.B., Hedlund, J., Horvath, J.A., Wagner, R.K., Williams, W.M., Snook, S.A., Grigorenko, E.L. (2000). *Practical intelligence in everyday life*. New York: Cambridge University Press.
- 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. (1997). Video-based situational testing. *Personnel Psychology*, 50, 25-49.
- Weekley, J.A. & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology*, 52, 679-700.

- Weekley, J.A., & Ployhart, R.E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance*, 18, 81-104.
- 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.