McDaniel, M.A. & Whetzel, D.L. (2007). Situational judgment tests. In D.L. Whetzel & G. R. Wheaton (Eds.). *Applied measurement: Industrial psychology in human resources management*. Erlbaum. 235-258.

CHAPTER NINE

Situational Judgment Tests

Michael A. McDaniel
Virginia Commonwealth University and Work Skills First, Inc

Deborah L. Whetzel
United States Postal Service

OVERVIEW

Situational judgment tests present applicants with scenarios that they might encounter on the job and ask applicants to evaluate various actions that might be taken in response to the situations. Thus, situational judgment tests are one type of job simulation that measures job knowledge. They rely on the principle that a person's behavior in past situations can predict how that individual is likely to behave in similar situations in the future. This means that one way to predict how effectively job applicants will perform on the job when they become employees is to measure how effectively they perform in a simulation of the job when they are applicants. People vary in the extent to which they acquire job knowledge from a series of events, and situational judgment tests are a method of assessing such differences.

Some job simulations are more realistic than others. For example, applicants for airplane pilot positions might be placed in a flight simulator that provides a very realistic reproduction of flying an airplane. Although such realistic simulations might be expected to have an advantage in their predictive potential, they can be so expensive to develop and administer that less realistic simulations often are an attractive alternative. Jobs for which high-fidelity simulations (e.g., flight simulators) are used to predict performance typically require some level of job experience. One advantage of low-fidelity simulations, such as situational judgment tests, is that they can be developed to predict performance in entry-level jobs in which no experience is required as well as in higher level jobs.

This chapter describes situational judgment tests in some detail and provides a brief history of their use. It then describes the characteristics along which situational judgment tests can vary, including test fidelity, cognitive complexity, and response instructions. The chapter then summarizes research evidence about their reliability, validity, and subgroup differences. Finally, the chapter describes procedures for developing situational judgment tests based on research findings.

WHAT IS A SITUATIONAL JUDGMENT TEST?

In a situational judgment item, a scenario is described and the applicant is required to evaluate several possible responses to the scenario. As mentioned earlier, situational judgment tests vary in their level of fidelity to the job. Typically, for entry-level selection (e.g., to an electrician apprenticeship training program), a situational judgment test might refer to general situations that might be encountered by employees. Such items might describe a problem with a boss or coworker. For higher level positions (e.g., promotion to journeyman), a situational judgment test might refer to circumstances that closely reflect issues or problems encountered in a specific job. An example of a situational judgment test item that might be used for promotion is shown here:

> You and another journeyman electrician from another crew are jointly responsible for coordinating a project involving both crews. This other person is not carrying out his share of the responsibilities. You would ...
>
1. Discuss the situation with your foreman and ask him to take it up with the other person's foreman.
2. Remind him that you need his help and that the project will not be completed effectively without a full team effort from both of you.
3. Tell him that he is not doing his share of the work, that you will not do it all yourself, and that if he does not start doing more, you will be forced to take the matter to his foreman.
4. Try to find out why he is not doing his share and explain to him that this creates more work for you and makes it harder to finish the project.
5. Get someone else from his crew to help with the project.

Motowidlo, Dunnette, and Carter (1990) used the term *low-fidelity simulation* to describe such an item. Sternberg et al. (2000) developed similar items and called them assessments of "practical intelligence." However, within the areas of human resource management and industrial psychology, these items typically are now referred to as "situational judgment."

Several situational judgment tests have been developed to predict job performance in managerial and supervisory positions (e.g., Bruce & Learner, 1958; Campbell, Dunnette, Lawler, & Weick, 1970; Corts, 1980; Mandell, 1950; Motowidlo et al., 1990; Wagner & Sternberg, 1991). They have also been developed to predict insurance agent turnover (Dalessio, 1992), success in engineer positions (Clevenger, Jockin, Morris, & Anselmi, 1999), performance in teams (Stevens &

Campion, 1999), success in telephone sales and collection positions (Phillips, 1992, 1993), and skill in managing conflict (Olson-Buchanan, Drasgow, Moberg, Mead, & Keenan, 1994). Thus, situational judgment tests are quite flexible and can be developed to predict performance in a variety of jobs.

A BRIEF HISTORY OF SITUATIONAL JUDGMENT TESTS

As noted by Weekley and Ployhart (2006), the earliest example of situational judgment tests depends on how they are defined. For example, a U.S. civil service exam used in 1873 for the Examiner of Trade-Marks, Patent Office contained the following: "A banking company asks protection for a certain device, as a trademark, which they propose to put upon their notes. What action would you take on the application?" (DuBois, 1970, p. 148). The 1905 Binet scale used to measure intelligence in children included questions such as, "When a person has offended you, and comes to offer his apologies, what should you do?" Although situations were presented, these early efforts did not include possible ways of handling the situation that were presented to the applicant.

As noted by McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001), the first widely used situational judgment test that contained response options was likely the George Washington Social Intelligence Test. One of the subtests, called Judgment in Social Situations, required "keen judgment, and a deep appreciation of human motives, to answer correctly" (Moss, 1926, p. 26). Several solutions to each situation were offered in a multiple-choice format, only one of which was correct. In an early review of empirical studies, Thorndike and Stein (1937) criticized the test, claiming that correlations between the test and other tests of presumed social attributes were very low.

During World War II, Army psychologists attempted to assess the judgment of soldiers (Northrop, 1989). These judgment tests were comprised of scenarios and a number of alternative responses to each scenario. Solutions were based on the person's ability to use common sense, experience, and general knowledge, rather than logical reasoning. Starting in the 1940s, a number of situational judgment tests were developed to measure supervisory potential. These included the Practical Judgment Test (Cardall, 1942), How Supervise? (File, 1945; File & Remmers, 1948), Supervisory Practices Test (Bruce & Learner, 1958), Business Judgment Test (Bruce, 1965), Supervisory Judgment Test (Greenberg, 1963), and the Supervisory Inventory on Human Relations (Kirkpatrick & Planty, 1960). In the late 1950's and early 1960's situational judgment tests were also used by large organizations as part of selection test batteries to predict managerial success. For example, the Standard Oil Company of New Jersey designed a program of research called the Early Identification of Management Potential to identify employees who have potential to be successful in management (Campbell, Dunnette, Lawler, & Weick, 1970).

Recently, there has been renewed interest in the use of situational judgment measures for predicting job performance. For example, the United States Office of Personnel Management designed Test 905 to assess the human relations capacity

and potential of applicants for promotion to first-line federal trades and labor supervisory positions (Corts, 1980). Motowidlo et al. (1990) examined the use of a situational judgment test, referred to as a low-fidelity simulation, for selecting entry-level managers. In validation studies with samples of managers from seven different companies, correlations between the test and various job performance criteria ranged from the .20s to the .40s. Wagner and Sternberg (1991) published a test called the Tacit Knowledge Inventory for Managers (TKIM). This measure is based on their theory of tacit knowledge, or "... practical know-how that usually is not openly expressed or stated and which must be acquired in the absence of direct instruction" (Wagner, 1987, p. 1236). The TKIM presents scenarios that require respondents to choose a course of action from a list of alternatives. These scenarios differ from those of previously mentioned tests in that the TKIM scenarios are considerably longer and more detailed. Wagner and Sternberg (1991) reported five studies examining the criterion-related validity of tacit knowledge measures in academic and business settings, although no validity was presented for the TKIM itself. They found moderate correlations between their measure and a variety of criteria, some of which would be considered job performance measures. Sternberg et al. (1995) also reported that these measures were unrelated to measures of general cognitive ability. This conclusion should be tempered by the fact that their samples (e.g., Yale undergraduate students) are likely to have substantial range restriction on measures of general cognitive ability, thus reducing observed relationships on the restricted predictor.

Finally, in investigating a situational judgment test, Smith and McDaniel (1998) found the largest correlates were with age and length of job experience. From this, they inferred that the test measured job-related knowledge and skills gained through life and work experiences. The test also correlated with the personality dimensions of conscientiousness (r = .32) and emotional stability (r = .22) as well as with measures of general cognitive ability (mean r = .22). Smith and McDaniel concluded that the situational judgment test assessed multiple job-related constructs.

THE STRUCTURE AND FORMAT OF SITUATIONAL JUDGMENT TESTS

Situational judgment items vary with respect to several characteristics (McDaniel & Nguyen, 2001; McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006). Knowledge of the various situational judgment test formats can assist situational judgment test developers in making informed decisions. Also, some of the characteristics of situational judgment tests have implications for test validity and the degree of mean subgroup differences in test scores. Here, some of the major characteristics along which situational judgment tests vary are reviewed. These include test fidelity, cognitive complexity, and response instructions.

Test Fidelity

Test fidelity refers to the extent to which the test format mirrors how a situation would be encountered in a work setting. A higher fidelity situational judgment

test may involve presenting situations using a short video, whereas a lower fidelity situational judgment test involves presenting situations in a written format (paper-and-pencil or computer presentation of text). Also, there are levels of fidelity within types of presentation. A situational judgment test with a written format might have more fidelity if the situation were described using technical terms common to the job. A situational judgment test with a video format might have less fidelity to the extent the item situation differs from aspects of the actual work situation. For example, if the work procedures shown in the video are not the most current procedures, then the situational judgment test may have less fidelity. This can happen when work procedures change over time but the video-based situational judgment test is not updated.

Video-based situational judgment tests are likely to reduce the reading and other cognitive demands relative to paper-and-pencil situational judgment tests. Consequently, video-based situational judgment tests typically produce lower correlations with cognitive ability, as well as smaller mean ethnic differences, when compared to paper-and-pencil situational judgment tests (Chan & Schmitt, 2002; Whetzel, McDaniel, & Nguyen, 2005). Although the smaller mean ethnic differences in video situational judgment tests are an advantage, the reduced correlations of such tests with cognitive ability may result in lower validities. Additional research is needed to evaluate this possibility.

Cognitive Complexity

Situational judgment items vary in their ***cognitive complexity***. Cognitively complex items require more cognitive resources to understand than less cognitively complex items. One factor likely to influence the cognitive complexity of the items is the length of the stem. Some situational judgment test items have very short stems (e.g., "You have encountered a problem that you cannot solve and you cannot locate your supervisor to help you with the problem"). Other situational judgment test stems are much longer, such as those in the TKIM (Wagner & Sternberg, 1991) mentioned earlier. Longer item stems likely increase cognitive complexity in part through increased demands on reading comprehension. Thus, items with longer item stems tend to be more difficult to comprehend than items with shorter stems.

A second factor that is likely to influence the cognitive complexity of the items is the complexity of the situation presented. Consider this sample stem again: "You have encountered a problem that you cannot solve and you cannot locate your supervisor to help you with the problem." This stem describes a relatively low-complexity situation with obvious potential responses. For example, the employee could seek assistance from a different supervisor or a knowledgeable coworker, or the employee could gain knowledge of the assignment from reading. In contrast, an example of a high-complexity stem would be: "You are supervised by two electricians who are not getting along. The electricians give you conflicting instructions and each demands that the work each assigns be given the highest priority." This stem describes a more complex situation in which the potential responses also may be complex.

	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 you would perform the response. Rank the responses from the most likely to the least likely.
Knowledge	Pick the best answer.What should you do?	Pick the best answer and pick the worst answer. Pick the best and second best.	Rate each response for effectiveness. Rank the responses from the best to the worst.

FIGURE 9.1 Taxamony of response instructions in situational judgement tests.

One can assess the cognitive complexity of a situational judgment test item by correlating that item with a cognitive ability test (McDaniel & Nguyen, 2001). This correlation might be based on data obtained from a pilot test or from operational use of the test. The cognitive complexity of a test has implications for mean racial differences and test validity. Given the large mean ethnic differences in cognitive ability (Roth, BeVier, Bobko, Switzer, & Tyler, 2001), a more cognitively complex test item is likely to increase the item's mean ethnic group differences. Because cognitive ability is one of the best predictors of job performance (Schmidt & Hunter, 1998), one might expect cognitively complex situational judgment tests to yield higher validities than less cognitively complex situational judgment tests. The implications of cognitive complexity for validity and subgroup differences as well as test development are discussed later in this chapter.

Response Instructions

There are many types of response instructions that can be used with a situational judgment test. McDaniel, Whetzel, and Nguyen (2006) offered a two-dimensional table of response instructions (see Fig. 9.1). The rows of the table are labeled "Behavioral Tendency" and "Knowledge." In a situational judgment test with behavioral tendency instructions, applicants are asked to report how they would typically behave in response to the situation. A situational judgment test with knowledge instructions asks applicants to evaluate the effectiveness of responses. The second dimension, defining the three columns of the table, lists the number of scoreable responses that can be obtained from the item. Some response instructions (e.g., "Pick the best answer") generate one scoreable response per item. Other response

instructions yield two dichotomous responses per item (e.g., "Pick the best answer and pick the worst answer"). Still other response instructions yield as many scoreable responses as there are response options (e.g., "Rate each response for effectiveness").

Response instructions have been shown to influence the construct as well as the criterion-related validity of situational judgment tests. Due to their influence on construct validity (i.e., the extent to which the tests are correlated with *g*), response instructions also affect subgroup differences. The implications of response instructions for validity and subgroup differences as well as test development are discussed later in this chapter.

WHAT DO SITUATIONAL JUDGMENT TESTS MEASURE?

Situational judgment tests are best viewed as measurement methods. Some situational judgment tests might emphasize technical knowledge whereas others might emphasize knowledge of how to work as part of a team. Although situational judgment tests might be developed to measure specific personality or ability variables, their unique format (presenting a hypothetical work situation and eliciting a hypothetical response to that situation) lends itself especially well to measuring various forms of job knowledge. As mentioned, the predictive principle behind all simulations, including situational judgment tests, is that people's behavior in the past can predict how they will behave in similar situations in the future.

One reason that people behave somewhat consistently in similar situations is that through experience they develop beliefs or knowledge about the best thing to do in certain situations in order to achieve desired results. Some people have better opportunities to have these experiences and some people are better able to take advantage of their experiences and learn from them. As a result, people who have acquired this situational knowledge over time should know better how to deal with certain situations and should be consistently more effective in those situations than people who, for whatever reason, have not acquired that knowledge.

In their book, *Practical Intelligence in Everyday Life*, Sternberg et al. (2000) asserted that there is a general factor of practical or tacit intelligence that is substantively distinct from general cognitive ability. The items used in Sternberg et al.'s "practical intelligence" tests are situational judgment items. Thus, rather than measuring some unique and previously unknown construct using a relatively novel measurement tool, Sternberg et al. were actually using situational judgment tests. Reviews by Gottfredson (2003), McDaniel and Whetzel (2005), and Ree and Earles (1993) noted that there is no support for a construct of practical intelligence. The practical intelligence construct also is critiqued in chapter 5.

There are several methods for empirically identifying constructs measured by a selection instrument. One involves correlating scores on one instrument with scores on another. McDaniel, Hartman, and Grubb (2003) correlated situational judgment test scores with scores on cognitive ability and scores on the Big 5 personality dimensions (Digman, 1990; Goldberg, 1993; John, 1990). In general, the results showed that situational judgment tests are correlated with measures of cognitive ability and personality.

TABLE 9.1
An Example of the Multidimensionality of Situational Judgment Test Items (Scenario A)

Scenario A

You assigned a very high profile project to one of your project managers. During each of the project update meetings, your project manager indicates that everything is going as scheduled. Now, one week before the project is due, your project manager informs you that the project is less than 50% complete.

	Correlation with:		
Responses:	*g* *(n = 448–450)*	*Conscientiousness* *(n = 1196–1222)*	*Agreeableness* *(n = 1196–1222)*
Personally take over the project and meet with the customer to determine critical requirements.	.10*	.01	–.13*
Meet with the customer to extend the deadline. Talk with the project manager about how the lack of communication has jeopardized the company's relationship with the customer.	.11*	–.03	–.05
Fire the project manager and take over the project yourself.	.08	.00	–.16*
Coach the project manager on how to handle the project more efficiently.	–.17*	.01	.09
Do not assign any high profile jobs to this project manager in the future.	.13*	.07	–.08

* indicates statistically significant correlations.

TABLE 9.2
An Example of the Multidimensionality of Situational Judgment Test Items (Scenario B)

Scenario B			
You lead a project that requires specific, accurate data to make decisions. The data-capturing method currently being used does not provide you with the information you need. Another department promised to provide you with the information, but failed to do so at the last minute. This setback delayed your project and you are certain that you will require the information to complete your project accurately.			
	Correlation with:		
Responses:	*g* *(n = 448–450)*	*Conscientiousness* *(n = 1196–1222)*	*Agreeableness* *(n = 1196–1222)*
Do the time-consuming work yourself even though it is not tecnically your responsibility.	.07	.11*	−.08*
Temporarily allocate some member of your team to capture the data.	−.01	.11*	.00
Ask the customer for a deadline extension and explain that the other department failed to provide the necessary information.	.12*	.06	−.02
Ask your manager to pressure the other department to deliver the information.	.17*	.02	−.10*

* indicates statistically significant correlations.

McDaniel and Whetzel (2005) reported correlations between measures of *g* and personality and response options in a situational judgment test. The correlations are shown in Tables 9.1 and 9.2. These items were developed for professional positions in a Fortune 500 corporation and are presented here with permission. Each item presents a scenario and several response options. The respondents were asked to rate the effectiveness of each response option for resolving the problem depicted in the scenario. Each response option was individually correlated with other variables (e.g., cognitive ability and personality) collected on each respondent. In scenario A (Table 9.1), the first response option was judged effective by those higher in *g* ($r = .10$) and lower in agreeableness ($r = -.13$). The second and fifth options were judged effective by those high in *g* ($r = .11$ and .13). The third option was judged effective by those low in agreeableness ($r = -.16$). The fourth option was judged effective by those low in *g* ($r = -.17$). Other correlations were suggestive of relationships with the effectiveness ratings but were not statistically significant. The correlations are all relatively low because they represent correlations with a single item with limited reliability. In scenario B (Table 9.2), the first two options were found effective by those higher in conscientiousness (both $r = .11$), the third and fourth options found effective by those higher in *g* ($r = .12$ and .17), and the fourth option was found effective by those low in agreeableness ($r = -.10$). In summary, the response options for these scenarios, like most SJT scenarios and response options, are often construct heterogeneous. Tests made up of such items measure multiple constructs and have loadings on multiple factors.

A second method for empirically identifying constructs measured by situational judgments tests involves conducting factor analysis. However, factor analysis has seldom proved useful in specifying the content of situational judgment tests. Clause, Mullins, Nee, Pulakos, and Schmitt (1998) found that in situational judgment tests, multidimensionality often occurs within individual items. When items are multidimensional, it is very difficult to specify their content through traditional, empirical means such as factor analysis. McDaniel and Whetzel (2005) reviewed the sparse literature on factor analyses of situational judgment test items and concluded that instances of interpretable factors are rare.

In summary, both the correlational and factor analysis methods have shown situational judgment tests to be multidimensional, even at the item level. They appear to measure a variety of constructs including cognitive ability and the Big 5 (i.e., conscientiousness, agreeableness, extroversion, emotional stability, and openness to experience).

PSYCHOMETRIC CHARACTERISTICS OF SITUATIONAL JUDGMENT TESTS

In this section psychometric properties of situational judgment tests, including reliability, construct validity, criterion-related validity, incremental validity beyond general cognitive ability, and subgroup differences, are discussed

Reliability

Computing the reliability of situational judgment tests is problematic for several reasons. First, the most readily available reliability estimate, Cronbach's alpha, may not be an appropriate reliability index because of the multidimensional nature of situational judgment tests (Cronbach, 1949, 1951). Test-retest reliability is rarely found in the literature on situational judgment tests because it requires at least two separate administrations of the same test to the same examinees. Parallel form reliability often is infeasible because it requires the use of different item content to measure the same constructs. Because it is difficult to isolate the particular constructs assessed using a situational judgment test, construct equivalence across forms can be problematic. Due to these test development and data collection problems, many researchers continue to provide internal consistency estimates while acknowledging that they underestimate the reliability (Chan & Schmitt, 1997; Pulakos & Schmitt, 1996; Pulakos, Schmitt, & Chan, 1996) of situational judgment tests. One notable exception is Chan and Schmitt (2002), who estimated parallel form reliability at .76. Clearly, more thought and research are needed on the appropriate methods for assessing reliability so that we have more and better estimates of the reliability of situational judgment tests.

Construct Validity

Three meta-analyses (McDaniel et al., 2001; McDaniel, Hartman, & Grubb, 2003; McDaniel & Nguyen, 2001) summarized the construct validity of situational judgment tests. McDaniel et al. (2003) provided the most comprehensive of these reviews. They found that situational judgment tests correlate in varying degrees with measures of three of the Big 5 personality traits (Digman, 1990) and with cognitive ability measures. The magnitude of these correlations is moderated by the situational judgment test response instructions, as shown in Table 9.3. Situational judgment tests with behavioral tendency instructions tend to be more correlated with personality than situational judgment tests with knowledge instructions. However, situational judgment tests with knowledge instructions correlate more highly with cognitive ability than do situational judgment tests with behavioral tendency instructions (.43 vs. .23).

As a result of these findings, McDaniel et al. (2003) suggested that it may be possible to change the construct validity of a situational judgment test by altering the response instructions. They could not empirically demonstrate this phenomenon because they had no studies in their sample that held the situational judgment test constant but varied the response instructions. However, since that time several researchers (Mary Doherty, personal communication, July 7, 2005; Hartman & Grubb, 2005; Nguyen, 2004; Nguyen, Biderman, & McDaniel, 2005; Vasilopoulos, Cucina, Hayes, & McElreath, 2005) found that when administering the same situational judgment test with varying response instructions, one can change the magnitude of correlations consistent with the findings of McDaniel et al. (2003).

TABLE 9.3
Meta-Analytic Correlations Between Situational Judgment Tests With Cognitive Ability and Big 5 Measures

	N	*k*	ρ
Cognitive ability	22,553	62	.39
Knowledge instructions	17,290	41	.43
Behavioral tendency instructions	5,263	21	.23
Agreeableness	14,131	16	.33
Knowledge instructions	8,303	5	.20
Behavioral tendency instructions	5,828	11	.53
Conscientiousness	19,656	19	.37
Knowledge instructions	13,754	8	.33
Behavioral tendency instructions	5,902	11	.51
Emotional stability	7,718	14	.41
Knowledge instructions	1,990	4	.11
Behavioral tendency instructions	5,728	10	.51
Extroversion	12,607	10	.20
Knowledge instructions	11,867	5	.21
Behavioral tendency instructions	740	5	.11
Openness to experience	874	5	.12
Knowledge instructions	160	1	.25
Behavioral tendency instructions	714	4	.09

Note. *N* is the number of subjects across all studies in the analysis; *k* is the number of studies; r is the population correlation. The first row in each analysis is the correlation between the situational judgment test and the Big 5 measure for both kinds of instruction.

Thus, test developers who are interested in assessing personality constructs may wish to use behavioral tendency instructions. However, one should note that behavioral tendency instructions are susceptible to faking (Nguyen et al., 2005). On the other hand, if one were interested in assessing cognitive ability, one might use knowledge instructions. The caution here is that a cognitively loaded situational judgment test is likely to result in greater subgroup differences. In summary, there are advantages and disadvantages to both kinds of response instructions and test developers need to carefully consider the consequences of their choices. Recommendations for test development resulting from these findings are provided later in this chapter.

Criterion-Related Validity

The criterion-related validity of situational judgment tests has been evaluated in many primary studies (Chan & Schmitt, 1997; Hanson & Borman, 1989; Motowidlo et al., 1990; Smith & McDaniel, 1998). Two meta-analyses examined the criterion-related validity of situational judgment tests (McDaniel et al., 2001; McDaniel et al., 2003). In the second and more recent meta-analysis, the overall

validity of situational judgment tests across 84 coefficients was .32 (N = 11,809). In addition to overall validity, the study evaluated response instructions as a moderator of validity. As mentioned earlier, knowledge instructions ask examinees to determine the effectiveness of various responses to a situation and behavioral tendency instructions ask examinees what they would do in various situations. Situational judgment tests with knowledge response instructions yielded higher validity (.33) than situational judgment tests with behavioral tendency instructions (.27). Although this is not a large magnitude moderator, it does lead to implications about the design of situational judgment tests. To maximize criterion-related validity, a test developer should consider using knowledge instructions; however, as mentioned before, the use of knowledge instructions is more likely to result in subgroup differences than the use of behavioral tendency instructions. These validity results are almost entirely based on concurrent validity studies (e.g., research typically conducted using job incumbents, rather than applicants, as subjects). Conclusions about the magnitude of the response instruction moderator should be reexamined as estimates of predictive validity (e.g., research typically conducted using applicants as subjects) become available.

Incremental Validity

Two meta-analyses (McDaniel et al., 2001; McDaniel et al., 2003) and several primary studies (Chan & Schmitt, 2002; Clevenger et al., 2000; O'Connell, McDaniel, Grubb, Hartman, & Lawrence, 2002; Weekley & Jones, 1997, 1999) examined the incremental validity of situational judgment tests over measures of cognitive ability. The research is consistent in showing that situational judgment tests provide incremental validity over cognitive ability. As measurement methods, situational judgment tests can assess different constructs to varying degrees, and the degree of incremental validity over cognitive ability will vary depending on the correlation between the situational judgment test and the measure of cognitive ability. Situational judgment tests with high cognitive ability correlations likely will have less incremental validity over cognitive ability than situational judgment tests with low cognitive ability correlations.

Few studies have examined the incremental validity of situational judgment tests over both cognitive ability and personality. One study (O'Connell et al., 2002) reported incremental validity of the situational judgment test over cognitive ability but found little incremental validity over both cognitive ability and personality. However, Weekley and Ployhart (2005) discussed a situational judgment test that provided incremental validity beyond cognitive ability, personality, and experience. More research is needed before one can draw compelling conclusions about the incremental validity of situational judgment tests over both cognitive ability and personality.

Subgroup Differences

Whetzel et al. (2005) examined ethnic and gender subgroup differences in situational judgment test scores. Typically, African Americans scored lower on average

TABLE 9.4
Vector Correlations Between Ethnic and Gender Differences and Constructs Correlated With Situational Judgment Tests

	Cognitive Ability	*Conscientiousness*	*Agreeableness*	*Emotional Stability*
African American/White difference	.88 (18)	.13 (9)	−.38 (9)	−.89 (6)
Male/female difference	−.05(19)	.36 (9)	.38 (10)	.37 (7)

Note. Numbers in parentheses are the numbers of coefficients contributing data to the vector correlations.

than Whites with a mean *d* of .39, where *d* is a standardized mean difference (a *d* of 0 indicates no mean difference between two groups). Differences were larger for situational judgment tests in a paper-and-pencil format ($d = .40$) in comparison to a video format ($d = .33$). These differences were almost entirely moderated by the extent to which the situational judgment tests were correlated with measures of cognitive ability. One can also compute the vector correlations (Jensen, 1998) between the effect size (i.e., difference between groups) and the cognitive loading of the test. As shown in Table 9.4, the vector correlation between the effect size for African Americans and Whites and the *g*-loading of the situational judgment test was .88. This suggests that as the correlation of the situational judgment test with a measure of general cognitive ability increases (i.e., as the cognitive complexity increases), the mean ethnic score difference also increases.

There also was a moderating effect related to the personality variables of agreeableness and emotional stability. As shown in Table 9.4, as the correlation of the situational judgment test with agreeableness and emotional stability increased, the magnitude of the mean African American versus White score difference decreased. In brief, situational judgment tests show larger ethnic differences when the situational judgment test is positively related to cognitive ability and negatively related to agreeableness and emotional stability.

Whetzel et al. (2005) also reported that the gender difference was small ($d = .14$) and favored females. This difference was moderated somewhat by the correlation of the situational judgment test with conscientiousness, agreeableness, and emotional stability. As shown in Table 9.4, as the correlation between these personality variables and the situational judgment test increased, the gender difference favoring females also increased. These findings suggest that females obtained

[1]Portions of this section of the chapter were taken directly, and with permission, from Motowidlo, Hanson, and Crafts (1997).

slightly higher situational judgment test scores to the extent that the tests were correlated with conscientiousness, agreeableness, and emotional stability.

HOW TO DEVELOP A SITUATIONAL JUDGMENT TEST[1]

Although situational judgment tests come in a variety of formats, they all have the common feature that they present a description of a situation representing a problem or challenge that might be encountered at work. The items ask applicants how they would respond to the situation. The rest of this chapter offers suggestions for developing a situational judgment test. This developmental strategy combines practices that have been successfully followed in the past to develop demonstrably valid situational judgment tests. A shortcut approach also is provided. Other ways to build situational judgment tests also are possible, and they might be as good or even better than the approach described in this chapter.

There are three general stages for developing a situational judgment test. First, a panel of subject matter experts (SMEs) generates descriptions of problem situations that might happen at work. Second, the SMEs write multiple-choice response alternatives for each problem situation. Third, a scoring key is developed.

Develop Situational Item Stems

Situational item stems should represent classes of events that actually happen on the job. They should represent classes of problems or challenges that people have to handle effectively or their job performance will suffer. They do not have to reflect matters of critical or monumental importance, but they should not be so minor or trivial that it does not matter how people deal with them. Furthermore, they should be difficult enough that there are meaningful differences in how effectively different people handle them.

The item stems should be described in enough detail to provide the cues necessary to distinguish more effective from less effective ways of dealing with the situations, but not in so much detail that the cues point to a single correct response that will be obvious to everyone. These cues should be general enough so that they can be correctly interpreted even by people who have never encountered the situation, as long as they have encountered similar situations in different contexts.

The first step is to assemble groups of SMEs into a workshop and ask them to write critical incidents (see chap. 3). Using the electrician example, groups of experienced electricians or journeymen would be assembled. If results of a job analysis (see chap. 3) are available, performance dimensions (i.e., the competencies, knowledge, skills, or abilities needed for successful performance) are shared with workshop participants. The SMEs would then be asked to think about occasions when they, or someone they knew, encountered a problem in a situation that involved one of the performance dimensions. Using a critical incident form, they would be asked to write about each situational critical incident by: (a) describing the problem in full detail, (b) briefly noting how the electrician in the incident dealt with it, and (c) describing the results of the electrician's actions. If performance dimensions are

available, the participants also note which performance dimensions are related to the situation. Often a situation is related to several performance dimensions. An example of a critical incident, taken from chapter 3, is:

> The foreman of a job gave a print to an apprentice and said, "Tomorrow, lay this whole floor out and pipe it." The next day the apprentice realized that he did not know how to do the task and the foreman was not available. The apprentice reviewed available documentation until he learned what he needed to know. The apprentice successfully completed the job and felt proud.

The next step is to sort the critical incidents according to the content of the scenarios or problems that electricians must handle. Combining judgments of the different judges will show which situations tend to be grouped together most often and this will lead to definitions of situational categories. For instance, in the electrician example, two categories that could emerge might be Reading Blueprints and Completing Jobs on Time.

With these situational categories in hand, the next step is to select representative critical incidents from each one and edit them into situational judgment item stems. Selecting situations from each category helps to ensure that the final situational inventory will include examples of all the important kinds of situational problems that occur on the job. Normally, the final version of the situational judgment test will also contain multiple situations per situational category. The exact number of situations selected per category will be based on some rational procedure. For example, job experts may be asked to rate the importance of the situational categories, and situations from the more important categories will be more heavily represented on the test. Alternatively, individual situations may be mapped to the job performance categories, and importance ratings for these categories can be used to define a rule for selecting scenarios for the test. An example of a stem that could be developed from the previous critical incident is: "As an apprentice electrician you receive your work assignments and seek advice from your supervisor, the project foreman. You have encountered a problem that you cannot solve and you cannot locate your supervisor to help you." Note that some of the detail from the critical incident is omitted from the stem. The purpose is to make the stems general enough that most applicants will understand the content of the stem.

From a technical standpoint, the larger the number of situational judgment items in the final test the better, but practical considerations limit the number of items that can be included. Situational judgment items may take a minute or two to answer and the number of items in the final test should not exceed the time available to administer the test. If the final situational judgment test is to include no more than about 40 situational items, at least 50 to 60 problem situations should be prepared at this stage in the development process.

One can substantially shorten this process if a job analysis has been conducted that has identified performance dimensions. With verbally fluent incumbents, one can have the incumbents write item stems rather than critical incidents. Working from the list of performance dimensions, participants divide the dimensions among

themselves and write situational judgment stems. This abbreviated process assumes that the performance dimensions provide good coverage of the job and that the participants are willing and competent to write item stems. If situational item stems can be developed in this manner, it is sometimes possible to write all the item stems needed for the test in less than a day.

Develop Response Alternatives

Response alternatives should represent classes of broadly different strategies for handling each situation. The alternatives should all seem reasonable but some have to be more "correct" for the situation than others. The more correct alternatives should be more attractive to applicants with the best potential for success on the job.

One way to develop response alternatives with these characteristics is to collect responses to the situational item stems from job incumbents. This can be done by assembling the situational item stems one to a page in a questionnaire. The questionnaire should be administered to incumbents who would be asked to complete the questionnaire by writing a short description in the space provided of how they or someone they know would handle each problem. The goal is to provide a range of possible responses that vary in effectiveness. If there are too many problems for people to answer all of them, the problems can be divided into two or more shorter questionnaires, but at least five people should answer each problem to ensure that many potentially different kinds of responses are collected. Examples of responses to the stem above are:

1. Review the available documentation and identify the best approach.
2. Seek out another foreman to help you.
3. Work on something else until the foreman is available.
4. Try various approaches until you find the solution.
5. Go on break until the foreman is available.

Note that Response 1 is similar to that provided in the original critical incident.

Taking one situational problem at a time, the responses should be reviewed to identify a variety of strategies, without worrying at this point about which are the best and worst responses to the problem.

Develop a Scoring Key

Finally, a scoring key is developed by collecting judgments from SMEs about the effectiveness of the alternative response options for each situational judgment item. Typically, the test developer prepares a questionnaire asking SMEs to rate each response on an effectiveness scale. The questionnaire should be completed by individuals who are very experienced and knowledgeable about the job. A common procedure is to ask the most knowledgeable individuals possible to complete

Very Ineffective	Ineffective	Effective	Very Effective
This action is inappropriate. It will make the problem worse.	This is a poor action. It will not help solve the problem.	This is a reasonable action that would go far in resolving the problem.	This is one of the best and most effective actions of all possible actions.

FIGURE 9.2 Example of effectiveness rating scale.

−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

FIGURE 9.3 Scoring pattern for selecting the best or worst response.

the questionnaire. In the present example, they would be experienced journeymen or very senior electricians. In general, the more SMEs who contribute to the ratings, the more stable the ratings will be (i.e., the ratings will be less subject to individual idiosyncrasies). That said, approximately five to seven raters often are used. A possible effectiveness rating scale that the SMEs might use is shown in Fig. 9.2.

Using these expert judgments, the test developer computes the mean and standard deviations of the rating of each response option. The standard deviation is an indication of expert judgment agreement. Situational items for which there is little agreement among experts on the relative effectiveness of alternatives should be dropped. For remaining items, the experts' judgments would be used to identify the effective and less effective response alternatives for each situational judgment item.

Once the test developers have obtained mean ratings on each response, they can determine the scoring key. If a response instruction is used that asks the applicant to select one choice (e.g., pick the best response, what would you most likely do), the developer should declare the response with the highest mean effectiveness rating to be the correct response. For example, if the instructions ask the applicant to choose the best or worst response, the simple scoring pattern shown in Fig. 9.3 is recommended (McDaniel, Whetzel, & Nguyen, 2006). Likewise, if the response instruction asks for two responses (e.g., pick the best response and then pick the worst response; what would you most likely do and what would you least likely do), the mean effectiveness ratings should be used to identify the most effective and least effective response for the keyed responses.

−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

FIGURE 9.4 Scoring pattern for rating the effectiveness of possible responses.

On the other hand, if one seeks to have the applicant rate the effectiveness of each response, one could develop a scoring key based on the same 4-point scale shown in Fig. 9.2; however, the keying approach shown in Fig. 9.4 is recommended for several reasons (McDaniel, Whetzel, & Nguyen, 2006). First, it only requires that the incumbents who are providing ratings used to establish the key agree on whether the response option is an effective behavior or an ineffective behavior. Second, there are individual differences in how applicants understand relative statements (e.g., effective vs. very effective). Two incumbent raters might believe a given response option to be at the same level of effectiveness even though one rater describes it as "effective" whereas another describes it as "very effective." This difference is due to the rater's different interpretations of the word very. If the keying is based on this strategy, one can avoid dealing with the nuances of the word very.

However, McDaniel, Whetzel, and Nguyen (2006) recommended that a 4-point Likert rating scale, similar to the one shown in Fig. 9.2, be used in the actual situational judgment test instrument because applicants may feel constricted by 2-point, dichotomous rating scales. Thus, the use of a 2-point dichotomous rating scale is suggested (Fig. 9.4) for developing the answer key and a 4-point Likert scale (Fig. 9.2) is suggested for the actual situational judgment test instrument administered to examinees.

Another keying approach involves deviation scoring from the mean effectiveness rating (Legree, Psotka, Tremble, & Bourne, 2005). In this approach, the mean rating is used as the correct answer and ratings differing from the mean receive lower scores. For example, if the mean effectiveness rating of a response option is 1.5 and an applicant rates the response at "2," the applicant loses a half-point. Likewise, if the applicant rates the response a "1," the applicant also loses a half-point. Thus, the highest possible score is a 0 and the lowest possible is some negative number. One might want to add a positive number to all scores to make all scores positive.

SUMMARY

Situational judgment tests present descriptions of work situations that might happen on the job and ask applicants how they would handle them. They are based on the idea that people have different levels of knowledge about how best to handle

various work situations. By measuring this knowledge, situational judgment tests can predict job performance. The literature on the reliability of situational judgment tests is deficient because the measures tend to have heterogeneous content but measures of homogeneity are typically offered as estimates of reliability. The criterion-related validity of situational judgment tests is at useful levels. Both the criterion-related and construct validity of situational judgment tests are moderated by response instructions. Response instructions fall into two general categories, knowledge and behavioral tendency, and the choice of one or the other affects validity and the likelihood of subgroup differences.

Developing a situational inventory involves three general stages: (a) preparing descriptions of problem situations, (b) preparing multiple response alternatives for each problem, and (c) identifying the effectiveness of each response option. A set of procedures that could be followed in each stage to develop an effective situational judgment test is provided.

REFERENCES

Bruce, M. M. (1965). *Examiner's manual: Business Judgment Test.* Larchmont, NY: Author.

Bruce, M. M., & Learner, D. B. (1958). A supervisory practices test. *Personnel Psychology, 11,* 207–216.

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 Associates.

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.

Clause, C. S., Mullins, M. E., Nee, M. T., Pulakos, E. D., & Schmitt, N. (1998). Parallel test form development: A procedure for alternative predictors and an example. *Personnel Psychology, 51,* 193–208.

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

Clevenger, J. P., Jockin, T., Morris, S., & Anselmi, T. (1999, April). *A situational judgment test for engineers: Construct and criterion related validity of a less adverse alternative.* Paper presented at the 14th annual convention of the Society for Industrial and Organizational Psychology, Inc., Atlanta, GA.

Corts, D. B. (1980). *Development and validation of a test for the ranking of applicants for promotion to first-line federal trades and labor supervisory positions* (PRR-80-30). Washington, DC: U. S. Office of Personnel Management, Personnel Research and Development Center.

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

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

Dalessio, A. T. (1992, May). *Predicting insurance agent turnover using a video-based situational judgment test*. Paper presented at the seventh annual conference of the Society for Industrial and Organizational Psychology, Inc., Montreal, Canada.

Digman, J. M. (1990). Personality structure: Emergence of the five factor model. *Annual Review of Psychology, 41,* 417–440.

Doherty, M. (2005, July). Personal communication.

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. (1948). *How Supervise? Manual 1948 revision.* New York: The Psychological Corporation.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26–34.

Gottfredson, L. S. (2003). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence, 31,* 343–397.

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

Hanson, M. A., & Borman, W. C. (1989, April). *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 symposium conducted at the 4th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Hartman, N. S., & Grubb, W. L., III. (2005, November). *Situational judgment tests and validity: It's a matter of instruction*. Paper presented at the Southern Management Association, Charleston, SC.

Jensen, A. R. (1998). *The g factor*. Westport, CT: Prager.

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100) New York: Guilford Press.

Kirkpatrick, D. L., & Planty, E. (1960). *Supervisory Inventory on Human Relations.* Chicago: SRA.

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. 155–180). Berlin, Germany: Hogrefe & Huber.

Mandell, M. M. (1950). The administrative judgment test. *Journal of Applied Psychology, 34,* 145–147.

McDaniel, M. A., Hartman, N. S., & Grubb, W. L., III. (2003, April). *Situational judgment tests, knowledge, behavioral tendency, and validity: A meta-analysis*. Paper presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Inc., Orlando, FL.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: 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.

McDaniel, M. A., Whetzel, D. L., Hartman, N. S., Nguyen, N., & Grubb, W. L. (2006). Situational judgment tests: Validity and an integrative model. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 183–203). Mahwah, NJ: Lawrence Erlbaum Associates.

McDaniel, M. A., Whetzel, D. L., & Nguyen, N. T. (2006). *Situational judgment tests in personnel selection: A monograph for the International Personnel Management Association Assessment Council.* Alexandria, VA: International Personnel Management Assessment Council.

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

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640–647.

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

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

Nguyen, N. T. (2004, February). *Response instructions and construct validity of a situational judgment test.* Paper delivered at the 11th annual meeting of the American Society of Business and Behavioral Sciences, Las Vegas, NV.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking in a situational judgment test. *International Journal of Selection and Assessment, 13,* 250–260.

O'Connell, M. S., McDaniel, M. A., Grubb, W. L., III, Hartman, N. S., & Lawrence, A. (2002, April). *Incremental validity of situational judgment tests for task and contextual performance.* Paper presented at the 17th annual conference of the Society of Industrial and Organizational Psychology, Inc., Toronto, Canada.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., & Keenan, P. A. (1994, April). *The conflict resolution skills assessment: Model-based, multimedia measurement.* Paper presented at the ninth annual conference of the Society for Industrial and Organizational Psychology, Inc., Nashville, TN.

Phillips, J. F. (1992). Predicting sales skills. *Journal of Business and Psychology, 7,* 151–160.

Phillips, J. F. (1993). Predicting negotiation skills. *Journal of Business and Psychology, 7,* 403–411.

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.

Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of job performance ratings: An examination of ratee race, ratee gender, and rater level effects. *Human Performance, 9,* 103–119.

Ree, M. J., & Earles, J. A. (1993). *g* is to psychology what carbon is to chemistry: A reply to Sternberg and Wagner, McClelland, and Calfee. *Current Directions in Psychological Science, 2,* 11–12.

Roth, P. L., BeVier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54,* 297–330.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Smith, K. C., & McDaniel, M. A. (1998, April). *Criterion and construct validity evidence for a situational judgment measure.* Paper presented at the 13th annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., et al. (2000). *Practical intelligence in everyday life.* New York: Cambridge University Press.

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist, 50,* 912–927.

Stevens, M. J., & Campion, M. A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management, 25,* 207–208.

Thorndike, R. L., & Stein, S. (1937). An evaluation of the attempts to measure social intelligence. *Psychological Bulletin, 34,* 275–285.

Vasilopoulos, N. L., Cucina, J. M., Hayes, T. L., & McElreath, J. A. (2005, April). *Effect of situational judgment test response instructions on validity.* Paper presented at the 20th annual conference of the Society for Industrial and Organizational Psychology, Inc., Los Angeles.

Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology, 52,* 1236–1247.

Wagner, R. K., & Sternberg, R. J. (1991). *Tacit Knowledge Inventory for Managers: User manual.* San Antonio, TX: 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.

Weekley, J. A., & Ployhart, R. E. (2006). An introduction to situational judgment testing. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests.* Mahwah, NJ: Lawrence Erlbaum Associates.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. (2005). *Subgroup differences in situational judgment test performance: A meta-analysis.* Manuscript submitted for publication.