McDaniel, M.A., Whetzel, D.L., Hartman, N. S., Nguyen, N. & Grubb, W. L. (2006). Situational judgment tests: Validity and an integrative model. In R. Ployhart & J. Weekley (Eds). *Situational judgment tests: Theory, measurement, and application*. Jossey Bass. 183-204.

# 9

# Situational Judgment Tests: Validity and an Integrative Model

Michael A. McDaniel
*Virginia Commonwealth University*

Deborah L. Whetzel
*Work Skills First, Inc.*

Nathan S. Hartman
*John Carroll University*

Nhung T. Nguyen
*Towson University*

W. Lee Grubb, III
*East Carolina University*

This chapter offers insights and data concerning factors that can affect the construct and criterion-related validity of situational judgment tests (SJTs). First, we review the history of the debates concerning the validity of SJTs. Next, we review four characteristics of SJT items that are logically related to issues of construct and criterion-related validity. Then, we summarize evidence on construct validity, criterion-related validity, and incremental validity of SJTs. Next, we present a model that integrates the findings.

Finally, we offer topics for future research concerning the construct and criterion-related validity of SJTs.

## VALIDITY DEBATES CONCERNING SJTs ARE AS OLD AS SJTs

The development and use of SJTs have a long history in personnel selection. In fact, the use of SJTs dates back to the 1920s. For a detailed history of SJTs, the reader is referred to the first chapter of this book and McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001). Despite the long history of these tests, substantial attention to the construct and criterion-related validity has been a recent phenomenon. Still, some early work foreshadows recent debates about the validity of SJTs, particularly their construct validity. The first test known to use situational judgment items was the George Washington Social Intelligence Test. One of the subtests, called Judgment in Social Situations, included many items in work situations. This subtest required "keen judgment, and a deep appreciation of human motives, to answer correctly" (Moss, 1926, p. 26). Thorndike and Stein (1937) began the construct validity debate in SJTs by arguing that the George Washington Social Intelligence Test did not correlate with other tests designed to measure social intelligence and was highly correlated with measures of general intelligence. Cardall (1942) offered the Practical Intelligence Test, but some researchers criticized it for its high correlations with measures of general intelligence (Carrington, 1949; Taylor, 1949). Northrop (1989) reviewed World War II-era SJTs and made observations on their complex construct domain. More recently, Sternberg et al. (2000) made bold claims about the construct validity of practical intelligence measures that are composed of situational judgment items. These claims have received substantial critique (Gottfredson, 2003; McDaniel & Whetzel, 2005).

## CHARACTERISTICS OF SJTs THAT MIGHT AFFECT VALIDITY

There are several characteristics of SJTs that may affect their validity. This section draws on and expands the McDaniel and Nguyen (2001) discussion of the characteristics of SJTs. These characteristics include the degree of fidelity of the SJT, the cognitive complexity of the items, the content of the scenario, and the response instructions. A discussion of some of these characteristics is also relevant to issues of development and scoring of SJTs and is addressed in chapter 8 of this volume. In our chapter, these characteristics are presented as they relate to the construct and criterion-related validity of SJTs.

## Degree of Fidelity

SJTs vary in the degree of fidelity of presentation. Fidelity can be defined as the extent to which the test resembles the work required on the job. For example, work samples would have high fidelity and cognitive ability tests would have low fidelity. Typically, video-based SJTs are considered to have higher fidelity than written SJTs. However, within a given format, one might expect fidelity to vary across SJTs.

Video-based SJTs consist of a series of situations or scenarios in a video media as well as some number of plausible response options, also on video media. Several such measures have been developed to predict customer service performance. For example, Alignmark's AccuVision Customer Service System (Alignmark, 2001) is a video-based selection test that has been developed to measure customer service skills and abilities. Job-analysis information provided the basis for the design and content of the system and varying degrees of customer contact are represented by the different scenarios (e.g., employee-to-customer, employee-to-employee, and telephone interactions).

We believe video-based SJTs are popular for three reasons. First, video simulations have stronger visual and affective appeal than written SJTs. Typically, video-based SJTs are more interesting to view than written tests are to read. Thus, employers who use video-based SJTs can speak of how innovative and technologically sophisticated their selection systems are. Likewise, applicants can be impressed by the sophistication of the selection system and be attracted to the organization. Second, high-fidelity simulations are valued because of their philosophy of measurement. They reflect the sample approach to measurement as opposed to sign measurement (Wernimont & Campbell, 1968). In their review of inferential and evidential bases of validity, Binning and Barrett (1989) discussed this measurement approach as one seeking to demonstrate that the predictor is isomorphic with the performance domain. Based on this measurement philosophy, many academics and practitioners shun low-fidelity instruments such as cognitive ability tests and embrace higher fidelity methods (e.g, video SJTs, work sample tests, and assessment centers) because they are considered samples of job content and are viewed as intrinsically better selection tools. Sometimes this belief is based on evidence and sometimes it is not. Third, video-based based SJTs are popular because they are perceived to have lower Black–White racial differences. (Nguyen, McDaniel, & Whetzel, 2005) The rationale is that video-based testing formats reduce the cognitive variance in the test by reducing the reading demands. Given the large mean differences between Blacks and Whites in cognitive ability (Roth, BeVier, Bobko, Switzer, & Tyler, 2001), reducing

the cognitive variance in a test likely will reduce Black–White score differences on the test. Reducing the cognitive demands is a common approach to manipulating racial outcomes in test performance (Gottfredson, 1996). Reducing Black–White differences in test scores is a common motivation for developing video-based SJTs. We have often seen them used by employers facing race-based employment litigation. Developers of video-based SJTs promote them as tests designed to minimize adverse impact.

Typically, however, SJTs are presented in a written format, whether as paper-and-pencil or computer-administered tests. In written-format SJTs, the scenario is described rather than shown via video. Written SJTs are developed more frequently due to the cost differences in development between written and video formats and the need for expanded skills sets for video production. Although both formats share the cost of developing the scenarios and response options, video-based tests require costs associated with scripting, actors, and video production. These factors make video-based SJTs substantially more expensive to develop than paper-and-pencil measures. Personnel psychologists typically have the skills needed to develop SJTs in written format. Video-based SJTs typically would require the personnel psychologist to obtain assistance from those with video competencies. Separate from the cost considerations, video-based SJTs require the test-development project to involve more people and a longer time to completion. Thus, simplicity of production likely makes written format SJTs more common than video-based SJTs.

In summary, the fidelity of the SJT might affect both construct and criterion-related validity. A video-based SJT has a reasonable chance of reducing the cognitive variance of the SJT primarily by reducing the reading demands. Thus, in contrast to written SJTs, video-based SJTs can be expected to have a lower correlation with cognitive ability and yield lower Black–White test score differences. If the construct validity of the SJT changes with fidelity, one might observe criterion-related validity differences among SJTs of varying fidelity. Cognitive ability is the best predictor of job performance (Schmidt & Hunter, 1998). If one removes the cognitive variance from a test, one runs the risk of reducing its prediction of job performance. However, the remaining noncognitive variance may also have useful levels of validity. We suggest that if one removes cognitive variance from a SJT, the SJT will be correlated with personality traits such as conscientiousness, agreeableness, and emotional stability as well as job knowledge. All of these constructs can contribute to the prediction of job performance. Also, the high fidelity of the video might contextualize the measurement properties in a manner to enhance the validity of the test. We recognize that this assertion is speculative, but there is some evidence to suggest that adding contextualization to personality tests can improve

validity (Robie, Born, & Schmit, 2001; Robie, Schmit, Ryan, & Zickar, 2000; Schmit, Ryan, Stierwalt, & Powell, 1995). Thus, it is reasonable to suggest that if high-fidelity SJTs offer more work-related contextualization than low-fidelity SJTs, the greater contextualization of high-fidelity SJTs may add to their validity. However, we know of no evidence that examines the relationship between the fidelity of an SJT and its validity and find the topic worthy of investigation.

## Cognitive Complexity

SJTs vary in the level of cognitive complexity of the scenarios. We offer this example as a low-complexity situational judgment item:

> Everyone in your workgroup but you has received a new computer. Nobody has said anything to you about the situation.
>
> a. Assume it was a mistake and talk to your supervisor
> b. Take a computer from a coworker's desk
> c. Confront your boss about why you are being treated unfairly
> d. Quit

The Tacit Knowledge Inventory for Managers (TKIM; Wagner & Sternberg, 1991), on the other hand, presents scenarios that are considerably longer and more detailed than the typical SJT. The TKIM presents situations that are typically several paragraphs long and involve fairly complex interactions. Given the increased reading and reasoning requirements imposed by complex scenarios, they are likely to be more highly correlated with cognitive ability than other SJTs composed of lower complexity scenarios. McDaniel and Nguyen (2001) suggested that the complexity of the situation is related to the length of the scenario and that more words are typically required to describe complex situations than less complex situations. A related issue concerns the comprehensibility of the stems. It is harder to understand the meaning and import of some situations than other situations. Sacco et al. (2000) examined the comprehensibility of item stems using readability formulas and found variability in the reading demands. It is a reasonable assertion that the length, complexity, and comprehensibility of the situation are interrelated and may drive the cognitive loading of the situational stems.

The cognitive complexity of the items in a SJT can be expected to affect both construct and criterion-related validity. Cognitively demanding SJTs relative to less cognitively demanding SJTs can be expected to have relatively more of their variance driven by cognitive ability and less of their variance associated with noncognitive traits. This difference in the relative

degree of cognitive and noncognitive constructs assessed might impact the criterion-related validity. Whereas cognitive tests have higher validity than personality tests, on average, for most job performance criteria, increasing the cognitive loading of the SJT might increase its criterion-related validity. The Sacco et al. (2000) data supported this assertion. However, a SJT may not be the sole screening tool. If one increases the cognitive load of a SJT, the SJT may contribute little additional variance to a battery containing a cognitive ability test. In contrast, a SJT with lower cognitive load, may have a lower correlation with cognitive ability and provide better incremental validity to a battery containing a cognitive ability test.

### Content of the Scenario

There is variability in the content of scenarios of SJTs. Some scenarios use interpersonal situations that do not require any kind of job knowledge. Others involve job knowledge that is fairly generic (e.g., decision making or problem solving) and others may involve technical job knowledge (e.g., knowledge of electronics or mechanics).

We suggest that the content of the scenario affects both construct and criterion-related validity and offer the following observations. It is difficult to assess the correlates of situational judgment items by inspecting their content. One's response to an item concerning an interpersonal conflict would certainly reflect one's interpersonal competencies but it may also reflect other traits. In interpersonal conflicts some possible responses reflect stupid behavior and others reflect behavior driven by complex reasoning. Some responses might reflect low or high levels of conscientiousness or emotional stability. In our sample item concerning not receiving a new computer, endorsement of the response of taking a new computer from a co-worker's desk might simultaneously reflect low cognitive ability, low conscientiousness, and low emotional stability. However, it is reasonable that content must drive constructs assessed. SJT items that present interpersonal scenarios should tap interpersonal constructs to a greater degree than scenarios focused on resolving problems with automobile engines. It is also reasonable that the content assessed must drive the criterion-related validity. Some content is more job-related than others and some content will better predict job performance than other content.

### Response Instructions

Two different kinds of response instructions are typically used for SJTs: behavioral tendency and knowledge. Tests with behavioral tendency instructions ask respondents to identify how they would likely behave

in a given situation. These instructions, asking people about their typical behavior, look like personality measures that ask people to describe their behavioral inclinations. A variant of this approach is to ask the respondent to identify the response they would most likely perform and those they would least likely perform (Dalessio, 1994; Smith & McDaniel, 1998).

SJTs with knowledge instructions ask respondents to evaluate the effectiveness of possible responses to a particular situation. These judgments take the form of effectiveness ratings or rankings of best response and/or worst response. Alternatively, an applicant might be asked to identify the best response or the best and worst response.

McDaniel and Nguyen (2001) suggested that knowledge instructions make the SJT more faking resistant than behavioral tendency instructions. Consider, as a metaphor, the distinction between a personality item and a mathematical knowledge item. As with a SJT with behavioral tendency instructions, personality test instructions encourage respondents to describe their typical behavior. In the assessment of conscientiousness, an item might ask if the respondent maintains an orderly work area. A disorderly person could answer truthfully or could falsely indicate that he or she maintains a neat work area. As with a SJT using knowledge instructions, a respondent on a mathematics knowledge item is asked to provide the best answer. In the assessment of cube-root knowledge, a mathematics test might ask the respondent for the cube-root of 4,913. A mathematically deficient person cannot pick the correct answer with certainty. The individual might correctly guess the answer (17) but he or she cannot knowingly provide the correct answer as can the disorderly person who claims that he or she is orderly. Thus, in a SJT with behavioral intention instructions it is possible for an ill-suited applicant to answer in a manner that mimics an applicant with highly desirable behavioral tendencies. In a knowledge instruction SJT, it is more difficult for an applicant who lacks knowledge of the best responses to mimic a knowledgeable applicant. Nguyen, Biderman, and McDaniel (in press) provided preliminary evidence that faking can raise scores on a SJT with behavioral tendency instructions to a much greater extent than on a SJT with knowledge instructions.

Response instructions are likely to affect construct validity. Responses to behavior tendency items describe typical behavior when applicants are responding without the intent of distorting. These responses are similar to personality items and are likely to assess or be correlated with personality dispositions. On the other hand, knowledge items assess one's knowledge of the best way to behave. Assessments of knowledge generally have cognitive correlates. Job knowledge is gained as a result of opportunities and ability to learn. The opportunity to learn and the ability

to learn may be influenced by personality dispositions. For example, anxiety may impede learning or cause one to avoid situations that offer the opportunity to learn and introversion may impede one's knowledge acquisition through an avoidance of active learning situations (e.g., introverts may have less knowledge gained through public speaking). However, knowledge acquisition places large demands on cognitive skills. Thus, SJTs using behavioral tendency instructions may correlate higher with personality traits and personality-loaded criteria (contextual performance) than SJTs using knowledge instructions that may correlate more highly with cognitive ability tests and cognitively loaded criteria (task performance).

Response instructions likely influence criterion-related validity in two ways. The first way relates to differential faking between response formats. We made arguments and cited preliminary evidence that SJTs using behavioral tendency instructions may be more readily faked than SJTs using knowledge instructions. If this is true, and to the extent that some respondents fake and faking diminishes validity, one would expect SJTs with behavioral tendency instructions to have lower criterion-related validity than SJTs with knowledge instructions. This validity difference should be more pronounced in applicant samples than in incumbent samples because incumbents typically have less motivation to fake.

The second way in which response formats might affect criterion-related validity is through their effects on constructs assessed. Whereas cognitive ability tests have substantially larger correlations with job performance than personality tests (Schmidt & Hunter, 1998), SJTs with knowledge instructions, with their associated larger cognitive load, may yield higher validities than SJTs with behavioral tendency instructions. On the other hand, a SJT with behavioral tendency instructions might provide a reliable measure of cognitive ability and supplement cognitive variance with job-related personality variance. Thus, an SJT with behavioral tendency instructions may have similar levels of validity as a test battery containing both a cognitive ability test and a set of personality measures. Optimally weighted composites of cognitive and personality measures can yield higher validities than cognitive tests alone (Schmidt & Hunter, 1998).

## THE EVIDENCE FOR CONSTRUCT AND CRITERION-RELATED VALIDITY

So far, we have offered speculation on factors that may be associated with the construct and criterion-related validity of SJTs. There has been substantial research on the construct and criterion-related validity of SJTs that addresses some but not all of the factors on which we have speculated.

Here, we present a summary of empirical evidence addressing the construct and criterion-related validity of SJTs.

## Construct Validity Evidence

Several classic primary studies have been conducted documenting the validity of SJTs (e.g., Chan & Schmitt, 1997; Motowidlo, Dunnette, & Carter, 1990; Olson-Buchanan et al., 1998; Smith & McDaniel, 1998). Given the attention of SJTs in the recent psychological literature, meta-analyses have been conducted on the construct validity of SJTs (McDaniel, Hartman, & Grubb, 2003; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel & Nguyen, 2001). McDaniel et al. (2001) examined the cognitive correlates of SJTs. McDaniel and Nguyen (2001) summarized personality and job-experience correlates of SJTs. McDaniel et al. (2003) extended the construct-validity analyses of McDaniel and Nguyen (2001) and McDaniel et al. (2001) with more data and the inclusion of a response-instruction moderator. The construct validity highlights of the McDaniel et al. (2003) effort are shown in Table 9.1.

SJTs are shown to measure cognitive ability and the Big Five personality traits to varying degrees. The extent to which SJTs measure these constructs

**TABLE 9.1**

Meta-Analytic Results of Correlations Between Situational Judgment Tests and Cognitive Ability, Agreeableness, Conscientiousness, and Emotional Stability

| *Distribution of correlations with SJTs* | *N* | *No. of rs* | $\rho$ |
|---|---|---|---|
| Cognitive ability | 22,553 | 62 | .39 |
| Behavioral tendency instructions | 5,263 | 21 | .23 |
| Knowledge instructions | 17,290 | 41 | .43 |
| Agreeableness | 14,131 | 16 | .33 |
| Behavioral tendency instructions | 5,828 | 11 | .53 |
| Knowledge instructions | 8,303 | 5 | .20 |
| Conscientiousness | 19,656 | 19 | .37 |
| Behavioral tendency instructions | 5,902 | 11 | .51 |
| Knowledge instructions | 13,754 | 8 | .33 |
| Emotional Stability | 7,718 | 14 | .41 |
| Behavioral tendency instructions | 5,728 | 10 | .51 |
| Knowledge instructions | 1,990 | 4 | .11 |

is moderated by the SJT response instructions. McDaniel et al. (2003) found that for the three most researched personality constructs, SJTs with behavioral tendency instructions are more correlated with personality than SJTs with knowledge instructions (Agreeableness .53 vs. .20; Conscientiousness .51 vs. .33; Emotional Stability .51 vs. .11). SJTs with knowledge instructions are more correlated with cognitive ability than SJTs with behavioral tendency instructions (.43 vs. .23). They noted that some of the distributions had relatively small numbers of coefficients and that results should be replicated as more data cumulate.

As shown in Table 9.1, the primary correlates with SJTs are cognitive ability, agreeableness, conscientiousness, and emotional stability. Given the wide range of constructs correlated with SJTs, their multifaceted nature does not make it easy to target them to specific constructs to the exclusion of other constructs. Efforts at such targeting have not been fully successful (Ployhart & Ryan, 2000). Our reading of this literature is that SJTs can be targeted to assess specific constructs but that the targeted SJTs will continue to measure multiple constructs.

In summary, SJTs are typically correlated with cognitive ability, agreeableness, conscientiousness, and emotional stability. SJTs with behavioral tendency instructions are more correlated with personality than SJTs with knowledge instructions. SJTs with knowledge instructions are more correlated with cognitive ability than SJTs with behavioral tendency instructions. These findings suggest that one can change the construct validity of a situational judgment test by altering the response instructions. Furthermore, the finding that SJTs have moderate correlates with personality and cognitive ability suggests that the tests are best viewed as methods and not assessment of a judgment construct.

We note that the construct-validity data are drawn almost entirely from concurrent studies. Thus, for studies using incumbents, it is clear that SJTs with behavioral tendency instructions appear to assess more personality and less cognitive ability, whereas the opposite is found for SJTs with knowledge instructions. Applicant data, when they become available, may show a less clear distinction between behavioral tendency instructions and knowledge instructions. We believe the cause of any potential difference will be due to applicant faking. It is reasonable that some applicants will respond in a manner to make themselves look better than they are and other applicants will not. It is also reasonable that among those applicants who choose to fake, some will be better at faking than others. We argue that applicants will have more difficulty faking the SJTs using knowledge instructions than the SJTs using behavioral tendency instructions. Deceitful applicants may not provide a response indicative of their behavioral tendency but rather a response consistent with their perception of what is

the best response. As such, deceitful applicants will respond to a behavioral tendency SJT as if it had the instructions of a knowledge SJT. To the extent that this happens with some frequency among applicants, we anticipate that the construct validity differences between behavioral tendency and knowledge instruction SJTs will be smaller because the results for the behavioral tendency SJTs will become more similar to the results for knowledge instruction SJTs. Thus, an SJT that is more personality loaded in an incumbent sample, may become more cognitively loaded in an applicant sample.

### Criterion-Related Validity Evidence

As a result of the large number of SJT instruments and studies conducted to assess their validity, McDaniel et al. (2001) conducted a meta-analysis to determine the criterion-related validity of these kinds of instruments. McDaniel et al. (2003) re-analyzed and updated the 2001 data and found that knowledge-response instructions yielded higher validity (.33) than behavioral tendency instructions (.27). We note that this is not a large magnitude moderator and is based primarily on concurrent studies with incumbent data. Earlier, we speculated that the construct validity evidence based on applicant samples may be somewhat different from the construct validity data using incumbent samples. We hold the same caveat for the criterion-related validity data. To the extent that some applicants fake when completing SJTs, faking respondents on SJTs with behavioral tendency instructions may respond not with behavioral dispositions but with their perception of the best answer. To the extent that this happens, the criterion-related validity of behavioral tendency SJTs may approximate those of SJTs with knowledge instructions. We recognize that this argument may lead to the conclusion that faking can enhance validity. This is not a position with which we are comfortable or which we endorse and we encourage research to evaluate this possibility.

### Incremental Validity Evidence

The incremental validity of SJTs over measures of cognitive ability has been a topic of several primary studies (Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt Harvey, 2001; Chan & Schmitt, 2002; O'Connell, McDaniel, Grubb, Hartman, & Lawrence, 2002; Weekly & Jones, 1997, 1999) and two meta-analyses (McDaniel et al., 2001; McDaniel et al., 2003). All studies showed that SJTs provide incremental validity over cognitive ability. The incremental prediction is reasonable in that SJTs typically measure job-related personality traits including conscientiousness, agreeableness, and

emotional stability. Whereas SJTs are measurement methods and can measure a variety of constructs in varying magnitudes, the incremental validity of SJTs over cognitive ability can expect to vary with the cognitive saturation of the SJT. SJTs that are highly correlated with cognitive ability can not be expected to have much incremental validity over cognitive ability. SJTs that measure noncognitive job-related constructs can be expected to have useful levels of incremental validity over cognitive ability.

Data on incremental validity of SJTs over both cognitive ability and personality are rare. O'Connell et al. (2002) found incremental validity of the SJT over cognitive ability but very little incremental validity over both cognitive ability and personality. They did not report incremental validity of the SJT over personality alone. Whereas SJTs typically measure both personality and cognitive ability, one might expect an SJT to have incremental validity over cognitive ability or over personality. However, it is likely to be more difficult for SJTs to have incremental validity over both cognitive ability and personality. Still, Weekley and Ployhart demonstrated that a SJT provided incremental validity beyond cognitive ability, personality, and experience. Future research should examine whether this is a rare or a common event.

## A MODEL TO INTEGRATE CONSTRUCT AND CRITERION-RELATED VALIDITY EVIDENCE FOR SJTS

Figure 9.1 presents a model that integrates validity evidence for SJTs. The four personal traits that affect performance on SJTs are cognitive ability, agreeableness, conscientiousness, and emotional stability. This assertion is consistent with the correlations between measures of these constructs and SJTs. The four personal traits also affect the extent to which job knowledge is gained through education and training. For example, the smart and dependable gain more job knowledge through education and training than do the stupid and the slothful. The four personal traits also affect the extent to which one gains job knowledge through job experience. We have divided job knowledge into general job knowledge and technical job knowledge. We envision general job knowledge to be composed of basic knowledges common to most jobs and might be viewed as work socialization knowledge. These knowledges would include the value of showing up to work, being nice to co-workers, dressing appropriately, following the directions of one's supervisor, and refraining from employee theft and inappropriate language. These are knowledges that one might gain in one's initial job and might show more variance among applicants for entry-level positions than more senior positions. However, general job knowledge might

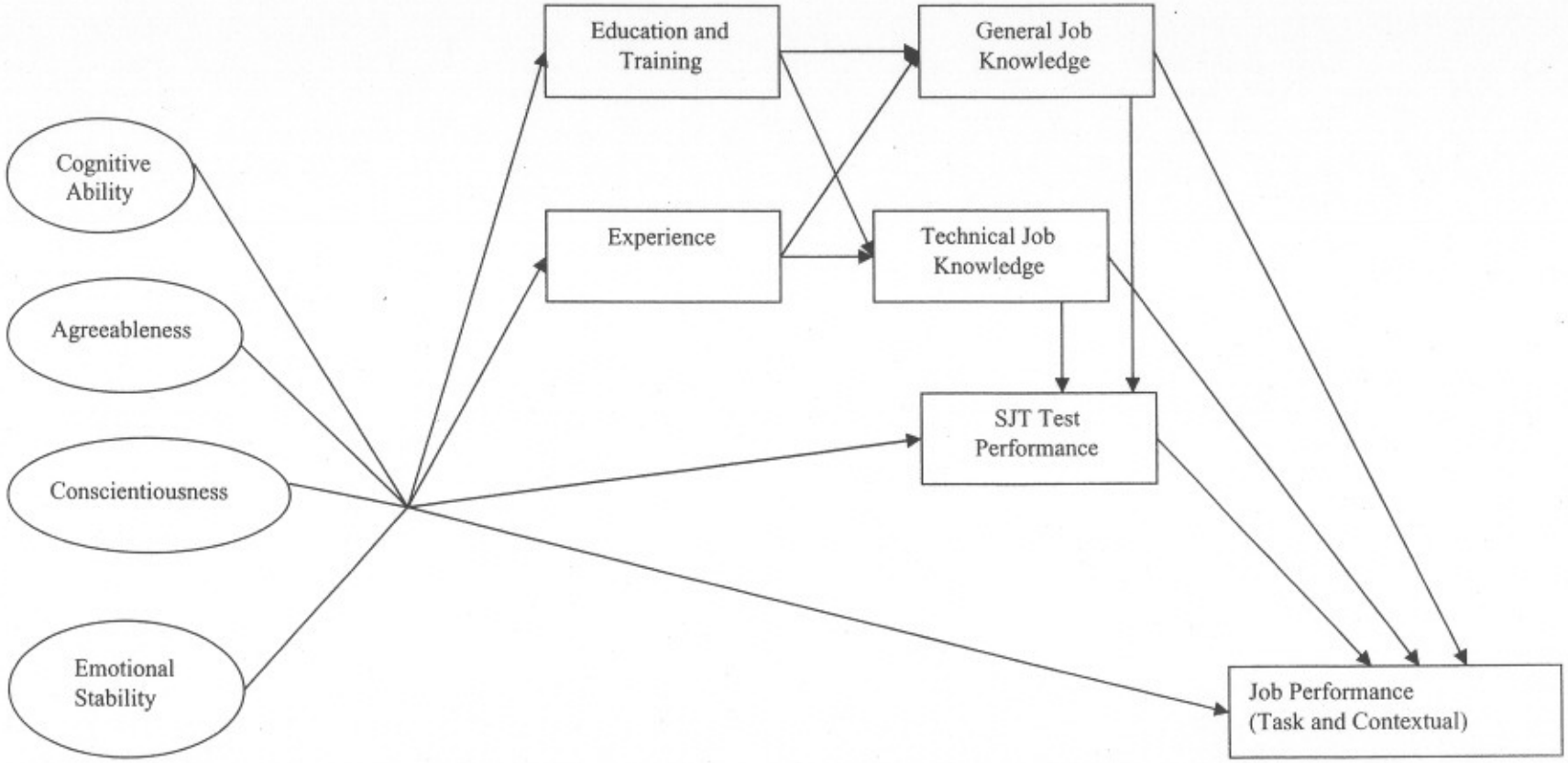


FIG. 9.1. A conceptual model of the factors affecting the construct and criterion-related validity of situational judgment tests.

also include some knowledges needed at supervisory levels. For example, some supervisors have more knowledge related to dealing with subordinates than others. Many SJTs are designed to tap general supervisory knowledge. Technical job knowledge is knowledge specific to a job or an industry and is gained through education, training, and experience. Some SJTs might tap technical knowledge such as techniques useful in closing a sale or knowledge related to managing a large project. There is little research addressing knowledge correlates of SJTs but SJTs are often assumed to measure knowledge, tacit or explicit (Schmidt & Hunter, 1993; Sternberg et al., 2000).

Our model assumes that test performance on SJTs is a function of cognitive ability, agreeableness, conscientiousness, and emotional stability, both directly and through their effects on job knowledge as mediated by education, training, and experience. Job knowledge, both general and technical, is assumed to directly affect performance on SJTs.

Job performance is predicted by cognitive ability, agreeableness, conscientiousness, emotional stability, SJTs, and job knowledge. Whereas different SJTs assess the four personal traits and job knowledge to varying degrees, the incremental validity of a specific SJT will vary based on its correlates. SJTs with substantial cognitive correlates may have little incremental validity over cognitive ability but substantial incremental validity over personality. SJTs with substantial noncognitive correlates may have little incremental validity over personality but substantial incremental validity over cognitive tests. SJTs with both high cognitive and noncognitive saturation, may offer little incremental validity over a battery of cognitive and noncognitive tests.

The model also recognizes that job performance can be measured with varying emphases on task and contextual performance. We anticipate that knowledge-based SJTs will be better predictors of task performance than behavioral tendency SJTs. If the job performance construct space were weighted to be more contextually oriented, one would expect an increase in the validity of behavioral tendency SJTs.

Our model is not rocket science, but it is meant as a heuristic for understanding the construct, criterion-related, and incremental validity of SJTs. The model's separation of job knowledge into general and technical is speculative. The model's description of SJT variance as a function of only cognitive ability, three personality traits and knowledge may be too restrictive. The model also needs to be expanded to consider differential loading of job performance on task and contextual factors. However, the model is consistent with available data, can aid in understanding the roles of cognitive and personality constructs in SJTs, and can facilitate understanding of the differences across studies in incremental validity. Certainly,

more theory and better models of the nomonological net of SJTs can be developed.

## FUTURE RESEARCH

Although we know much about the construct and criterion-related validity of SJTs, we believe that research on SJTs is in its infancy and that there are a large number of issues that need further attention. We offer eight areas in need of additional research.

### Response Instructions and Item Content

Our discussion of the effects of response instructions on criterion-related and construct validity assumes that there is nothing different in the content of the items in knowledge-instruction versus behavioral tendency instruction SJTs. It could be that all or part of the response instruction effect is actually due to differences in item content. Nguyen et al. (in press), using a single set of SJT items, found the expected effects with respect to faking and correlations with cognitive ability. Specifically, a SJT with knowledge instructions was less fakable and more correlated with cognitive ability than was the same SJT with behavioral tendency instructions. However, that study did not examine correlations with personality or job performance. The chapter authors have done a small sample study in which raters were asked to guess whether SJT items were from a knowledge instruction or a behavioral tendency instruction SJT and found that raters were not able to make accurate decisions. Thus, these results argue against an item-content confound. Despite this preliminary evidence against an item-content confound, more study of item-content and other possible confounds is warranted. We recognize that some will not find it credible that one can change the construct and criterion-related validity of a test simply by changing the response instructions. More research is needed to evaluate our assertions about the impact of response instruction effects on SJTs.

### Job Knowledge

Most SJT researchers assume that SJTs measure knowledge and some explicitly assert this (Schmidt & Hunter, 1993; Sternberg et al., 2000). Yet no research has adequately assessed job-knowledge correlates of SJTs. McDaniel and colleagues (McDaniel & Nguyen, 2001; McDaniel et al., 2003) have summarized the small literature that has related SJTs to length of job

experience as a remote surrogate measure of job knowledge. More research is clearly needed.

## Constructs of a Moving Target

Most researchers acknowledge that SJTs are measurement methods that can and do measure multiple constructs. Different SJTs can and do measure constructs differentially. The Stevens and Campion (1999) teamwork SJT has substantial correlates with cognitive ability, whereas other SJTs have much lower correlations with cognitive ability. Likewise, some SJTs have large correlations with some personality traits and others have lower correlations. Thus, any attempt to draw conclusions about the constructs assessed by SJTs must recognize that these are "constructs assessed on the average" and that any given SJT can deviate from these typical correlations. Much more work needs to be done to better target the constructs measured by SJTs. Ployhart and Ryan (2000) offered a very reasonable method for constructing SJTs to measure specific constructs. Although the SJT scales did measure the intended constructs, the correlations with other measures of the constructs were not large and the discriminant validity of the measures was low. Given that a reasonable approach to building SJTs to target specific constructs yielded less than satisfying results, we are not hopeful that SJTs can be readily targeted to specific constructs to the exclusion of others. However, research to refute our pessimism is encouraged.

## Search for More Constructs

We have offered evidence that SJTs in part assess cognitive ability, conscientiousness, agreeableness, and emotional stability. We have also speculated that SJTs tap job knowledge and encouraged more research of SJTs and job knowledge. Sternberg and colleagues (2000) argued that one can measure a general factor of practical intelligence with situational judgment items. Although we agree with Gottfredson (2003) and McDaniel & Whetzel (2005) that the available evidence finds no support for the existence of a practical intelligence construct let alone a general factor of practical intelligence, we do encourage more theory and research concerning other constructs that are or can be assessed by SJTs.

We note that our findings that SJTs typically are correlated with cognitive ability, conscientiousness, agreeableness, and emotional stability may be a function of the jobs examined to date and might differ for other jobs. For example, SJTs designed for sales jobs could reasonably have large magnitude

correlations with extroversion. It is worthwhile to examine the extent to which job content moderates the correlates of SJTs.

## Fidelity and Validity

All validity evidence cumulated by McDaniel and colleagues (McDaniel et al., 2001; McDaniel & Nguyen, 2001; McDaniel et al., 2003) has been restricted to written SJTs. As more validity data on video-based SJTs become available, cumulative evidence on their construct and criterion-related validity should be summarized. In this chapter, we have also restricted fidelity to describe the difference between video and written formats. This is a constrained definition of fidelity. We have also considered fidelity for its effect on the cognitive load of SJTs. We suspect fidelity has more import than its effect on cognitive load. We also speculated on the contextualization of items as an aspect of fidelity that might affect validity. The notion of fidelity and its relation to validity needs further examination.

## Applicant Data

Almost the entire knowledge base concerning the construct and criterion-related validity of SJTs is based on concurrent data. Concurrent studies use incumbents as respondents and the task of completing the SJT is typically explained as a research effort that does not affect the career of the respondents. Applicants complete SJTs under much different situations that serve to manipulate the motivation of the respondents. Compared to incumbents, applicants are much more likely to be concerned about the evaluations of their responses. Motivational and other differences between applicants and incumbents may affect the construct and criterion-related validity of SJTs. In this chapter, we speculated that the differences between behavioral tendency and knowledge instruction SJTs may be less pronounced in applicant data than the results we described.

## Cognitive Loading

We have speculated about cognitive loading of SJT items and its effect on Black–White score differences, construct validity, and criterion-related validity. There is very little research (Sacco et al., 2000; Nguyen et al. 2005) that has directly assessed cognitive loading of items on their racial differences and validity. Clearly, more research is needed.

### Publication Bias

Developments in meta-analytic methods in industrial/organizational (I/O) psychology has largely stalled. However, meta-analysis in medical research is a hot-bed of methodological developments. Foremost among these developments are statistical techniques to evaluate publication bias in literatures (Dickerson, in press; Halpern & Berlin, in press). Fail-safe Ns, the most common approach to publication bias assessment in I/O psychology has been shown to be a very poor procedure to assess publication bias (Becker, in press) and a number of more accurate and powerful approaches have been offered (Duval, in press; Hedges & Vevea, in press; Sterne & Egger, in press; Sutton & Pigott, in press). Personnel psychologists may ignore this research to their eventual detriment. Authors who have results counter to the typical results may be discouraged from submitting them for publication because the results do not fit the "known facts." Likewise, editors who receive small sample, low-validity studies are likely to reject them as flawed due to sample size and other factors because they are counter to "known facts." This is not a situation that is conducive to the advancement of cumulatetive live knowledge. Consistent with many publication bias investigations, one might expect small sample, low-validity studies to be suppressed. If this were the case, there would be an upward bias in the reported validity of SJTs and possibly some effect on the correlations with other measures relevant to construct validity. Current validity data on SJTs should be evaluated for potential publication bias and efforts to evaluate and prevent publication bias (Berlin & Ghersi, in press) in this and other I/O literatures is warranted.

## SUMMARY

This examination of SJT construct validity has focused primarily on cognitive ability, conscientiousness, emotional stability, and agreeableness. We suggest that job knowledge is also an important correlate of SJT performance. The criterion-related validity of SJTs is well established, which is reasonable given that the constructs associated with SJTs also tend to be useful predictors of job performance. We offer evidence that the instructions used with SJTs can moderate the construct and criterion-related validity of the measures. We also provide a heuristic model that helps to integrate the relations between other constructs, SJTs, and job performance. Finally, we note several areas of research that will help advance our knowledges of SJTs.

## REFERENCES


Alignmark. (2001). *AccuVision customer service system validation Report*. Maitland, FL: Author.

Becker, B. J. (in press). The failsafe N or file-drawer number. In H. Rothsetin, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments*. New York: Wiley.

Berlin, J. A., & Ghersi, D. (in press). Preventing publication bias: Registries and prospective meta-analysis. In H. Rothsetin, A.J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments*. New York: Wiley.

Binning, J. F., & Barrett, G.V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478–494.

Cardall, A. J. (1942). *Preliminary manual for the Test of Practical Judgment*. Chicago: Science Research Associates.

Carrington, D. H. (1949). Note on the Cardall Practical Judgment Test. *Journal of Applied Psychology, 33*, 29–30.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143–159.

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

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Schmidt Harvey, V. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410–417.

Dalessio, A. T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology, 9*, 23–32.

Dickerson, K. (in press). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. Rothsetin, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments*. New York: Wiley.

Duval, S. (in press). The "Trim and Fill" method. In H. Rothsetin, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments*. New York: Wiley.

Gottfredson, L. S. (1996). Racially gerrymandering the content of police tests to satisfy the U.S. Justice Department: A case study. *Psychology, Public Policy, and Law, 2*, 418–446.

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

Halpern, S. D., & Berlin, J. A. (in press). Beyond Conventional Publication Bias: Other Determinants of Data Suppression. In H. Rothsetin, A. J. Sutton, & M. Borenstein (Eds). *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments*. Wiley.

Hedges, L., & Vevea, J. (in press). The selection model approach to publication bias. In H. Rothsetin, A.J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments*. New York: Wiley.

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, 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. (in press). Situational judgment research: Informing the debate on practical intelligence theory. *Intelligence, 33,* 515–525.

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

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

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

Nguyen, N. T., McDaniel, M. A., & Whetzel, D. L. (2005, April). *Subgroup differences in situational judgment test performance: A meta-analysis.* Paper presented at the 20th annual conference of the society, for Industrial and Organizational Psychology, Los Angeles.

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

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 Organizational Psychology, Toronto, Canada.

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.

Ployhart, R. E., & Ryan, A. M. (2000, April). *Integrating personality tests with situational judgment tests for the prediction of customer service performance.* Paper presented at the 15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Porr, W. B., & Ployhart, R. E. (2004, April). *The validity of empirically and construct-oriented situational judgment tests.* Paper presented at the 19th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Robie, C., Born, M. P., & Schmit, M. J. (2001). Personal and situational determinants of personality responses: A partial reanalysis and reinterpretation of the Schmit et al. (1995) data. *Journal of Business & Psychology, 16,* 101–117.

Robie, C., Schmit, M. J., Ryan, A. M., & Zickar, M. J. (2000). Effects of item context specificity on the measurement equivalence of a personality inventory. *Organizational Research Methods,* 34, 348–365.

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.

Sacco, J. M., Scheu, C. R., Ryan, A. M., Schmitt, N., Schmidt, D. B., & Rogg, K. L. (2000, April). *Reading level and verbal test scores as predictors of subgroup differences and validities of situational judgment tests.* Paper presented at the15th Annual Conference of the Society of Industrial and Organizational Psychology, New Orleans, LA.

Schmidt, F. L., & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge. *Current Directions in Psychological Science,* 2, 8–9.

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.

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology,* 80, 607–620.

Smith, K. C., & McDaniel, M. A. (1998, April). *Criterion and construct validity evidence for a situational judgment measure*. Paper presented at the 13th annual convention 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., Snook, S. A., & Grigorenko, E. L. (2000). *Practical intelligence in everyday life*. New York: Cambridge University Press.

Sterne, J. A. C. & Egger, M. (in press). Regression methods to detect publication and other bias in meta-analysis. In H. Rothsetin, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments*. New York: Wiley.

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.

Sutton, A. J., & Pigott, T. D. (in press). Bias in meta-analysis induced by incompletely reported studies. In H. Rothsetin, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments*. New York: Wiley.

Taylor, H. R. (1949). Test of practical judgment. In O. K. Buros (Ed.), *The third mental measurements yearbook* (pp. 694–695). New Brunswick, NJ: Rutgers University Press.

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

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.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372–376.