Human Resource Management Review 19 (2009) 188–202

Contents lists available at ScienceDirect

Human Resource Management Review

journal homepage: www.elsevier.com/locate/humres

Situational judgment tests: An overview of current research

Deborah L. Whetzel [a,*], Michael A. McDaniel [b]

[a] Human Resources Research Organization (HumRRO), United States
[b] Virginia Commonwealth University, United States

ARTICLE INFO

Keywords:
Situational Judgment Test
Validity
Reliability
Subgroup differences
Faking
Applicant reaction
Coaching

ABSTRACT

Situational judgment tests (SJTs) are popular personnel selection tests. To aid researchers, the paper summarizes the current knowledge and where knowledge gaps exist. To guide practice, the paper provides evidence-based recommendations. The paper begins with a brief history of SJTs, presents likely reasons for the resurgence of SJT research and practice, and summarizes the theoretical basis of SJTs. Then, the distinction between personnel selection methods and constructs is reviewed as it is particularly important in understanding SJTs. SJT research relevant to reliability and validity is summarized as is research relevant to the implementation of SJTs. The paper concludes with recommendations for practice and an agenda for future research.

1. Introduction

Situational judgment tests (SJTs) are designed to assess an applicant's judgment regarding a situation encountered in the work place (Weekley & Ployhart, 2006). SJT items present respondents with work-related situations and a list of plausible courses of action. Respondents are asked to evaluate each course of action for either the likelihood that they would perform the action or the effectiveness of the action. An illustrative SJT item is presented here:

You are facing a project deadline and are concerned that you may not complete the project by the time it is due. It is very important to your supervisor that you complete the project by the deadline. It is not possible to get anyone to help you with the work.

A. Ask for an extension of the deadline.
B. Let your supervisor know that you may not meet the deadline.
C. Work as many hours as it takes to get the job done by the deadline.
D. Explore different ways to do the work so it can be completed by the deadline.
E. On the day it is due, hand in what you have done so far.
F. Do the most critical parts of the project by the deadline and complete the remaining parts after the deadline.
G. Tell your supervisor that the deadline is unreasonable.
H. Give your supervisor an update and express your concern about your ability to complete the project by the deadline.
I. Quit your job.

This paper is designed to provide the reader with an overview of current and needed research in SJTs. For researchers, the paper summarizes the current knowledge and where knowledge gaps exist. For practitioners, the paper provides evidence-based recommendations for practice. We begin the paper with a brief history of SJTs and discuss reasons for the resurgence of SJT research and practice. This is followed by a discussion of the theoretical basis of SJTs. Next, we orient the reader to the SJT literature in two ways. First, we draw distinctions between personnel selection methods and constructs. This distinction is particularly important in understanding SJTs. We also orient the reader by identifying two major limitations of the knowledge base for SJTs. These limitations have implications for the strength of conclusions that can be drawn from the SJT literature. We offer these

[*] Corresponding author.
E-mail address: DWhetzel@humrro.org (D.L. Whetzel).

1053-4822/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.hrmr.2009.03.007

limitations early in the paper and revisit them when drawing conclusions about SJTs. We then offer a summary of SJT research knowledge grouping it into two broad categories. The first category concerns research relevant to reliability and validity. The second category concerns research relevant to the implementation of SJTs. These research topics include sub-group differences, the use of video SJTs, scoring methods, applicant reaction, retest effects, coaching, and faking. We conclude the paper with recommendations for practice and an agenda for future research.

1.1. History of SJTs

Several reviews have been conducted describing the history of SJTs (McDaniel & Whetzel, 2007; Weekley & Ployhart, 2006). As noted by Weekley and Ployhart (2006), the earliest example of SJTs depends on how they are defined. For instance, a U.S. civil service examination implemented in 1873 for the Examiner of Trade-Marks, Patent Office contained the following: “A banking company asks protection for a certain device, as a trade-mark, which they propose to put upon their notes. What action would you take on the application?” (DuBois, 1970, p. 148). Further, the 1905 Binet scale that measured intelligence in children asked questions such as, “When a person has offended you, and comes to offer his apologies, what should you do?” Although scenarios were presented, these early measures did not provide respondents with alternative options for handling the situations.

As noted by McDaniel, Morgeson, Finnegan, Campion and Braverman (2001), the first widely used SJT containing response options was likely a subtest of the George Washington Social Intelligence Test. The subtest 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. An early review of empirical studies criticized the test, claiming that correlations between the test and other measures of presumed social attributes were very low (Thorndike & Stein, 1937).

Army psychologists attempted to assess the judgment of soldiers during World War II (Northrop, 1989). These tests were comprised of scenarios and alternative responses to each scenario. Solutions were based on the person's ability to use common sense, experience, and general knowledge. Starting in the 1940s, a number of SJTs were developed to measure supervisory potential, such as the *Practical Judgment Test* (Cardall, 1942), and the *Supervisory Practices Test* (Bruce & Learner, 1958). In the late 1950's and early 1960's SJTs 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 the potential to be successful in management (Campbell, Dunnette, Lawler, & Weick, 1970).

More 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 developed 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, Dunnette and Carter (1990) renewed interest in SJTs when they examined “low-fidelity simulations” 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 typical SJTs in that the TKIM scenarios are considerably more lengthy and detailed. Wagner and Sternberg (1991) reported the conduct of five studies examining the criterion-related validity of the TKIM in academic and business settings, although no validity data were presented. Sternberg et al. (2000) also reported that these measures were unrelated to measures of general cognitive ability. We note that their samples included Yale undergraduate students, who are likely to have substantial range restriction on measures of general cognitive ability, thus reducing observed relationships on the restricted predictor.

Research on SJTs indicates that they are effective and are frequently used selection tools both in the U.S. and Europe (McDaniel et al., 2001; Salgado, Viswesvaran, & Ones, 2001). We attribute the resurgence of SJT research and practice to several factors. First, the Motowidlo et al. (1990) article was the first article concerning SJTs in a major personnel selection journal and generated substantial interest. Second, meta-analytic summarizes of research have documented that SJTs have useful levels of validity as predictors of job performance (McDaniel et al., 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007). Third, researchers and practitioners have long sought valid measures with lower sub-group differences than general cognitive ability. Research has demonstrated that SJTs have less race-based adverse impact than cognitive measures (Chan & Schmitt, 1997; Motowidlo & Tippins, 1993; Motowidlo et al., 1990; Whetzel, McDaniel, & Nguyen, 2008; Weekley & Jones, 1997, 1999). Fourth, SJTs have face and content validity because they describe work-related situations (Motowidlo, Hanson, & Crafts, 1997; Salgado et al., 2001). This makes SJTs appealing to staffing decision makers and applicants alike.

1.2. Theory relevant to SJTs

Personnel selection research emerged from an applied need to screen job applicants. Therefore an expansive theoretical basis for the research is often lacking. SJT research is no exception to this paucity of theory development in personnel selection but there are two notable areas of theory, both associated with Motowidlo and his colleagues (Motowidlo et al., 1990; Motowidlo, Hooper, & Jackson, 2006). In his 1990 article, Motowidlo noted that SJTs emanate from the tenet of behavioral consistency (i.e., that past behavior is the best predictor of future behavior). That is, by eliciting a sample of current behavior, one can predict how someone will behave in the future (Wernimont & Campbell, 1968). SJT items are samples of behavior in that the respondent is presented with a job situation and asked to evaluate various behavioral responses. Motowidlo et al. (1990) called SJTs low-fidelity simulations because these tests are intended to simulate job situations. The second major contribution to SJT theory is the concept of implicit trait policy (Motowidlo et al., 2006). Implicit trait policies are inherent beliefs about causal relationships between personality traits and behavioral effectiveness. Motowidlo et al. (2006) argued that individual differences in personality traits affect judgment of the effectiveness of behavioral episodes that express those personality traits. For example, if actions in SJT response options that express high agreeableness are truly more effective than actions that express low agreeableness, more agreeable people will weigh those response options more heavily than those low in agreeableness. We address this theory in more detail in our discussion of construct validity.

1.3. Method vs. construct

The distinction between methods and constructs has been an area of confusion in personnel selection for decades (Arthur & Villado, 2008; Hunter & Hunter, 1984). Constructs refer to the psychological or behavioral domain being sampled (e.g., general mental ability, conscientiousness, and psychomotor ability). Methods refer to the specific process or technique by which constructs are measured (e.g., interviews, paper-and-pencil tests, simulation modes of assessment) (Arthur & Villado, 2008). Some personnel selection measures can be clearly defined with respect to constructs. For example, cognitive ability tests measure the construct of general cognitive ability and conscientiousness tests measure the construct of conscientiousness. However, other common selection measures are best classified as measurement methods that simultaneously measure multiple constructs. Such selection measures include biodata measures, assessment centers, and employment interviews. For example, an employment interview targeted at measuring conscientiousness would also likely assess oral communication skills. This lack of precision regarding assessed constructs poses two problems. First, it is difficult to explain why a measure predicts job performance. Second, it may be difficult to build new measures that have the same characteristics as existing measures.

The distinction between methods and constructs is critical to understanding SJT research. Some researchers argue that SJTs are measures of a single construct. For example, Sternberg et al. (2000) argued that SJTs, which he had called practical intelligence tests, measured a single construct distinct from general cognitive ability. However, these assertions were shown to be unsubstantiated (Gottfredson, 2003; McDaniel & Whetzel, 2005) because SJTs measure multiple constructs and have some relationship with general cognitive ability. Many multiple-construct measures, such as Big 5 personality inventories, can be readily factored into construct relevant factors. For example, a factor analysis of a Big 5 measure will typically yield five factors corresponding to the constructs of conscientiousness, agreeableness, emotional stability, extroversion, and neuroticism. SJTs seldom yield interpretable factors (McDaniel & Whetzel, 2005). We concur with Schmitt and Chan (2006) who "propose[d] that SJTs, like the interview, be construed as a method of testing that can be used to measure different constructs but that the method places constraints on the range of constructs measured" (p. 149). Extending the argument, Arthur and Villado (2008) have suggested that other scenario-based tests, such as situational interviews, are all SJTs that only differ in their method, such as video-based, oral, or paper-and-pencil administration.

1.4. Limitations of the SJT knowledge base

Reviews of a literature, such as this article, are particularly valuable when they can draw well-supported evidence-based conclusions. The current paper draws such conclusions but, as with any review, the reader should be well aware of any limitations that may temper the confidence one may place in the conclusions. As with most personnel selection validity literatures, most SJT validity studies rely on concurrent designs. In such designs, respondents are incumbents who typically have little motivation to distort their responses. These incumbents are commonly told that their test results are part of a research project to evaluate an employment test and that their test results will have not be used to make decisions about their careers. However, operationally, tests are given to job applicants who, on average, may be motivated to distort their responses (i.e., fake to look good) because the test scores are used in determining whether they get hired. Thus, because SJT research primarily uses concurrent studies, it is possible that some of the conclusions drawn in this review may not hold for SJTs used to screen job applicants. Whether the results actually differ in applicant samples is an empirical question to be answered by future research.

Our conclusion that SJTs are methods also places constraints on the interpretation of validity and other research evidence concerning SJTs. Much of the research presented in this review is based on meta-analyses of results from primary studies. These meta-analytic results are best viewed as typical of most SJTs. However, because SJTs are measurement methods and different SJTs may measure different constructs, the meta-analytic results may not accurately describe the validity of a specific SJT. This is a concern particularly if the SJT taps different constructs than the constructs common to the SJT studies summarized in the meta-analytic studies or if the SJT is using some novel method of construction or presentation. In summary, while the meta-analytic conclusions are likely correct on average, some SJTs may display different construct or criterion-related validity than would be expected based on the cumulative literature.

2. Reliability and validity

2.1. Reliability

Estimating the reliability of SJTs is problematic for several reasons. First, SJTs typically assess multiple constructs and are often construct heterogeneous at the item level (McDaniel & Whetzel, 2005). In the sample item presented earlier, response option H is an example of an item that is likely to be construct heterogeneous. The item states: "Give your supervisor an update and express your concern about your ability to complete the project by the deadline." A respondent might believe that this is an effective action because the respondent is intelligent and/or because the respondent is conscientious. This could cause the item to be correlated with both general cognitive ability and conscientiousness. Such items often do not have unidimensional loadings when examined with factor analysis. When items lack clear factor loadings, it makes the creation of homogenous scales difficult. The scale and item heterogeneity makes Cronbach's alpha an inappropriate reliability index (Cronbach, 1949a, 1951). Test–retest reliability is a more appropriate reliability estimate for SJTs but it is rarely reported. Parallel form reliability also is rare because it requires the use of different item content to measure the same constructs. Because it is difficult to identify particular constructs assessed using SJTs, construct equivalence across forms can be problematic. Due to these test development and data collection problems, many researchers continue to provide internal consistency estimates with or without acknowledging that they underestimate the reliability (Chan & Schmitt, 1997; Pulakos & Schmitt, 1996; Pulakos, Schmitt, & Chan, 1996) of SJTs. One notable exception is Chan and Schmitt (2002), who estimated parallel form reliability at .76.

Two studies have sought to identify methods for constructing alternative forms of SJTs (Lievens & Sackett, 2007b; Oswald, Friede, Schmitt, Kim, & Ramsay, 2005). Lievens and Sackett (2007) examined three approaches for developing SJT items using item generation theory (Irvine & Kyllonen, 2002). The Random Assignment Strategy occurs when a large enough pool of SJT items is developed for a particular domain and they are randomly assigned to alternative forms. The Incident Isomorphism Strategy occurs when pairs of items are developed from the same critical incident (e.g., a physician dealing with a patient who refuses medication) and one of each pair is assigned to a form. The Item Isomorphism strategy is used when items reflect the same domain, the same incident, and the same context of the item stem and responses. Thus, the only differences across pairs of items are wording and grammar. These strategies form a continuum of item similarity across forms with random assignment being at the low end and item isomorphism being at the high end.

The effects of these approaches on alternate forms reliability were studied in a high-stakes context (i.e., admission to medical college). Because only students who failed the test the first time participated in the retest, they corrected the observed correlations for indirect range restriction (i.e., candidates were selected on the basis of a third variable). For the domain of interpersonal/communication skills, corrected correlations were .34 ($N=703$) for the random assignment strategy, .56 ($N=1385$) for the incident isomorphic strategy, and .68 ($N=1273$) for the item isomorphic strategy. These reliability estimates seem low compared to those typically achieved with measures of cognitive ability in which the expected reliability estimates are .80 and above. However, because several assumptions of test–retest reliability were violated due to the high-stakes nature of the test (e.g., the fact that only those who failed the first time took the second test), these numbers are not considered test–retest reliability. Thus, as benchmark, they computed similar coefficients for general mental ability (GMA) (only computing reliability on those who took the test the second time after failing it the first time). None of the coefficients for GMA fell above .70. Only SJTs developed using the item isomorphic approach yielded consistency values (.68) that were similar to those of the GMA (.67).

Oswald et al. (2005) took a different approach. They combined items selected randomly from each of 12 content domains (similar to the Random Assignment strategy described by Lievens and Sackett, 2007) to create large numbers of parallel forms. Parallel forms had to pass the following criteria: the means had to be similar within $|d|\leq.05$, alpha reliabilities were to be at or above .70, and criterion-related validity with GPA was to be at or above .15. Further, they trimmed the outlying 20% of standard deviation values (10% of each tail) out of the distribution. Of the 10,000 alternative forms tested, 144 remained. This is an empirical approach assuming a large number of items are available.

In sum, test–retest and parallel forms reliability are the most appropriate reliability estimation methods for SJTs. The reviewed literature on developing parallel forms indicates that it is a challenging task. Unless researchers have evidence that their SJT is homogeneous, they should stop estimating SJT reliability with coefficient alpha.

2.2. Validity

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) and the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, Inc., 2003) offer professional guidance on the concept of validity. The *Standards* define validity as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (American Educational Research Association et al., 1999, p. 184). The *Principles* adopt the same definition. The *Standards* and the *Principles* consider validity as a unitary concept and varying sources of evidence can contribute to an understanding of the inferences that can be drawn from a test score. There are two primary types of evidence related to the validity of SJT scores. The first is evidence related to the constructs measured by SJTs. The second is evidence concerning the prediction of job performance. These two types of evidence are presented below.

2.2.1. Validity evidence related to constructs

Three meta-analyses (McDaniel et al., 2001; McDaniel et al., 2007; McDaniel & Nguyen, 2001) summarized validity evidence relevant to the constructs assessed by SJTs. McDaniel et al. (2001) reported that the SJTs reviewed had average observed correlations of .31 with general cognitive ability indicating that SJTs typically assess cognitive ability to some degree but there is substantial variance in SJTs that is not associated with general cognitive ability. McDaniel and Nguyen (2001) identified two categories of SJT response instructions that moderate the validity of SJTs: knowledge and behavioral tendency. Knowledge response instructions ask respondents to select the correct or best possible response or to rate the effectiveness of responses.

Behavioral tendency response instructions ask the respondent to select the response that represents what the respondent would likely do or to rate the likelihood that they would perform an action. McDaniel and Nguyen (2001) hypothesized that response instructions may influence construct and criterion-related validity.

McDaniel et al. (2007) followed up on this line of research and provided the most comprehensive of these reviews. They compared the two types of response instructions and correlated scores using each type with scores on cognitive ability and personality. They found that SJTs 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 SJT response instructions, as shown in Table 1. SJTs with behavioral tendency instructions tend to be more correlated with personality than SJTs with knowledge instructions. However, SJTs with knowledge instructions correlate more highly with cognitive ability than SJTs with behavioral tendency instructions.

These results suggest that one can change the construct validity of a SJT by altering its response instructions. However, it is possible that the construct differences attributed to response instructions might be due to differences in the content of the SJTs. That is, SJTs with knowledge instructions may tend to have one type of item content while SJTs with behavioral tendency instructions may tend to have a different type of content. To address this issue, they conducted a second meta-analysis ($k=8$) of studies that held the SJT content constant but varied the response instructions (i.e., the same SJT items were administered twice, once with knowledge instructions and once with behavioral tendency instructions). The analyses for cognitive ability correlates showed that SJTs with knowledge instructions had substantially larger correlations with cognitive ability than the same SJTs administered with behavioral tendency instructions (.28 vs. .17). SJTs with behavioral tendency instructions had larger correlations with the Big 5 than the same SJTs administered with knowledge instructions for Agreeableness (.20 vs. .14), for Conscientiousness (.33 vs. .21), for Emotional Stability (.13 vs. .02), for Extraversion (.07 vs. .02), and for Openness to Experience (.10 vs. .05). Thus, they found that when administering the same SJT with varying response instructions, one can change the construct being measured.

Thus SJTs can measure both cognitive ability and personality, but the emphasis differs. Test developers who wish to emphasize the assessment of personality constructs in an SJT may wish to use behavioral tendency instructions. However, one should note that behavioral tendency instructions are susceptible to faking (Nguyen, Biderman, & McDaniel, 2005). On the other hand, if one were interested in maximizing cognitive ability variance within a SJT, a test with knowledge instructions may be more appropriate. The caution here is that a cognitively loaded SJT 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.

An important caveat to our conclusion concerning response instructions and constructs assessed is the issue of study design. When discussing the limitations of the SJT knowledge base, we noted that almost all SJT studies in the literature are based on concurrent designs. Applicants are more likely than incumbents to distort their responses to make a favorable impression. Additional research using applicants is warranted to determine if the conclusions concerning response instructions and constructs measured is replicable in applicant samples.

In addition to the meta-analytic findings concerning constructs assessed with SJTs, a number of studies have addressed efforts to develop SJTs to measure targeted constructs. Ascalon (2005) attempted to measure cross-culture social intelligence. The underlying

Table 1
SJT construct correlations.

Instruction type	N	k	Mean r	ρ
Cognitive ability	30,859	95	.29	.32
Knowledge instructions	24.656	69	.32	.35
Behavioral tendency instructions	6.203	26	.17	.19
Agreeableness	25,473	51	.22	.25
Knowledge instructions	17,115	34	.17	.19
Behavioral tendency instructions	8,358	17	.33	.37
Conscientiousness	31,277	53	.23	.27
Knowledge instructions	23,043	38	.21	.24
Behavioral tendency instructions	8,234	15	.30	.34
Emotional stability	19,325	49	.19	.22
Knowledge instructions	11,067	33	.10	.12
Behavioral tendency instructions	8,258	16	.31	.35
Extroversion	11,351	25	.13	.14
Knowledge instructions	9,533	14	.14	.15
Behavioral tendency instructions	1,818	11	.07	.08
Openness to experience	4,515	19	.11	.13
Knowledge instructions	2,921	11	.12	.14
Behavioral tendency instructions	1,594	8	.09	.11

Adapted from McDaniel et al. (2007).
Note. N is the number of subjects across all studies in the analysis; k is the number of studies; Mean r is the observed mean. ρ is the estimated mean population correlation which were corrected for measurement error in the personality and cognitive ability measures. The first row in each analysis is the correlation between the situational judgment test and the Big 5 measure for both kinds of instructions.

dimensions of the measure were empathy and ethnocentrism. The findings supported the classification of empathetic-non-ethnocentric style as theoretically the best strategy, and the non-empathetic-ethnocentric style as the worst strategy. There was partial support for construct validity, however, due to the small sample ($N=74$), the results were likely affected by sampling error.

Becker (2005) developed a SJT designed to measure employee integrity. Using non-transparent workplace dilemmas, he assessed whether SJT scores would predict integrity-relevant outcomes. For the combined sample of fast food service employees ($N=86$), engineers ($N=59$), and production workers ($N=128$), integrity scores were correlated with managerial ratings of career potential, leadership activities, and job performance. Integrity was not related to the quality of interpersonal relationships.

Teamwork is another construct that researchers have attempted to measure using SJTs. Mumford, Morgeson, Van Iddekinge, and Campion (2008) developed a SJT called the Team Role Test to measure knowledge of ten roles relevant to the team context (e.g., coordinator, critic, contributor, and completer). In a sample of academic project teams ($N=93$), team role knowledge predicted team member role performance ($r=.34$) and had incremental validity beyond general mental ability and the Big 5. As shown in a second study, the predictive validity of role knowledge generalized to team members in a work setting ($N=82$; $r=.30$).

Teamwork knowledge, using an SJT, also was researched in a manufacturing organization (a Midwest mill of a national steel corporation; Morgeson, Reider, & Campion, 2005) in which teams were used to perform production-related tasks, with most decisions made at the team level. They found that when teamwork knowledge (measured with a SJT) was entered into the equation last, after social skills (measured with a structured interview) and personality, teamwork knowledge accounted for an additional 8% of the contextual performance criterion variance ($N=90$). The SJT was most highly correlated with conscientiousness ($r=.23$) and the remainder of the personality correlates were .03 or less.

In an effort to understand the relationships among ability, experience, and personality with SJTs, Weekley and Ployhart (2005) used path analysis to identify alternative models showing antecedents and relationships with performance. In their study with 271 employees in two levels of loss prevention management jobs within a large mass merchandizing retail organization, they found that SJTs scores mediated the effects of ability and experience on performance and partially mediated the effects of personality on performance. Their results suggested that the SJT assesses general forms of knowledge (rather than job-specific knowledge). Importantly, the SJT showed incremental validity above cognitive ability, personality, and experience measures.

The ability to judge effective actions in various situations also can provide boundary conditions around certain personality constructs. Chan (2006) investigated the interactive effects of SJT effectiveness and proactive personality (i.e., defined by Crant [2000] as taking initiative to improve current circumstances or to create new ones) on work perceptions and work outcomes. Specifically, using a sample of 139 employees at a large rehabilitation agency, proactive personality positively predicted the work outcomes (job satisfaction, organizational commitment and job performance) for individuals with high situational judgment effectiveness, but negatively predicted the work outcomes for those with low levels of situational judgment effectiveness. Adding the interaction term increased the proportion of criterion variance accounted for by 10% for job satisfaction and 6% for job performance. One can imagine times when it is inappropriate to be proactive and those without the ability to judge such times may be deemed to have lower levels of job performance than those who can identify when it is appropriate to be proactive.

Although not explicitly a construct validity study, Motowidlo et al. (2006) studied people's implicit trait policies (ITP) and how they affect SJT scores in an examination of ITP. Their ITP hypothesis was that personality traits have causal effects on implicit beliefs about the importance of those traits for behavioral effectiveness. In their sample of 96 students, correlations between ITPs and their associated personality traits support the hypothesis for extraversion ($r=.39$) and agreeableness ($r=.30$), but much less so for conscientiousness ($r=.15$). In a second study, using an SJT prepared for operational use in a government agency, they attempted to replicate these results for agreeableness and conscientiousness. For a sample of 100 students, agreeableness results were statistically significant and conscientiousness results were not. Motowidlo et al. (2006) argued that when SJTs are scored by comparing applicants' responses to those of experts about the effectiveness of response options, they measure procedural knowledge. However, SJTs can be built to measure other constructs. Specifically, if response options are developed to measure high or low levels of targeted attributes, and if respondents rate each response option for its effectiveness, and if SJTs are scored by computing estimates of the magnitude of the trait for which it was developed, then SJTs can measure implicit trait policies for targeted traits.

2.2.2. Validity evidence concerning prediction of job performance

Having reviewed the first primary type of validity evidence (constructs), we now review the second primary type of validity evidence (prediction of criteria). The criterion-related validity of SJTs 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 SJTs (McDaniel et al., 2001, 2007). In the second and more comprehensive meta-analysis, the overall validity of SJTs across 118 coefficients was .26 ($N=24{,}756$), regardless of instruction type. These validity results are almost entirely based on concurrent validity studies (e.g., research typically conducted using job incumbents, rather than applicants, as subjects). The authors suggested that 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. We join those authors in calling for such research.

2.3. Incremental validity

Two meta-analyses (McDaniel et al., 2001, 2007) and several primary studies (Chan & Schmitt, 2002; Clevenger & Haaland, 2000; O'Connell, McDaniel, Grubb, Hartman, & Lawrence, 2002; Weekley & Jones, 1997, 1999) examined the incremental validity of SJTs over measures of cognitive ability.

McDaniel et al. (2007) estimated the incremental validity of SJTs over cognitive ability, the Big 5, and a composite of cognitive ability and the Big 5. They also examined the extent to which cognitive ability and the Big 5 add incremental validity to SJTs. To build correlation matrices needed for these analyses, they used other data to estimate the criterion-related validity of cognitive ability and the Big 5, and the intercorrelations among all measures. They conducted the incremental validity analyses by running hierarchical linear regressions using a correlation matrix of all variables. All incremental validity analyses were based on observed correlations that were not corrected for measurement error or range restriction. They showed the correlation between cognitive ability (*g*) and the SJT used in each of three analyses (SJTs using knowledge instructions [.32], SJTs using knowledge instructions without an outlier [.37], and SJTs using behavioral tendency instructions [.17]). They also show the criterion-related validity values used for *g*, the SJT, and the Big 5 (.25, .20 and .16 respectively).

Next, they examined the validity of various composites of *g*, SJT, and the Big 5. These validities are multiple *R*s from regressions. The incremental validity of the SJT over cognitive ability (*g*) is .03. That number was calculated by subtracting the validity of *g* alone ($r = .25$) from the multiple *R* where cognitive ability and SJT were optimally weighted in a regression to predict job performance ($R = .28$). Thus, by adding a SJT to a predictor battery already containing *g*, the SJT incremented the uncorrected validity by .03.

In all three response instruction scenarios, SJTs provided incremental validity over cognitive ability ranging from .03 to .05. The largest incremental validity is for SJTs with behavioral tendency instructions (.05 vs. .03). Such SJTs have the lowest correlations with *g* and thus have a higher probability of predicting over and above *g*. Although this finding is consistent with the idea that knowledge instructions are more *g* saturated, this difference may not be practically meaningful.

In all three response instruction scenarios, SJTs provide incremental validity over a composite of the Big 5, ranging from .06 to .07. Because SJTs with behavioral tendency instructions have more personality saturation than SJTs with knowledge instructions, it is reasonable that SJTs with behavioral tendency instructions offer lower incremental validity (.06) over the Big 5 than knowledge instruction SJTs (.07). Although the direction of the moderating effect is consistent with expectations, the magnitude of the moderating effect is very small.

In all three scenarios, SJTs offer incremental validity over a composite of *g* and the Big 5 with incremental values ranging from .01 to .02. The response instruction moderator does not appear to meaningfully moderate the incremental validity. McDaniel et al. noted that although these observed incremental values are small, few predictors offer incremental prediction over an optimally weighted composite of six variables (i.e., *g* and the Big 5). As with evidence related to construct correlations and zero-order correlations with job performance, we remind the reader that most of the studies related to validity are based on concurrent samples. We encourage that our conclusions be re-evaluated as more applicant sample data become available.

3. Review of research regarding implementation of SJTs

We now turn to issues surrounding the implementation of SJTs. These issues include subgroup differences, the use of video-based SJTs, scoring methods, applicant reaction, retest effects, coaching, and faking.

3.1. Subgroup differences

Subgroup difference are a key concern when implementing any selection system. Whetzel et al. (2008) provided a systematic review of mean race and sex differences in SJT performance. The mean differences were expressed as standardized mean differences (*d*). A *d* of one indicates that one group is one standard deviation above the mean of another. Their meta-analysis showed that, on average, White test takers performed better on SJTs than Black ($d = .38$), Hispanic ($d = .24$), and Asian ($d = .29$) test takers. Female examinees performed slightly better than male test takers ($d = -.11$). They investigated two moderators of these differences: 1) loading of *g* or personality on the SJT, and 2) response instructions. The loading of *g* on the SJT, also called the *cognitive loading*, was operationalized as the correlation between a measure of cognitive ability and the SJT. The personality loading of the SJT was similarly operationalized as the correlation between a personality scale and a SJT. As shown in Table 2, mean race differences between Black, Hispanic, Asian and White examinees on SJTs were explained largely by the cognitive loading of the SJT such that the larger the cognitive load, the larger the mean race differences. Regarding the effect of personality loadings of SJTs on race differences, Black-White and Asian-White differences were smaller to the extent that SJTs were correlated with emotional stability and Hispanic-White differences were smaller to the extent that SJTs were correlated with conscientiousness and agreeableness.

Table 2
Vector correlations between ethnic and gender differences and constructs correlated with situational judgment tests.

	Cognitive ability	Conscientiousness	Agreeableness	Emotional stability
Black/White comparison	.77 (30)	−.01 (30)	.14 (29)	−.20 (23)
Hispanic/White comparison	.49 (21)	−.45 (25)	−.24 (24)	.07 (21)
Asian/White comparison	.40 (14)	−.11 (19)	.14 (18)	−.37 (17)
Male/Female comparison	.08 (33)	−.37 (35)	−.49 (35)	.06 (29)

Adapted from Whetzel et al. (2008).
Note. Numbers in parentheses are the numbers of coefficients contributing data to the vector correlations.

For gender differences, the vector correlation between the cognitive loading of the SJTs and the mean sex differences is .08, indicating that cognitive loading has minimal impact on sex differences in SJTs. For both conscientiousness and agreeableness, the higher the loading of the construct, the higher the scores achieved by women (−.37 for conscientiousness and −.49 for agreeableness). These results show that sex differences can be explained, in part, by the loading of personality in the SJT. That is, the more a SJT is positively correlated with conscientiousness and agreeableness, the greater the sex differences favoring women.

In sum, SJTs are popular selection methods because of their validity. Although they have some adverse impact on minority groups, the magnitude is less than that typically found with cognitive ability tests. Further, they have some, albeit minimal, incremental validity over and above other commonly used selection devices, including cognitive ability and personality.

3.2. Video-based SJTs

A number of studies have been conducted comparing video-based vs. written SJTs. Chan and Schmitt (1997) conducted a laboratory experiment comparing both media and found that a video-based SJT had significantly less adverse impact than a written SJT and students perceived the video-based SJT to have more face validity than the written SJT. Richman-Hirsch, Olson-Buchanan, and Drasgow (2000) found similar results in that students reacted more favorably to a multimedia format of a conflict resolution skills SJT than to a written format of the same test. There has been some question about the extent to which contextualization provided by SJTs is helpful to examinees by providing cues unavailable in written form or hinders their performance by providing too much information. Olson-Buchanan and Drasgow (2006) stated:

> [In video-based SJTs assessees] see and hear people interacting, perceive, or more importantly, fail to perceive their emotions and stress, and confront dilemmas about one's choice of action or inaction …. With this format, we may be able to better understand how the assessee will interpret verbal and nonverbal behaviors of others in the workplace and choose to respond. (p. 253)

On the other hand, it has been mentioned that video-based SJTs might insert irrelevant contextual information and unintentionally bring more error into SJTs (Weekley & Jones, 1997).

More recently, Lievens, Buyse and Sackett (2005b) examined the incremental validity of a video-based SJT over cognitive ability for making college admission decisions ($N = 7197$). They found that when the criterion included both cognitive and interpersonal domains, the video-based SJT showed incremental validity over cognitively-oriented measures for curricula that included interpersonal courses, but not for other curricula. This study demonstrates the importance of differentiating not only predictor constructs, but criterion domains.

Lievens and Sackett (2006) also studied the predictive validity of video-based and written SJTs of the same content (interpersonal and communication skills) in a high-stakes testing environment ($N = 1159$ took the video-based SJT; $N = 1750$ took the written SJT). They found that the video-based SJT had a lower correlation ($r = .11$) with cognitive ability than the written version ($r = .18$). For predicting interpersonally-oriented criteria, the video-based SJT had higher validity ($r = .34$) than the written version ($r = .08$).

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

3.3. Scoring of SJTs

Similar to biodata items, SJT items are frequently written so that there is no definitive “correct” answer, as defined by a body of knowledge or expertise. Consequently, there are several ways to score SJTs, each of which may lead to different estimates of validity (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). In their analysis of various scoring approaches, Bergman et al. (2006) reviewed empirical scoring (in which items or options are scored according to their relationships with a criterion measure) (Hogan, 1994), theoretical scoring (in which items are constructed to reflect theory, or theory can be used to identify the best and worst options), and expert-based scoring (in which SMEs make judgments about the items). They investigated each scoring method, as well as several combinations of methods, using a 21-item Leadership Skills Assessment SJT to determine the incremental validity of the SJT over cognitive ability (the Wonderlic Personnel Test) and personality (Sixteen Personality Factor Questionnaire) ($N = 123$). They found that cognitive ability accounted for 10% of the variance in leadership ratings and that personality factors were not significantly related to the leadership criterion. Incremental validity of each of the keys was entered as a third step in the hierarchical regressions. Their results showed that the empirical key accounted for an additional 2.3% of the criterion variance, the hybrid initiating structure key accounted for an additional 3.0%; the hybrid participation key accounted for an additional 1.7% of the variance; and the SME approach accounted for an additional 4.9% of the variance in leadership ratings. None of the keys showed subgroup differences by sex and all of the keys showed discriminant validity. The authors concluded that the validity of an SJT depends on part on its scoring and that poor choices could lead to the conclusion that SJTs are not valid when it may only be that the scoring key is not valid.

Legree, Psotka, Tremble, and Bourne (2005) reviewed a scoring approach, known as Consensus Based Measurement (CBM), and compared it with expert-based scoring to determine the effect of both methods on validity. Their CBM procedure allowed examinee responses to be scored as deviations from the consensus defined by the response distributions of the examinee sample. Their reasoning is that, "…knowledge domains may exist that are lodged in opinion and have no objective standard for verification other than societal views, opinions, and interpretations." (p. 103).

They cited several studies comparing CBM and expert judgment. In their earlier work with SJTs (Legree, 1995; Legree & Grafton, 1995), they compared the opinions of a small number of experts and the mean ratings across examinees ($N=198$ U.S. Air Force recruits) using an SJT that consisted of 49 scenarios and listed a total of 202 alternatives in which the examinees rated the effectiveness of each alternative. They found an observed correlation of .72 ($r=.95$ corrected for attenuation of the reliability of each set of observations) between the two scoring methods.

In another study comparing expert and examinees ratings, they used the Tacit Knowledge for Military Leadership (TKML) scale (Hedlund et al., 2003). The two comparison groups were 50 Lieutenant Colonels (with an average of 18 years' experience) and 355 West Point cadets (with no military experience). The two sets of scoring standards correlated .96. While certainly there is some similarity of backgrounds (e.g., military training regarding interpersonal events and problems, and issues of authority and obedience), one might expect there to be some differences in responses between the two groups. Thus, they found that use of the cadet group's average as the standard was indistinguishable from an expert-based score.

Krokos, Meade, Cantwell, Pond, and Wilson (2004) studied the use of empirical scoring. Their reasoning was that SJTs are typically designed to capture complex, social or practical aspects of performance in work situations which are indistinguishable from tests of tacit knowledge (McDaniel & Whetzel, 2005). When SJTs are based on a scoring key developed based on the consensus judgment of SMEs, the likelihood that the correct answers will be the most transparent options is increased. Thus, more transparent items are more likely to be retained when SMEs are asked to determine the correct answer. Contrary to Legree et al. (2005), they contend that if respondents' answers are used to create the scoring key, the "SMEs" consist of both high and low applicants/performers. Thus, low validity coefficients for SME-scored SJTs could be due to differences in perceptions of the construct among respondents (e.g., job applicants).

Krokos et al. (2004) noted that, in contrast, when empirical keying is used, the "SMEs" are the high performing respondents as measured by the criterion of interest. Using empirical keying, the most transparent option (the one that seems the best) may be endorsed by a majority of respondents, but if both high and low performing respondents equally endorse the response option, it will not differentiate between criterion groups and therefore, will not be weighted. On the other hand, a response option that is endorsed less frequently, but only by the high performing respondents, will be weighted more heavily with empirical scoring.

To evaluate their ideas, Krokos et al. (2004) developed an SJT that consisted of three detailed scenarios, each of which had five response options. Examinees were to select the option they would most likely do and the option they would least likely do. Responses were correlated with ratings made by a director. Their results showed that although the empirical keying approaches had large significant correlations with performance in the calibration sample, only one method was significantly related to performance in the cross-validation sample. We note that the sample size for the cross-validation sample was 75. Such a low number of examinees is indicative of greater sampling error and great likelihood of Type II error. Thus, it may be premature to reject their empirical keying methodology. Further, the low number of SJT items may also have affected results such that the instrument may have had low reliability.

In sum, as with response instructions leading to the measurement of different constructs (e.g., cognitive ability vs. personality) (McDaniel et al., 2007), different scoring keys can lead to an SJT assessing different constructs. Currently, there is insufficient research to judge one scoring strategy to be substantially better than other. More research is clearly needed in this area.

3.4. Applicant reaction

When implementing a selection system, applicant reactions can affect a number of variables. One of the most consistent findings in applicant reaction research is that the outcome applicants receive (e.g., pass or fail) or their perceptions of how well they did on a selection hurdle affects applicant perceptions (Ryan & Ployhart, 2000). For example, Gilliland (1994) found that the actual hiring outcome for applicants was related to perceptions of process and outcome fairness. Bauer, Maertz, Dolen and Campion (1998) found that applicants' passing or failing accounted for significant variance in outcomes such as the perceived fairness of testing and organizational attractiveness.

Truxillo, Seitz, and Bauer (2008) studied the role of cognitive ability in assessing one's own performance in a SJT. They reasoned that because the process of judging one's own performance is essentially a cognitive task, those higher in cognitive ability should be better able to assess their own performance on a selection test than those with lesser cognitive ability. Their study included 108 undergraduate students who took the Wonderlic and a video-based SJT. They also responded to questions about their perceived test performance as well as a three-item questionnaire on self-efficacy (Bauer et al., 1998). Their results showed that there was a strong relationship between actual and perceived test performance for those high in cognitive ability, but there was no relationship for those low in cognitive ability. Participants lower in cognitive ability were less able to assess their performance accurately, which suggests that cognitive ability may be related to certain meta-cognitive skills (Kraiger, Ford, & Salas, 1993). These results are consistent with other research showing that cognitive ability is related to knowledge acquisition (Hunter, 1983; Ree, Carretta, & Teachout, 1995) and accuracy in estimating others' performance (Hauenstein & Alexander, 1991).

Kanning, Grewe, Hollenberg, and Hadouch (2006) investigated applicant reactions to video vs. written SJTs. Specifically, they compared the stimulus component (video vs. text), the response component (video vs. text), and the interactivity of the test. The interactive SJTs took into account an applicant's previous answers when presenting the next item. The applicant chose one of two behavioral alternatives and that choice dictated which item was administered next. Thus, for each "stem," there were two items, the second of which was dictated by the response to the first item. Thus, they studied applicant reaction to two variables related to fidelity: 1) video vs. written, and 2) interactivity of the SJT. In their first study ($N=284$ police officers), they found that interactive items received higher ratings for acceptance, but not for the other perception variables (e.g., transparency and job-relatedness), however this study did not include comparisons with text-based stimulus. In the second study ($N=82$ police officers), the research design included all combinations of stimulus and response (video and text based) and interactivity. Their analyses showed the superiority of interactive; stimulus and response (video) compared with the other item types. Compared to other item types, this type received more positive values for usefulness, acceptance, and job relatedness. For ratings of perceived fairness, applicant reactions were not significantly different from SJTs that were not interactive. In general, items that were interactive and used video for both the stimulus and responses received higher positive emotional reactions and better acceptance than the other types of items. Of course, one would need to consider the cost of developing such items. They would be much more expensive to develop than a pencil and paper SJT, but an interactive SJT would be difficult to administer, unless the test was administered on computer with the use of branching programs.

In sum, there are at least two variables that have received recent attention in the research regarding applicant reactions to SJTs. First, self-efficacy research shows that those with higher cognitive ability are more able to judge their own test performance than those with lower cognitive ability. Second, higher levels of fidelity (e.g., video based SJTs and the use of interactive items) resulted in higher levels of user acceptance than SJTs using lower levels of fidelity.

3.5. Retest effects

Retesting occurs when examinees re-take an examination, presumably after not achieving a passing score the first time they took it. There appears to be at least two broad rationales behind retesting (Lievens, Buyse, & Sackett, 2005a). First, retesting may be warranted due to the possibility that the initial assessment was an error, either due to a temporary characteristic of the examinee (e.g., illness), a temporary characteristic of the testing situation (e.g., deviations from the standardized test administration procedures, such as poor lighting), or to random measurement error. Retesting effects can be defined as test score differences after prior exposure to an identical test or to an alternate form of the test under standardized conditions.

Lievens et al. (2005a) examined retest effects for three types of tests (knowledge, cognitive ability, and situational judgment) in a high-stakes testing environment (admission to medical school) using within-person and between-person designs. Within-person retest effects refer to effects associated with the same group of individuals who retake an identical test or an alternate form of a test. Between-person retest effects refer to the comparison between first-time test takers as a group and repeat test takers as a group.

The SJT consisted of short video-taped scenarios of interpersonal situations that physicians are likely to encounter with patients (e.g., handling patient complaints or conveying bad news). After viewing each scenario, the students were to choose the most effective response provided in text format.

Analyses of within-person retest effects showed that the mean scores of repeat test takers were 1/3 of a standard deviation higher for their second administration of the SJT and 1/2 of a standard deviation higher for their second administration of the cognitive ability test. Using the between-persons design, there was no mean difference in SJT validity between first-time test takers and repeat test takers ($r=.20$ corrected, $N=348$ and $r=.28$ corrected, $N=140$, respectively). Unlike SJTs, the validity of the cognitive ability test was significantly higher for the first-time test takers ($r=.28$ corrected, $N=1237$) than for the repeat test takers ($r=-.03$, $N=556$). The authors present possible reasons for these differences. For SJTs, in which the validity is about equal for first-time test takers and repeaters, yet the scores are different for repeaters, this combination of evidence seems to suggest that there is genuine improvement in the construct of interest and/or criterion-relevant change (e.g., reduced unfamiliarity with the second administration). This explanation may not be true for all test takers, but they may apply to the repeat test-takers in general.

3.6. Coaching

An issue related to retesting is the notion of coaching. Often, in between test administrations, examinees take a coaching course in order to improve their scores. Most research has found that coaching leads to either no improvement or slight improvement in the predictive validity of cognitive ability tests (Allalouf & Ben-Shakhar, 1998).

Cullen, Sackett and Lievens (2006) examined the coachability of two SJTs, the College Student Questionnaire (CSQ; Sternberg and the Rainbow Project Collaborators, 2002) and the Situational Judgment Inventory (SJI; Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004), both of which were developed for consideration as selection instruments in the college admission process. They developed a "coaching course" for each test by identifying test taking strategies, much as a test preparation organization would. Using pre and posttests and a control group that did not receive the coaching course, the compared the test scores of 395 students. They also obtained measures of cognitive ability (ACT scores) and grade point averages of the student participants.

For the CSQ, their results showed that participants in the training condition obtained scores that were higher by .24 standard deviations than those in the control group, whereas participants in the SJI training condition obtained scores that were actually lower by .22 standard deviations, however the validity of these tests was unaffected. In an attempt to understand the score differences, they also administered surveys asking participants about the level of effort required to understand the test taking strategies. Responses showed that the SJI strategies were more difficult to learn. One reason participants may have had to exert greater effort to learn the strategies is that they were more situation specific than the CSQ strategies. The CSQ strategies were fairly straightforward (i.e., be organized, take responsibility, and confront problems). In contrast, the SJI strategies were not as simplistic and there were qualifications to many of the strategies. This finding was substantiated by their analysis in which they regressed posttest scores in both training conditions on pretest scores and then on ACT scores. Their results showed that cognitive ability was an important determinant of the ability to absorb and successfully use the SJI strategies. Thus, given certain rules, the CSQ was more coachable than the SJI. Thus, SJTs, as typically constructed, using SME judgment, or empirical keying, rather than a set of well-defined rules, is less coachable than a test in which the strategies are easily defined.

As a side note, the finding that the CSQ was correlated with cognitive ability ($r=.32$) was contrary to claims by Sternberg and the Rainbow Project Collaborators (2002) that the CSQ and other tacit knowledge measures are not cognitively loaded, but are measure of "practical know-how." This is consistent with the findings of McDaniel and Whetzel (2005) in which they pointed to the multi-faceted nature of SJTs as evidence that SJTs capture some cognitive variance.

Another interesting finding was the scale effect for the CSQ. Cullen et al. (2006) simulated what would happen if the training course included an instruction not to endorse extreme values by changing the 1 and 2 responses to a 3 and by changing the 6 and 7 responses to a 5. They discovered that scores could be improved by 1.57 standard deviations if examinees did not endorse extreme answers (e.g., 1 or 2 and 6 or 7).

One of the most important practical implications of this study is that the validity of both the CSQ and the SJI in predicting freshman GPA was unaffected by the coaching interventions. However, use of certain strategies would change the rank order of candidates using the CSQ and that would have important implications in a high-stakes testing situation.

3.7. Faking

Faking on a selection measure can be defined as an individual's conscious distortion of responses to score favorably (e.g., Dwight, 1999; McFarland & Ryan, 2000). Faking has been studied extensively in non-cognitive measures, such as personality tests, biodata inventories, and integrity tests (Alliger & Dwight, 2000; Douglas, McDaniel, & Snell, 1996; Graham, McDaniel, Douglas, & Snell, 2002; Kluger & Collela, 1993; McFarland, Ryan, & Aleksander, 2002; Ones & Viswesvaran, 1998). In fact, there has been an ongoing debate in the literature about the extent to which applicants actually engage in faking in employment settings. Some studies have reported that applicants do not fake personality tests, and even if they do, it does not negatively affect their validity (Abrahams, Neumann, & Githens, 1971; Ellingson, Smith, & Sackett, 2001; Hough, 1998; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990, McCrae & Costa, 1983, Ones & Viswesvaran, 1998). Other studies have found that faking occurs in selection settings and that it attenuates the criterion-related validity of personality tests (Douglas et al., 1996; Dunnette, McCartney, Carlson, & Kirchner, 1962; Kluger, Reilly, & Russell, 1991; Schmit & Ryan, 1992; Zickar, 1997). Regarding practical outcomes, researchers have found that faking can have a significant affect on who is hired (Rosse, Stechner, Levin, & Miller, 1998; Weiner & Gibson, 2000).

Nguyen et al. (2005) suggested that the response instructions provided to examinees affect the extent to which SJTs are fakable. In their study, 203 student participants indicated both the best and worst response (i.e., knowledge) and the most likely and least likely response (i.e., behavioral tendency) to each situation. They also varied whether people were asked to fake good first or respond honestly first. Using a within subjects design, they found that the faking effect size for the SJT behavioral tendency response format was .34 when participants responded first under honest instructions and .15 when they responded first under faking instructions. The knowledge response format results were inconsistent, probably because it is difficult to "fake" knowledge (i.e., either one knows the answer or one does not). They also found that honest condition knowledge SJT scores were more highly correlated with cognitive ability ($r=.56$) than were behavioral tendency SJT scores ($r=.38$).

Peeters and Lievens (2005) studied the fakability of a SJT using college students. Their SJT consisted of 23 items related to student issues (e.g., teamwork studying for exams, organizing, accomplishing assignments, interpersonal skills, social responsibility, perseverance, and integrity) and students were asked how they would respond (a behavioral tendency instruction in the McDaniel et al., 2007 taxonomy). Their results showed that students in the fake condition ($N=153$) had significantly higher SJT scores than students in the honest condition ($N=138$). To assess whether the faking effect was practically significant, they computed the effect size which was about one standard deviation ($d=.89$) with women ($d=.94$) being better able to fake than men ($d=.76$). They also identified how many "fakers" were in the highest quartile to simulate the effect of a selection ratio of .25. They found that 76% of fakers and 24% of honest respondents were in the highest quartile. The lowest quartile consisted of 69% honest respondents and 31% fakers. This shows that faking on an SJT has substantial effects on who is selected.

Regarding the effect of faking on criterion-related validity, they found that the correlation between the SJT and GPA was significantly larger for the honest group ($r=.33$) than the faking group ($r=.09$), which indicated that faking had a negative effect on the criterion-related validity of the SJT. They also assessed the incremental validity of the SJT in honest and faking conditions over cognitive ability and personality. Their results showed that in the honest condition, the SJT explained an additional 5.1% of the variance in GPA whereas in the fake condition, the SJT did not add incrementally to the prediction of GPA over cognitive ability and personality.

The implications of this study are that when students fake, and they probably do in a selection context, the SJT with behavioral tendency instructions has limited validity because the correlation coefficient dropped to .09. Both the criterion-related validity and incremental validity of the SJT were affected by faking such that the use of SJTs with behavioral tendency instructions in a high-stakes testing situation may be problematic. One possible remedy for faking is to use knowledge instructions (Nguyen et al., 2005), rather than behavioral tendency questions.

Kanning and Kuhne (2006) studied the effect of social desirability on SJT items. They included an 11-item SJT using behavioral tendency instructions (what would you do) and a measure of social desirability in a test battery designed to select police officers. They used three groups consisting of real applicants, students instructed to present a positive impression, and students instructed

Table 3
Number of students who passed and did not pass in fake good and honest conditions.

	Fake good	Honest
Passed	46	51
Not passed	9	4

Adapted from Kanning & Kuhne (2006).

to respond honestly ($N=55$ for each group). They found that the means were nearly equal in the two student samples (honest and faking). Further, the number of students who "passed/did not pass" the battery was not much different between the honest vs. faking condition, as shown in Table 3. One plausible explanation is that to the extent that the scenarios were based on police situations, the students may not have known how to fake. In addition, there were no significant correlations between the SJT and social desirability scales.

In sum, these studies show mixed results regarding the fakability of SJTs. However, two general themes can be identified. First, people can fake. Consistently, people instructed to fake can do so and to the extent that this changes the rank order of candidates in a high-stakes selection situation, this ability has serious implications for operational use of tests. Second, one may be able to reduce faking by using knowledge-based instructions (e.g., what is the correct thing to do?, how effective is the behavior?). These kinds of instructions ask for the correct answer and thus make the items maximal performance measures (Cronbach, 1949b, 1984; McDaniel et al., 2007). As with all maximal performance tests, SJTs with knowledge instructions assess how one respondent performs when doing their best. Thus, both honest and faking respondents have the same motivation to respond and faking should not be possible.

3.8. Topics for future research and recommendations for practice

SJTs are a popular topic of research. In this section, we offer suggestions for the most pressing research needs in this area. Ten suggestions are presented. We then offer recommendations for practice.

We suggest that most pressing need in future SJT research is to determine the extent to which conclusions, largely based on concurrent samples, will generalize to applicant samples. Applicants complete SJTs under high stakes situations that likely have impact on their motivation. For example, applicants will likely fake more than incumbents. Such faking may make distinctions between behavioral tendency and applicant response instructions less important. For example, some applicants may ignore behavioral tendency instructions to report how they typically behave and instead report what they would believe would be the most desirable behavior. This would cause these applicants to resemble those applicants completing the SJT with a knowledge instruction (i.e., identify the best response to the situation).

Second, research concerning differences between responses instructions in SJTs seldom control for SJT content. Very few studies (see McDaniel et al., 2007) have held SJT content constant while manipulating response instructions. Thus, future studies should examine response instruction differences while controlling for the content of the SJT.

Third, we have noted that SJTs are measurement methods and the constructs assessed by SJTs vary across tests. Thus, meta-analytic conclusions about the criterion and construct validity of SJTs may be right, on average, but may be incorrect for a specific SJT. It might be useful to identify the characteristics of SJTs that do not conform to the conclusions drawn from the meta-analytic literature to see if any consistent pattern emerges. If such a pattern emerges, it can inform science on SJT characteristics (yet to be discovered) that influence validity.

Fourth, most knowledge about SJTs are based on convenience samples and seldom do SJT researchers consider the limitations of their samples on their conclusions about validity. Many selection instruments (e.g., personality tests, cognitive ability tests) yield varying validities depending on the jobs examined or the demands (e.g., cognitive) of the jobs. Future research should focus on potential job content moderators of the validity of SJTs.

Fifth, as noted by Ployhart and Weekley (2006), our knowledge of construct validity is primarily limited to cognitive ability and the Big 5. SJT responses can reasonably be expected to be a function of generic and domain-specific job knowledge gained through experience or formal education.

Sixth, the small amount of evidence collected to date suggests that some types of SJTs (i.e., those with knowledge instructions) may be resistant to faking. Our field is sorely in need of ways of assessing personality and other non-cognitive constructs in a faking-resistant manner. Thus, we encourage additional research on faking and SJTs.

The seventh research need is to study the effects of various scoring strategies. For example, when using Likert rating scales with SJTs, we suspect that most use raw scores while some (Legree et al., 2005) suggest that the scores be subject to a within-person standardization to remove individual differences in the use of the Likert scales related to scatter (i.e., rating variance) and elevation (i.e., some respondents use one end of the scale more than others) (Cronbach & Gleser, 1953). Does one approach yield scores that predict better than others?

Our eighth suggestion for future was based on remarks by a reviewer. In some ways, SJT scenarios are similar to problem analysis exercises such as those used as stand alone work samples (Callinan & Robertson, 2000) or as part of an assessment center. Often when scoring a problem analysis exercise, the evaluator is interested not only in the offered solution but in the process of analysis that lead to the solution. Currently, SJTs are only scored based on judgments of the solutions and are not based on any evaluation of the judgment process underlying the solution. Research into the factors considered by a respondent in evaluating an item may prove useful in understanding the constructs assessed by the item and may lead to novel ways of scoring SJT items.

Our ninth suggestion is that more research is needed to determine the extent to which knowledge instructions are faking resistant. The Nguyen et al. (2005) paper provided preliminary empirical evidence and there are rational arguments indicating that it is difficult to fake knowledge. However, more research is clearly needed. Given that faking is a concern of many in the assessment of self-reported personality, SJTs with knowledge instructions may prove to be a viable way of assessing personality in a faking resistant manner.

Last, but certainly not least, we need new insights into understanding the constructs assessed by SJTs. Although it is useful to know that SJTs tap both general cognitive ability and personality, there is substantial variance in SJTs that remains unexplained. Until we can better understand the constructs assessed, our knowledge of how to improve SJTs through assessing more predictive constructs will largely be guess work.

Our recommendations for practice are both research and experience driven. We have a strong preference for the use of knowledge instructions (e.g., how effective is this response?) primarily because the evidence to date, although limited, suggests that SJTs with knowledge instructions are faking resistant. We also have a preference for asking the respondent to rate each response option on a Likert scale of effectiveness. Thus, if the test has 10 scenarios each with five response options, one has 50 potentially scorable items when the respondent rates each response. If the respondent is asked for the best response, one would only have 10 potentially scorable responses. In the absence of large samples suitable for empirical keying, we suggest using a group of subject matter experts to take the test individually and then discuss discrepancies in their ratings and when possible arrive at an answer key through consensus. If one has sufficient scenarios and response options, we usually drop response options where the subject matter experts substantially disagree. An alternative approach is to identify the source of the disagreement and edit the response option such that a clear consensus on the effectiveness of the response option is reached. Although we recommend the use of multi-point Likert scales because respondents often wish to express some nuance of degree regarding their judgment of effectiveness, we often recode the responses into dichotomies such that the responses are identified as either effective or ineffective. A Likert scale with an even number of ratings points (a four-point scale or a six-point scale) facilitates the dichotomization (when dichotomizing an item, it is difficult to know whether to assign the middle point of an odd number Likert scale to effective or ineffective). We prefer to treat the items as dichotomous for three reasons. First, we would not wish to explain the distinction between *very effective* and *extremely effective* in an adversarial situation such as might be encountered in employment litigation. Second, there are likely to be individual differences in how respondents use Likert scales. For example, some may prefer using more extreme scale anchors (1 and 6 on a six point Likert scale), while others prefer using more moderate scale anchors (2 and 5 on a six point Likert scale). Dichotomizing the items removes these individual differences in use of rating scales and it is our unevaluated assumption that these rating scale preferences are not criterion-relevant and are best removed. Third, dichotomizing the items should lessen the effectiveness of the coaching advice (Cullen et al., 2006) concerning avoiding the use of extreme responses. We encourage research in evaluating our recommendations.

In summary, there is a substantial amount of SJT research. However, there is also a pressing need for additional research in several areas. This review has helped identify the state of the research and the gaps in the knowledge base.

Acknowledgements

The authors thank James Breaugh and George Thornton for their helpful feedback on the manuscript and Kathryn McDaniel for her assistance in Spanish language translation.

References

Abrahams, N. M., Neumann, I., & Githens, W. H. (1971). Faking vocational interests: Simulated versus real life motivation. *Personnel Psychology*, *24*, 5—12.

Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement*, *35*, 31—47.

Alliger, G. M., & Dwight, S. A. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. *Educational and Psychological Measurement*, *60*, 59—72.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Arthur, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*, 435—442.

Ascalon, E. M. (2005). Improving expatriate selection: Development of a situational judgment test to measure cross-cultural social intelligence. *Dissertation Abstracts International*, *65*, 4880.

Bauer, T. N., Maertz, C. P., Dolen, M. R., & Campion, M. A. (1998). A longitudinal assessment of applicant reactions to an employment test. *Journal of Applied Psychology*, *83*, 892—903.

Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment*, *13*, 225—232.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14*, 223—235.

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

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection & Assessment*, *4*, 248—260.

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. (2006). Interactive effects of situational judgment effectiveness and proactive personality on work perceptions and work outcomes. *Journal of Applied Psychology*, *91*, 275—281.

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

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

Clevenger, J. P. & Haaland, D. E. (2000). *Examining the relationship between job knowledge and situational judgment performance*. Paper presented at the15th Annual Conference of the Society of Industrial and Organizational Psychology. New Orleans. April.

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.* Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center (PRR-80-30).

Crant, J. M. (2000). Proactive behavior in organizations. *Journal of Management*, *26*, 435—462.

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

Cronbach, L. J. (1949b). *Essentials of psychological testing.* New York: Harper & Row.

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

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

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarities between profiles. *Psychological Bulletin*, *50*, 456—473.

Cullen, M. J., Sackett, P. R., & Lievens, F. P. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment*, *14*, 142—155.

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

Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996, August). *The validity of non-cognitive measures decays when applicants fake.* Paper presented at the annual conference of the Academy of Management, Cincinnati, OH.

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

Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology*, *15*, 13—24.

Dwight, S.A. (1999). An assessment of the different effects of warning applicants not to fake. Unpublished doctoral dissertation, State University of New York at Albany.

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology*, *86*, 122—133.

Gilliland, S. W. (1994). Effects of procedural and distributive justice on reactions to a selection system. *Journal of Applied Psychology*, *79*, 691—701.

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

Graham, K. E., McDaniel, M. A., Douglas, E. F., & Snell, A. F. (2002). Biodata validity decay and score inflation with faking: Do item attributes explain variance across items? *Journal of Business and Psychology*, *16*, 573—592.

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 4th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, Georgia.

Hauenstein, N. M. A., & Alexander, R. A. (1991). Rating ability and performance judgments: The joint influence of implicit theories and intelligence. *Organizational Behavior and Human Decision Processes*, *50*, 300—323.

Hedlund, J., Forsythe, G. B., Horvath, J. A., Williams, W. M., Snook, S., & Sternberg, R. J. (2003). Identifying and assessing tacit knowledge: Understanding the practical intelligence of military leaders. *Leadership Quarterly*, *14*, 117—140.

Hogan, J. B. (1994). Empirical keying of background data measures. In G. S. Stokes & M. D. Mumford (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 69—107). Palo Alto, CA: CPP Books.

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance*, *11*, 209—244.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, *75*, 581—595.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, and job performance, and supervisory ratings. In F. Landry, S. Bedeck & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 257—266). Hillsdale, NJ: Erlbaum.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72—98.

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation and test development* Mahwah, NJ: Erlbaum.

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subject's point of view: Reactions to different types of situational judgment items. *European Journal of Psychological Assessment*, *22*, 168—176.

Kanning, U. P., & Kuhne, S. (2006). Social desirability in a multimodal personnel selection test battery. *European Journal of Work and Organizational Psychology*, *15*, 241—261.

Kluger, A. N., & Collela, A. (1993). Beyond the mean bias: The effect of warning against faking on biodata item variances. *Personnel Psychology*, *46*, 763—780.

Kluger, A. N., Reilly, R. R., & Russell, C. J. (1991). Faking biodata tests: Are option-keyed instruments more resistant? *Journal of Applied Psychology*, *76*, 889—896.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, *78*, 311—328.

Krokos, K., Meade, A.W., Cantwell, A.R., Pond, S.B., Wilson, M.A. (2004). *Empirical keying of situational judgment tests*. Paper presented at the 16th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor. *Intelligence*, *21*, 247—266.

Legree, P. J., & Grafton, F. C. (1995). *Evidence for an interpersonal knowledge factor: The reliability and factor structure of tests of interpersonal knowledge and general cognitive ability.*Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences Technical Report No. 1030).

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *Emotional Intelligence: An International Handbook* Cambridge, MA: Hogrefe & Huber.

Lievens, F., Buyse, T., & Sackett, P. R. (2005a). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, *58*, 981—1007.

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

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, *91*, 1181—1188.

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

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology*, *51*, 882—888.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology*, *60*, 63—91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730—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. (2007). Situational judgment tests. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied Measurement: Industrial Psychology in Human Resources Management* (pp. 235—257). Mahwah, NJ: Lawrence Erlbaum and Associates.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across non-cognitive measures. *Journal of Applied Psychology*, *85*, 812—821.

McFarland, L. A., Ryan, A. M., & Aleksander, E. (2002). Item placement on a personality measure: Effects on faking behavior and test measurement properties. *Journal of Personality Assessment*, *78*, 348—369.

Morgeson, F. P., Reider, M. H., & Campion, M. A. (2005). Selection individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. *Personnel Psychology*, *58*, 583—611.

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: Davis Black.

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

Motowidlo, S. J., & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology*, *66*, 337−344.

Mumford, T. V., Morgeson, F. P., Van Iddekinge, C. H., & Campion, M. A. (2008). The team role test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology*, *93*, 250−267.

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

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 and Organizational Psychology, Toronto, Canada.

Olson-Buchanan, J. B., & Drasgow, F. (2006). Multimedia situational judgment tests: The medium creates the message. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 253−278). San Francisco: Jossey-Bass.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, *11*, 245−269.

Oswald, F. L., Friede, A. J., Schmitt, N., Kim, B. K., & Ramsay, L. J. (2005). Extending a practical method for developing alternate test forms using independent sets of items. *Organizational Research Methods*, *8*, 149−164.

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

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement*, *65*, 70−89.

Ployhart, R. E., & Weekley, J. A. (2006). Situational judgment: Some suggestions for future science and practice. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* Mahwah, NJ: Lawrence Erlbaum Associates.

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., Carretta, T. R., & Teachout, M. S. (1995). Role of ability and prior knowledge in complex training performance. *Journal of Applied Psychology*, *80*, 721−730.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology*, *85*, 880−887.

Rosse, J. G., Stechner, M. D., Levin, R. A., & Miller, J. L. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, *83*, 634−644.

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management*, *26*, 565−606.

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

Schmit, M. J., & Ryan, A. M. (1992). Test-taking dispositions: A missing link? *Journal of Applied Psychology*, *77*, 629−637.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 135−156). Mahwah, NJ: Lawrence Erlbaum.

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, Inc., Dallas, TX.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures*, 4th Edition Bowling Green, OH: Author.

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. and the Rainbow Project Collaborators (2002). Enhancing the SAT through assessments of analytical, practical, and creative skills. Unpublished manuscript.

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

Truxillo, D. M., Seitz, R., & Bauer, T. N. (2008). The role of cognitive ability in self-efficacy and self-assessed test performance. *Journal of Applied Social Psychology*, *38*, 903−918.

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: 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.

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.

Weiner, J.A., & Gibson, W.M. (2000, April). *Practical effects of faking on job applicant attitude test scores*. Paper presented at 15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

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

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*, 291−309.

Zickar, M.J. (1997, April). *Computer simulation of faking on a personality test*. Paper presented at the 12th annual conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.