PERSONNEL PSYCHOLOGY
2010, 63, 83–117

SITUATIONAL JUDGMENT TESTS: CONSTRUCTS ASSESSED AND A META-ANALYSIS OF THEIR CRITERION-RELATED VALIDITIES

MICHAEL S. CHRISTIAN
Eller College of Management University of Arizona

BRYAN D. EDWARDS
William S. Spears School of Business
Oklahoma State University

JILL C. BRADLEY
Craig School of Business
California State University, Fresno

Situational judgment tests (SJTs) are a measurement method that may be designed to assess a variety of constructs. Nevertheless, many studies fail to report the constructs measured by the situational judgment tests in the extant literature. Consequently, a construct-level focus in the situational judgment test literature is lacking, and researchers and practitioners know little about the specific constructs typically measured. Our objective was to extend the efforts of previous researchers (e.g., McDaniel, Hartman, Whetzel, & Grubb, 2007; McDaniel & Ngyuen, 2001; Schmitt & Chan, 2006) by highlighting the need for a construct focus in situational judgment test research. We identified and classified the construct domains assessed by situational judgment tests in the literature into a content-based typology. We then conducted a meta-analysis to determine the criterion-related validity of each construct domain and to test for moderators. We found that situational judgment tests most often assess leadership and interpersonal skills and those situational judgment tests measuring teamwork skills and leadership have relatively high validities for overall job performance. Although based on a small number of studies, we found evidence that (a) matching the predictor constructs with criterion facets improved criterion-related validity; and (b) video-based situational judgment tests tended to have stronger criterion-related

Authors' Note. This paper is based in part on the master's thesis of Michael S. Christian, which was chaired by Bryan D. Edwards. An earlier version of this paper was presented at the 2007 Annual Conference of the Society for Industrial and Organizational Psychology.

We are grateful to Winfred Arthur, Jr., Ronald Landis, Michael Burke, and Filip Lievens for reviewing previous drafts of this article and providing valuable suggestions. We also thank Michael McDaniel, Phillip Bobko, and Edgar Kausel for their helpful comments and suggestions on this project. Finally, we acknowledge the work of Jessica Siegel, Helen Terry, and Adela Garza.

Correspondence and requests for reprints should be addressed to Michael S. Christian Eller College of Management, University of Arizona, Department of Management and Organizations, McClelland Hall, PO Box 210108, Tucson, AZ 85721-0108; msc@email.arizona.edu.

validity than pencil-and-paper situational judgment tests, holding constructs constant. Implications for practice and research are discussed.

Situational judgment tests (SJTs) have a long history of use for employee selection (e.g., File, 1945; File & Remmers, 1971; Motowidlo, Dunnette, & Carter, 1990; Weekley & Jones, 1997, 1999). An SJT is a measurement method typically composed of job-related situations or scenarios that describe a dilemma or problem requiring the application of relevant knowledge, skills, abilities, or other characteristics (KSAOs) to solve. SJT items may be presented in written, verbal, video-based, or computer-based formats (e.g., Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001; Motowidlo et al., 1990), and usually contain options representing alternative courses of action from which the test taker chooses the most appropriate response.

Despite the widespread use of SJTs, even a cursory review of the literature reveals that test developers and researchers often give little attention to the constructs measured by SJTs and instead tend to report results based on overall (or composite) SJT scores. Nevertheless, because SJTs are measurement methods, understanding the constructs a given SJT measures is vitally important for interpreting its psychometric properties such as reliability, validity, and subgroup differences (e.g., Arthur & Villado, 2008; Lievens & Sackett, 2007; McDaniel, Morgeson, Finnegan, Campion, & Braverman 2001; Motowidlo et al., 1990; Schmitt & Chan, 2006; Smith & McDaniel, 1998). Hence, to understand *how* and *why* SJTs work in a selection context, there is a critical need for the identification of the constructs typically assessed using SJTs. This need has been highlighted recently by researchers who have called for an increased research focus on the constructs being measured by predictors of job performance (e.g., Arthur & Villado, 2008; Hough & Ones, 2001; McDaniel et al., 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007; Ployhart, 2006; Roth, Bobko, McFarland, & Buster, 2008). Indeed, there has been "virtually no direct investigation of the relationships linking SJT scores and test content" (Schmitt & Chan, 2006, p. 147). This is a critical oversight because test content is an important consideration in establishing construct validity (AERA, APA, & NCME, 1999; Binning & Barrett, 1989; Schmitt & Chan, 2006) and helps explain why constructs measured by SJTs are related to performance.

Therefore, as noted by others (e.g., Lievens, Buyse, & Sackett, 2005; Ployhart & Ryan, 2000a; Schmitt & Chan, 2006), we feel that SJT research could benefit from developing a more theoretically driven construct-level framework. Hence, the primary objectives of this research were to (a) discuss the advantages of attending to and reporting SJT construct-level versus method-level results; (b) develop a typology of constructs that

have been assessed by SJTs in the extant literature; and (c) undertake an initial examination of the criterion-related and incremental validity of the identified constructs and to investigate moderators of these validities.

Advantages of a Construct-Based Approach

The construct-based approach refers to the practice of evaluating and describing properties of SJTs in terms of the constructs measured (i.e., test content) rather than in terms of the method of measurement. The lack of attention to SJT constructs is arguably the result of the way in which SJTs are typically developed, applied, and reported in the literature (Arthur & Villado, 2008; Schmitt & Chan, 2006). Many selection tests are construct centered in that they are designed to measure a specific construct (e.g., cognitive ability, conscientiousness, integrity) and are therefore labeled based on the constructs that they measure (e.g., cognitive ability test). In contrast, predictors such as interviews, work samples, and SJTs are often described in method-based terms and are frequently developed using a job-centered approach in which the tests are designed to simulate aspects of the work itself rather than measure a specific predictor construct (Roth et al., 2008). In most selection contexts, simulation-based tests are developed to closely match job performance, which is reflected in the "sample" approach to measurement where tests are developed to have isomorphisim with the performance domain (Binning & Barrett, 1989; Wernimont & Campbell, 1968). Therefore, SJTs often are designed by collecting critical incidents of job performance in a particular setting and in doing so tap a number of predictor constructs simultaneously (e.g., Chan & Schmitt, 1997; McDaniel et al., 2001; McDaniel & Nguyen, 2001; Motowidlo & Tippins, 1993). Furthermore, many studies either fail to report the constructs measured by SJTs (e.g., Chan, 2002; Cucina, Vasilopoulos, & Leaman, 2003; Dicken & Black, 1965; McDaniel, Yost, Ludwick, Hense, & Hartman, 2004; Pereira & Harvey, 1999) or simply report composite method-level scores rather than scores for the specific constructs (e.g., Chan & Schmitt, 2002; Motowidlo et al., 1990; Smith & McDaniel, 1998; Swander, 2000; Weekley & Jones, 1997, 1999).

Reporting results at the construct level offers theoretical and practical advantages. First, from a theoretical perspective, the goal should not just be to show that a measure predicts job performance but also *why* that measure or construct predicts job performance (Arthur & Villado, 2008; Messick, 1995). Hence, identifying the constructs measured by selection tests such as SJTs is important for theory testing and understanding why a given test is or is not related to the criterion of interest. Second, a focus on reporting constructs also allows researchers to make more precise comparisons between various selection methods. The lack of attention

to constructs in the extant SJT literature leads to scores that are difficult to compare to scores derived from other selection methods or constructs (e.g., Arthur & Villado, 2008). For example, empirical investigations comparing the criterion-related validity, incremental validity (e.g., Clevenger et al., 2001), and subgroup differences[1] (e.g., Pulakos & Schmitt, 1996; Sackett, Schmitt, Ellingson, & Kabin, 2001) of SJTs to other predictor measures are difficult to interpret without specifying the construct(s) measured (Arthur, Day, McNelly, & Edens, 2003; Arthur & Villado, 2008). Likewise, comparisons of predictive validity between different SJT formats (e.g., pencil and paper, video-based) are more meaningful when constructs are held constant, as we detail later.

Third, specification of the KSAOs measured by SJTs helps to reduce contamination in test scores resulting from the measurement of unintended, non-job-relevant constructs. Fourth, in terms of job relevancy, a compelling argument must be made that the validity evidence available for the measure justifies its interpretation and use (Messick, 1995). Fifth, when practitioners and researchers are uncertain of the reason a test predicts a particular outcome, their ability to generalize findings across time and context is hindered. The construct-based approach allows for the development of SJTs that can be used to predict performance across many different jobs. In contrast, it would be difficult to transport SJTs across contexts if validity data are reported only at the composite or method level. Finally, by identifying the constructs measured by SJTs, practitioners can enhance predictive validity by theoretically matching predictor and criterion constructs (Bartram, 2005; Hogan & Holland, 2003; Mohammed, Mathieu, & Bartlett, 2002; Moon, 2001; Paunonen, Rothstein, & Jackson, 1999).

The Current State of Knowledge About the Constructs Typically Measured Using SJTs

In spite of (or perhaps *because* of) the lack of attention in many primary studies to constructs, some researchers have speculated about the constructs measured by SJTs. For instance, Sternberg and Wagner (1993) posited that SJTs measure tacit knowledge, whereas Schmidt and Hunter (1993) argued that they primarily measure job knowledge, and McDaniel and Nguyen (2001) suggested that some SJTs may predominantly

[1]A recent article by Whetzel, McDaniel, and Nguyen (2008) provides an interesting alternative to examining the degree to which subgroup differences are affected by variance attributed to cognitive ability. Using vector analysis, they show that as the correlation with cognitive ability of an SJT increases, standardized mean race differences on the SJT increase. However, this approach is still conducted at the method-level using composite scores for SJTs, whereas we focused more on a construct-based approach.

measure cognitive ability and personality. More recently, Schmitt and Chan (2006) argued that SJTs measure constructs like adaptability and contextual knowledge. Nevertheless, with the exception of the work by McDaniel and colleagues, we found little compelling empirical evidence for these suppositions. In a series of meta-analytic studies, McDaniel and colleagues assessed the construct saturation of SJTs and indicated that SJTs measure cognitive ability (M_ρ = .33–.46), Agreeableness (M_ρ = .27–.31), Conscientiousness (M_ρ = .25–.31), Emotional Stability (M_ρ = .26–.30), Extraversion (M_ρ = .30), and Openness (M_ρ = .13; McDaniel et al., 2001, 2007; McDaniel & Nguyen, 2001). These studies provide critical information by identifying a set of constructs typically associated with SJTs from which to build a more comprehensive typology.

This research extends McDaniel and colleagues' work by taking an alternative methodological approach. Specifically, we broadened the list of constructs generated by McDaniel and colleagues by using content analysis to investigate additional constructs such as leadership, social skills, and job knowledge. As noted by Schmitt and Chan (2006), this approach will help to reveal whether SJTs, as a measurement method, inherently tap certain constructs or whether the SJT content can be modified to assess these constructs to a greater or lesser extent. Further, this approach allows for holding predictor constructs constant, which facilitates comparisons of criterion-related validity between SJTs and other predictor methods. Hence, as we explain, our approach was to identify primary studies that *do* report construct or content information in order to develop a typology of the constructs typically assessed by SJTs.

Identifying the Constructs Assessed by SJTs

Our approach to understanding which constructs are typically measured by SJTs is consistent with research investigating other job-centered selection methods such as interviews (Huffcutt, Conway, Roth, & Stone, 2001), assessment centers (Arthur et al., 2003), and work samples (Roth et al., 2008), which often share the same construct/method confound as SJTs. We followed the suggested steps of Huffcutt et al. (2001), which involved SJT construct identification, classification, and frequency assessment in addition to collecting and reporting criterion-related validity information for the constructs typically assessed by SJTs.

We reviewed the selection literature for existing typologies and found the work of Huffcutt et al. (2001) to be the most suitable construct classification framework for SJTs. The construct categories in their typology include mental capability, knowledge and skills, basic personality tendencies, applied social skills, interests and preferences, organizational fit, and physical attributes. We chose this framework for several reasons.

First, Huffcutt et al.'s typology provided an adequate summary of the primary psychological characteristics that could be measured using SJTs. For example, McDaniel and colleagues have shown that SJTs measure cognitive ability and personality (McDaniel et al., 2001, 2007; McDaniel & Nguyen, 2001; Whetzel, et al., 2008), indicating that these constructs are often deliberately assessed using SJTs. Further, the SJT method enables the presentation of complicated social situations rich in contextual details. For this reason, we posited that SJTs would be frequently developed to measure applied social skills such as interpersonal skills, teamwork skills, and leadership. In addition, this typology has been used as a framework for classifying the constructs assessed by other job-centered, method-based predictors such as employment interviews (Huffcutt et al., 2001) and work sample tests (Roth et al., 2008).

Adding to our rationale for the typology, SJTs share basic similarities with interviews and work samples; in that they are methods that can be designed to tap a variety of job-relevant constructs. Further, these methods often measure constructs embedded within "clusters" of KSAOs designed to sample particular work-related characteristics (Huffcutt et al., 2001; Roth et al., 2008). Given that SJTs are typically composed of job-related scenarios designed to simulate the job, the scenarios may often measure multiple constructs because most job behaviors require multiple KSAOs. Therefore, we followed the lead of Roth et al. (2008) and analyzed the SJT content in terms of its saturation with predominant higher-order construct domains.[2]

The idea of construct saturation is useful because job-centered methods often do not "cleanly" assess one specific construct. Saturation refers to the extent to which a given construct influences (or saturates) complex measures like SJTs. Therefore, when the reported constructs for a given SJT were homogeneous relative to a particular construct domain, we considered the SJT "saturated" with this domain. For example, Weekley and Jones (1999) developed an SJT measuring coworker interaction skills, customer interaction skills, and loss-prevention behaviors in customer service, which are constructs related to interpersonal skills. Although the situational item content referenced incidental job-specific constructs, all items described interpersonal interactions; therefore this SJT was clearly saturated with the construct domain interpersonal skills.

Next, we conducted a comprehensive review of the SJT literature to identify the constructs and construct domains researchers reported measuring (see Table 1). We conceptualized construct *categories* as the highest

[2]Although Huffcutt et al. (2001) coded at the dimension level in their meta-analysis of interviews, we followed the lead of Roth et al. (2008), and coded studies at the test level. This was because virtually no studies of SJTs included scores at the dimension level.

TABLE 1
Typology of SJT Construct Domains

Construct category and domain	Construct	*k*	% of total
		136	100.00
Knowledge and skills			
Job knowledge and skills		4	2.94
	Knowledge of the interrelatedness of units	1	
	Pilot judgment (knowledge content)	1	
	Managing tasks	1	
	Team role knowledge	1	
Applied social skills			
Interpersonal skills		17	12.50
	Ability to size up personalities	1	
	Customer contact effectiveness	1	
	Customer service interactions	3	
	Guest relations	1	
	Interactions	2	
	Interpersonal skills	3	
	Negotiations	1	
	Service situations	1	
	Social intelligence	2	
	Working effectively with others	1	
	Not specified (interpersonal skills content)	1	
Teamwork skills		6	4.41
	Teamwork	3	
	Teamwork KSAs	3	
Leadership		51	37.50
	Administrative judgment	1	
	Conflict resolution for managers	2	
	Directing the activities of others	2	
	Handling people	3	
	General management performance	2	
	Handling employee problems	2	
	Leadership/supervision	4	
	Managerial/supervisory skill or judgment	5	
	Managerial situations	1	
	Supervisor actions dealing with people	2	
	Supervisor job knowledge	1	
	Supervisor problems	1	
	Supervisor Profile Questionnaire	1	
	Managing others	5	
	Not specified (leadership content)	19	

TABLE 1 (continued)

Construct category and domain	Construct	*k*	% of total
Basic personality tendencies		13	9.56
Personality composites	Conscientiousness, Agreeableness, Neuroticism	3	
	Adaptability, ownership, self-initiative, teamwork, integrity, work ethic	1	
Conscientiousness	Conscientiousness	7	
Agreeableness	Agreeableness	1	
Neuroticism	Neuroticism	1	
Heterogeneous composites		45	33.09

Note. Frequencies represent the number of independent effects.

order classification with construct *domains* falling under the categories, and *constructs* falling under construct domains. Following this conceptualization, we documented the extent to which researchers reported the constructs assessed and how commonly SJTs in the literature measured specific construct domains. Finally, we calculated initial meta-analytic estimates of the criterion-related validity of each construct domain[3] and examined the impact of both a construct-level moderator (i.e., job performance facets) and a method-level moderator (i.e., SJT format) on these validities.

Moderator Analyses and Hypotheses

Job performance facets. Our critique of research that focuses on composite predictor scores also applies to job performance criteria. The typical practice when conducting meta-analyses of predictors is to calculate validities based on ratings of job performance, collapsing across specific performance dimensions (e.g., Arthur et al., 2003; McDaniel et al. 2001, 2007; Roth et al., 2008). Although the practice of meta-analyzing criterion-related validity using a broad and inclusive criterion is useful in some respects (e.g., Viswesvaran, 2001), partitioning the criterion domain into specific aspects of job performance (e.g., facets) provides clarity of

[3]In order to estimate incremental validity using meta-analysis, one must obtain primary studies that report correlations between (a) more than one SJT construct domain and the criterion, and (b) the intercorrelations among the SJT construct domains. Unfortunately, almost no studies in the extant literature provided the information necessary to perform such analyses, so we were unable to accomplish this goal.

how and why predictor constructs relate to criteria (e.g., Bartram, 2005; Campbell, 1990; Hogan & Holland, 2003; Hurtz & Donovan, 2000; Motowidlo & Van Scotter, 1994; Rotundo & Sackett, 2002; Tett, Jackson, Rothstein, & Reddon, 1999). For example, Lievens and colleagues (2005) found that an SJT measuring interpersonal skills predicted an interpersonal performance criterion, whereas it had no relationship with an academic performance criterion. Indeed, evidence suggests that matching predictor constructs with theoretically appropriate criterion facets may result in stronger validities (Bartram, 2005; Hogan & Holland, 2003; Mohammed et al., 2002; Moon, 2001; Paunonen et al., 1999). Therefore, we examined specific performance facets as a moderator of the criterion-related validity of the constructs measured by the SJTs in our dataset.

In order to partition the job performance criterion, we reviewed a number of potential performance categorization models from the extant literature (see Viswesvaran & Ones, 2000 for a review), and from these we chose as a starting point the higher-order classification scheme suggested by Borman and Motowidlo (1993) of task versus contextual performance. Task performance is a measure of the degree to which individuals can perform the substantive tasks and duties that are central to the job. Contextual performance is not task specific and relates to an individuals' propensity to behave in ways that facilitate the social and psychological context of an organization (Borman & Motowidlo, 1993). Contextual performance consists of interpersonal and motivational components (e.g., Borman & Motowidlo, 1997; Chan & Schmitt 2002; Rotundo & Sackett, 2002). In addition, we included ratings of managerial performance as another job performance facet. Many SJTs are developed to predict managerial or supervisory performance, which arguably is distinct from task and contextual performance in that many elements of managerial performance involving behaviors typically considered to be contextual (e.g., interpersonal facilitation) are actually core management duties (e.g., Borman & Motowidlo, 1993; Conway, 1999; Scullen, Mount, & Judge, 2003; Witt & Ferris, 2003). Therefore, we defined managerial performance as behaviors central to the job of a manager or leader, including administrative and interpersonal behaviors. In sum, we divided the job performance criterion into three facets[4]: (a) task performance, (b) contextual performance, and (c) managerial performance.

Partitioning the performance criterion allowed a set of hypotheses to be developed based on prior research for expected magnitudes of the

[4]Although it could have been informative to provide information for studies that assessed supervisor ratings of overall global job performance (rather than assessing multiple dimensions and collapsing across them to create composites ratings), we did not find appropriate numbers of studies to facilitate such an analysis.

relationships between SJT predictor construct domains and specific criterion facets. As we detail next, we expected that the benefits of the construct-based approach to classifying SJTs would be realized in terms of stronger validities for construct domain-specific SJTs when appropriately matched to criterion facets, compared with weaker validities for heterogeneous composite SJTs correlated with the same facets.

Contextual performance includes a combination of social behaviors (e.g., teamwork, helping, cooperation, and conflict resolution) and motivational behaviors (e.g., effort, initiative, drive; Borman & Motowidlo, 1997; Van Scotter & Motowidlo, 1996). Applied social skills such as interpersonal skills, teamwork skills, and leadership skills should relate to contextual performance ratings to the degree that they reflect the ability to perceive and interpret social dynamics in such a way that facilitates judgments regarding the timing and appropriateness of contextual behaviors (Ferris, Witt, & Hochwarter, 2001; Morgeson, Reider, & Campion, 2005; Witt & Ferris, 2003). Interpersonal skills should predict contextual performance because interpersonally oriented individuals will be more likely to perform behaviors involving helping and social facilitation. Likewise, the social awareness associated with teamwork skills should translate into a propensity to perform behaviors that maintain the psychological and social context of organizational teams (Morgeson et al., 2005; Stevens & Campion 1999). Also, workers who are not managers but have strong leadership skills should have the ability and the motivation (Chan & Drasgow, 2001) to exhibit socially facilitative behaviors such as helping and motivating others. In addition, theoretical models and empirical evidence support a relationship between contextual performance and interpersonal skills (Ferris et al., 2001; Witt & Ferris, 2003), teamwork skills (Morgeson et al., 2005; Stevens & Campion, 1999), and leadership skills (Conway, 1999; Mumford, Zaccaro, Connelly, & Marks, 2000). Hence, we hypothesized:

Hypothesis 1: For contextual performance, SJTs measuring interpersonal skills will have stronger relationships than heterogeneous composite SJTs.

Hypothesis 2: For contextual performance, SJTs measuring teamwork skills will have stronger relationships than heterogeneous composite SJTs.

Hypothesis 3: For contextual performance, SJTs measuring leadership skills will have stronger relationships than heterogeneous composite SJTs.

To the extent that task performance requires an understanding of the skills needed to perform job-specific tasks, it should be related to job knowledge and skills (e.g., Borman, White, & Dorsey, 1995; Kanfer

& Ackerman, 1989; Van Scotter & Motowidlo, 1996). Indeed, empirical evidence supports a relationship between task performance and job knowledge (Borman et al., 1995; Schmidt, Hunter, & Outerbridge, 1986). Therefore, we hypothesized:

Hypothesis 4: For task performance, SJTs measuring job knowledge and skills will have stronger relationships than heterogeneous composite SJTs.

Finally, to the extent that managerial performance requires interpersonally oriented behaviors, it should be related to applied social skills such as leadership and interpersonal skills. There is considerable conceptual overlap for SJTs assessing leadership with managerial performance ratings. Indeed, leadership skills such as motivating and managing others, handling people, and directing and structuring subordinate activities are core aspects of management performance (Borman & Brush, 1993; Mumford et al., 2000). In addition, interpersonal skills that are not specific to leadership *per se* should be related to managerial performance. Effective management involves interpersonal interactions including communication, conflict resolution, and negotiations (Borman & Brush, 1993; Campbell, McCloy, Oppler, & Sager, 1993; Conway, 1999; Katz, 1955; Mumford et al., 2000). As such, managers proficient in many of the interpersonal skills assessed by SJTs will likely be rated favorably on managerial performance. Consistently, empirical evidence supports the relationship between managerial performance and leadership skills (Connelly, et al., 2000; Conway, 1999) and interpersonal skills (Conway, 1999; Scullen et al., 2003). Hence we hypothesized:

Hypothesis 5: For managerial performance, SJTs measuring leadership skills will have stronger relationships than heterogeneous composite SJTs.

Hypothesis 6: For managerial performance, SJTs measuring interpersonal skills will have stronger relationships than heterogeneous composite SJTs.

SJT format. Conceptually, SJTs can be developed to measure any construct. Realistically however, some methods lend themselves to measurement of certain constructs more readily than others. In addition, the nature of our arguments for maintaining the method/construct distinction in SJT research suggests that different SJT formats may be a potential moderator of SJT construct–performance relationships. Indeed, administration method can affect the equivalence of tests developed to measure the same construct (Edwards & Arthur, 2007; Ployhart, Weekley, Holtz, & Kemp, 2003; Sackett et al., 2001; Schmitt, Clause, & Pulakos, 1996).

Therefore, important questions in the SJT literature concern whether different constructs are measured with different test delivery formats (e.g., video-based; paper-and-pencil) and whether test format moderates the criterion-related validity of SJT construct domains (McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006; Weekley & Jones, 1997).

Our literature review revealed that SJTs are most commonly paper-and-pencil and video-based formats. Video-based formats are arguably higher in physical and psychological fidelity than paper-and-pencil formats (cf. Bass & Barrett, 1972) because video-based SJTs are more likely to depict ambient contextual details and hence should more realistically reproduce the job performance content domain. Therefore, because higher-fidelity simulations more closely model actual work behaviors than paper-and-pencil tests, video-based SJTs should be more strongly related to actual job performance (McDaniel et al., 2006; Motowidlo et al., 1990; Weekley & Jones, 1997). In addition, video-based SJTs should enhance applicant test perceptions (e.g., face validity), which are positively related to performance (e.g., Chan & Schmitt, 1997; Edwards & Arthur, 2007; Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993).

Further, a video-based format is likely to facilitate the measurement of constructs that rely on subtle social cues and complex contextual information such as applied social skills. These cues can be replicated and transmitted using a video-based SJT more easily than a paper-and-pencil SJT. Therefore, we expect that validities for applied social skills constructs will be higher for video-based SJTs than for paper-and-pencil SJTs. Conversely, constructs such as job knowledge and skills do not typically require such a high level of contextual information, as they rely more on the depiction of task elements rather than social elements. Therefore, there is likely no difference in validities between video-based and paper-and-pencil SJTs that measure job knowledge and skills. As such, we hypothesized:

Hypothesis 7: For the domains of interpersonal skills, teamwork skills, and leadership skills, video-based SJTs will have stronger relationships with job performance than paper-and-pencil SJTs.

Method

Literature Search

An extensive search of computerized databases (PsycINFO, Social Sciences Citation Index, ABInform, and Google Scholar) along with the reference lists of previous meta-analyses of SJTs was conducted to identify

studies that reported the use of SJTs. In addition, unpublished manuscripts were requested from researchers identified as having presented a relevant paper at the annual conference of Academy of Management and the Society for Industrial and Organizational Psychology or having published research on SJTs in 2005–2008. Finally, authors of articles in which the descriptive statistics were not reported in the manuscript were contacted to obtain these data. Based on our search, we obtained 161 manuscripts and articles.

Inclusion Criteria

Studies were included if they used any format of delivery for the SJT (e.g., video based, Web based, computer based, paper and pencil, interview). We omitted studies that measured predictor or performance constructs relevant only to students (e.g., study skills, GPA). After implementing these criteria, we obtained 136 independent data points from 85 studies for the typology, of which 134 independent data points from 84 studies were useable (i.e., provided the data necessary) for the meta-analysis of criterion-related validity.

Coding of Constructs

With respect to recording the information for both the typology and meta-analysis (e.g., correlations, artifacts, construct information), articles were independently coded by two of the study authors. In case of disagreement, all three authors went back to the primary study to reach consensus. Initial agreement for the predictor construct coding was 95%. Constructs were recorded at the lowest level of specificity from studies reporting the constructs assessed or providing enough content information for us to determine constructs measured. Some studies provided lists of the KSAOs measured in addition to broad constructs for which a score was provided (e.g., *giving advice* and *demonstrating empathy* were KSAOs bundled under the construct label *working effectively with others* in one study; O'Connell, McDaniel, Grubb, Hartman, & Lawrence, 2007). We classified SJTs into our typology based on the lowest level construct for which a score was provided (i.e., score-level constructs). The remainder of this paper refers to the construct labels at the lowest level of specificity for which a score was provided (i.e., score level) as *constructs.* For example, although O'Connell et al. (2007) reported measuring *giving advice* and *demonstrating empathy*, scores were not provided at this level. Instead, a score was provided that was the composite of these two constructs. O'Connell et al. labeled the composite *working effectively with others*,

which was the construct we coded. In the absence of a label, we looked for KSAOs or item-level content to determine the construct.

Construct Domains and Construct Typology

We sorted all of the SJT constructs into the construct domains in Huffcutt et al.'s (2001) typology. Of the 136 independent effects in the typology, 36 had different construct labels, which were sorted into the eight domains shown in Table 1. These domains were subsumed under the construct categories, knowledge and skills, applied social skills, and basic personality tendencies from Huffcutt et al. (2001). We found no studies measuring domains subsumed under Huffcutt et al.'s categories of mental ability, interests and preferences, organizational fit, and physical attributes. Finally, we created a fourth category for SJTs that were unclassifiable because they solely reported effect sizes based on composite scores. We labeled the fourth category *heterogeneous composites*. Although we argued that heterogeneous composites are not theoretically meaningful in a construct-based approach, we reported these data to illustrate the frequency with which composites are used and to compare the criterion-related validity of composites and specific construct domains across the performance dimensions. Our typology and the results of our construct sort are presented in Table 1.

Knowledge and skills. The knowledge and skills category was defined as consisting of three potential construct domains by Huffcutt et al. (2001): job knowledge and skills, education and training, and work experience. We found one of these three construct domains, *job knowledge and skills*, to be representative of the constructs that we sorted into this category. The construct domain job knowledge and skills includes constructs that assessed declarative or procedural knowledge. Included within this construct domain, for example, are SJTs designed to assess knowledge of military procedures, knowing how units are interrelated in the military, or knowledge of how to prioritize job tasks.

Applied social skills. Applied social skills contains constructs related to a respondent's skills to function effectively in social situations and interpersonal contexts (Huffcutt et al., 2001). We found three construct domains to be subsumed within this category: *interpersonal skills, teamwork skills, and leadership skills*. Interpersonal skills were defined as social skills that relate to an individual's skill in interacting with others. Within this construct domain, we included constructs that measured customer service skills, interaction skills, and negotiations. The second construct domain contained within the applied social skills category that we identified was teamwork skills. Teamwork skills involve skills specific to team settings that may promote team effectiveness. For example,

teamwork may involve a combination of collaborative problem solving, coordination, and team planning (e.g., Ellis, Bell, Ployhart, Hollenbeck, & Ilgen, 2005; Morgeson et al., 2005). Applied social skills constructs such as "working effectively with others" were not classified into the teamwork skills category unless they specifically referenced team settings. The third construct domain contained within the applied social skills category we identified was leadership skills. This domain included SJT constructs designed to assess general management skills, such as leadership, supervision, or administrative skills (e.g., Campbell et al. 1993), as well as more specific leadership skills such as resolving conflicts among subordinates, organizing and structuring work assignments, or handling employee problems.

Basic personality tendencies. A number of SJTs identified in the literature measured personality constructs. Consistent with current research on personality, we conceptualized personality as relatively enduring dispositional factors that relate to how employees act in the workplace. The only personality construct domain for which there were a reasonable number of effects to perform a meta-analysis was Conscientiousness with a *k* of seven data points. Four other SJTs that measured personality represented a combination of different personality construct domains, including various composites of Conscientiousness, Agreeableness, Emotional Stability, adaptability, and integrity. We included these in our meta-analysis for illustrative purposes.

Heterogeneous composites. Many researchers provided detailed heterogeneous lists of the constructs measured by an SJT but did not provide construct level scores. In many cases, a heterogeneous composite score for each SJT was reported, making it impossible to sort them into the three categories. For example, if an SJT appeared to measure communication skills Conscientiousness, and leadership ability but only a composite score was reported, we considered the composite score uninterpretable for the purposes of construct-level information. We identified 45 effects that either did not specify the constructs measured by SJTs or collapsed across several constructs and reported a composite score. We included these in Table 1 to document the number of SJTs that did not identify the constructs measured or presented method-level composite scores that for comparative purposes were uninterpretable.

Criterion Types

Consistent with other meta-analyses of predictor methods, in our first set of analyses we combined all performance criteria into an omnibus analysis of the criterion-related validity for each SJT construct domain. We also separated task, contextual, and managerial performance facets

into distinct categories using two decision rules. First, we sorted the performance facets using operationalizations consistent with definitions for task and contextual performance provided by Motowidlo and Van Scotter (1994), Van Scotter and Motowidlo (1996), and Hurtz and Donovan (2000): (a) *task performance*, or the degree to which an individual can perform tasks that are central to the job (e.g., technical skill, sales performance, use of technical equipment, job duties, or core technical proficiency); (b) *contextual performance*, or behaviors not formally required by the job, including interpersonal facilitation (e.g., building and mending relationships, cooperation, helping, consideration, interpersonal relations) and job dedication (e.g., motivation, effort, drive, initiative); and (c) *managerial performance*, or ratings by peers or supervisors of an individual's behaviors related to leadership, interpersonal management skills, management, or administrative duties that constitute core management responsibilities (Borman and Brush, 1993; Conway, 1999; Scullen et al., 2003).

For studies that did not report the specific performance label, we assessed the extent to which the job description or job title indicated that a particular facet of performance was a core task. As noted by others (e.g., Borman & Motowidlo, 1993; Conway, 1999; Witt and Ferris, 2003), the distinction between facets of performance may be blurred depending on the job context. Therefore, in some cases we used the job title to determine whether ratings were most appropriately labeled as task, contextual, or managerial performance. For example, when an interpersonal criterion was rated in a customer service job (e.g., customer contact skills), we classified it as task performance because we assumed that customer contact skills are core technical requirements of this job. Conversely, an interpersonal criterion assessed in a manufacturing job (e.g., ratings of cooperativeness) was considered contextual performance because we assumed that interpersonal skills are not formally required in many manufacturing jobs. Finally, when an interpersonal criterion that is a core part of a managerial job was used for a management position (e.g., communication skills), we classified it as managerial performance. Initial agreement in coding of the criterion facets (before reaching consensus) was 92%.

Criterion-Related Validity Analyses

We used meta-analysis (e.g., Hunter & Schmidt, 2004; Raju, Burke, Normand, & Langlios, 1991) to calculate corrected mean population-level estimates of the criterion-related validity of each construct domain. This procedure allows for the correction of effect sizes for measurement error and other statistical artifacts, based on the idea that differences in

TABLE 2
Mean Sample-Based Reliability Estimates Used for Analyses

Analysis	*k*	*N*	Estimate of reliability
Job performance	22	2,169	.58
Task performance	6	819	.59
Contextual performance	5	642	.51
Managerial performance	5	288	.67

Note. Each artifact distribution was calculated as a sample-size weighted average for all studies that reported interrater reliability information. Estimates of range restriction were not available in the primary studies.

the results of primary studies are due to statistical artifacts rather than actual differences within the population. For this analyses, we calculated approximately defined artifact distributions using sample-size weighted estimates taken from the sample of studies for each estimated effect, a practice that generates slightly more accurate estimates of the mean and variance of rho than population-level standard errors (Raju et al., 1991). We corrected for unreliability in the criterion (i.e., operational validity) using estimates of interrater reliability, which account for more sources of measurement error than internal consistency (Schmidt, Viswesvaran, and Ones, 2000). When a study reported an interrater reliability estimate, this was used in our corrections for that study; however when a study did not report interrater reliability, we corrected using the assumed reliability estimate found in Table 2. These assumed values were computed using a sample-weighted mean estimate from the distribution of studies for each effect. No corrections for range restriction were made because these estimates were not available from most studies.

In addition, we utilized a random effects model, which results in more accurate Type I error rates and more realistic confidence intervals than does a fixed effects model (e.g., Erez, Bloom, & Wells, 1996; Overton, 1998). Therefore, we placed a 95% confidence interval around each mean-corrected effect, which represents the extent to which the corrected effect may vary if other studies from the population were included in the analysis (for elaboration, see Burke & Landis, 2003). We also calculated credibility intervals, which indicate the extent to which correlations varied across studies for a particular analysis distribution; that is, 80% of the values in the population are contained within the bounds of the interval (Hunter & Schmidt, 2004). Finally, in many cases studies provided multiple correlations from the same sample and the same predictor construct or criterion construct, which were nonindependent (e.g., Lipsey & Wilson, 2001). In such cases, we created a single effect to represent the range of nonindependent effects using sample size-weighted composite correlations.

TABLE 3
Omnibus Analysis of Criterion-Related Validities of SJT Construct Domains for Job Performance

Construct category and domain	k	N	M_r	M_ρ	95% CI L	95% CI U	$SE_{M\rho}$	80% CV L	80% CV U	SD_ρ
Knowledge and skills										
Job knowledge and skills	4	695	.15	**.19**	.07	.32	.06	.08	.30	.09
Applied social skills										
Interpersonal skills	17	8,625	.19	**.25**	.20	.31	.03	.12	.39	.11
Teamwork skills	6	573	.29	**.38**	.26	.52	.07	.24	.53	.11
Leadership	51	7,589	.21	**.28**	.24	.32	.02	.15	.40	.10
Basic personality tendencies										
Personality composites	4	423	.30	**.43**	.30	.57	.07	.35	.52	.14
Conscientiousness	7	908	.19	**.24**	.13	.34	.05	.12	.35	.14
Heterogeneous composites	45	9,681	.21	**.28**	.24	.31	.02	.19	.36	.07

Note. k = the number of independent effect sizes included in each analysis; N = sample size; M_r = mean sample-weighted uncorrected correlation; M_ρ = operational validity (corrected for criterion unreliability); $SE_{M\rho}$ = standard error of M_ρ; SD_ρ = standard deviation of rho.

Results

Classification Typology and Construct Frequency

The results of the construct classification are shown in Table 1. The column on the left indicates the construct category and domains into which the constructs were sorted. The middle column contains the construct labels recorded from each study. The two columns on the right represent the number of effects at the construct level and the percent of the SJTs that measured a specific construct domain, respectively. Of these studies, the majority measured leadership (37.50%), followed by interpersonal skills (12.50%), basic personality tendencies (9.56%), teamwork skills (4.41%), and job knowledge and skills (2.94%). SJTs that were unclassifiable (i.e., they reported method-level composite effects) constituted 33.09% of the data points.

Validity of SJTs for Construct Domains

Table 2 presents the artifact distributions used in each analysis. Unless reported otherwise, for all mean effects reported below, the 95% confidence interval did not overlap zero. Specific information on the confidence intervals and other meta-analytic findings are reported in Tables 3, 4 and 5.

TABLE 4
Criterion-Related Validities of SJT Construct Domains for Job Performance Facets

Construct category and domain	k	N	M_r	M_ρ	95% CI L	95% CI U	$SE_{M\rho}$	80% CV L	80% CV U	SD_ρ
Contextual performance										
Knowledge and skills										
Job knowledge and skills	1	83	.27	-	-	-	-	-	-	-
Applied social skills										
Interpersonal skills	3	1,364	.17	**.21**	.04	.39	.09	.02	.40	.15
Teamwork skills	6	573	.27	**.35**	.23	.47	.06	.24	.46	.09
Leadership	5	3,034	.19	**.24**	.18	.31	.03	.17	.32	.06
Heterogeneous composites	8	2,387	.14	**.19**	.12	.25	.03	.12	.25	.05
Task performance										
Knowledge and skills										
Job knowledge and skills	1	82	.39	-	-	-	-	-	-	-
Applied social skills										
Interpersonal skills	6	1,818	.19	**.25**	.14	.36	.06	.10	.40	.12
Teamwork skills	3	232	.39	**.50**	.32	.68	.09	.36	.64	.11
Leadership	9	4,039	.17	**.21**	.15	.28	.03	.10	.33	.09
Basic personality tendencies										
Personality composites	3	316	.33	**.45**	.31	.60	.07	.37	.53	.13
Conscientiousness	3	268	.30	**.39**	.28	.49	.05	-	-	.00
Heterogeneous composites	19	5,416	.21	**.27**	.21	.33	.03	.13	.41	.11
Managerial performance										
Knowledge and skills										
Job knowledge and skills	2	931	.19	**.23**	.13	.34	.06	.16	.31	.06
Applied social skills										
Interpersonal skills	2	297	.29	**.36**	.21	.51	.08	.29	.43	.06
Leadership	17	3,769	.24	**.29**	.24	.35	.03	.17	.41	.09
Basic personality tendencies										
Conscientiousness	2	174	.05	**.06**	.00	.12	.03	-	-	.00
Heterogeneous composites	5	1,282	.10	**.12**	.08	.16	.02	-	-	.00

Note. k = the number of independent effect sizes included in each analysis; N = sample size; M_r = mean sample-weighted uncorrected correlation; M_ρ = operational validity (corrected for criterion unreliability); $SE_{M\rho}$ = standard error of M_ρ; SD_ρ = standard deviation of rho. Credibility values were not computed for effects that had zero estimates for SD_ρ. Job knowledge and skills estimates were not computed for task or contextual performance due to lack of data, however single data points are presented.

We do caution the reader that in some cases the estimates are based on low *k*s and should be interpreted with caution.

Table 3 presents the results of the omnibus analysis of criterion-related validity for each of the SJT construct domains. As shown in Table 3, SJTs that measured teamwork skills had a mean validity of .38. SJTs that assessed leadership skills had a mean validity of .28. SJTs assessing

TABLE 5
Effects of SJT Format on the Criterion-Related Validities of SJT Construct Domains (Across Job Performance Facets)

Construct category and domain	k	N	M_r	M_ρ	95% CI L	95% CI U	$SE_{M\rho}$	80% CV L	80% CV U	SD_ρ
Applied social skills										
Interpersonal skills										
Paper-and-pencil	15	8,182	.20	**.27**	.22	.32	.03	.16	.38	.08
Video-based	2	437	.36	**.47**	.39	.55	.04	-	-	.00
Leadership										
Paper-and-pencil	47	6,938	.21	**.27**	.23	.31	.02	.14	.41	.10
Video-based	4	651	.25	**.33**	.25	.40	.04	-	-	.00
Heterogeneous Composites										
Paper-and-pencil	40	7,316	.20	**.25**	.22	.29	.02	.17	.33	.06
Video-based	5	2,365	.28	**.36**	.30	.42	.03	.31	.41	.04

Note. k = the number of independent effect sizes included in each analysis; N = sample size; M_r = mean sample-weighted uncorrected correlation; M_ρ = operational validity (corrected for criterion unreliability); $SE_{M\rho}$ = standard error of M_ρ; SD_ρ = standard deviation of rho. Credibility values were not computed for effects that had zero estimates for SD_ρ.

interpersonal skills had a mean validity of .25; and SJTs assessing Conscientiousness had a mean validity of .24. Although based on only four studies each, SJTs measuring job knowledge and skills had a mean validity of .19, and personality composites had a mean validity of .43. Finally, for heterogeneous composite SJTs, we obtained a mean validity of .28.

Moderator Analyses

Criterion-related validity by criterion facet. Table 4 presents the results of the analyses of criterion-related validity for each construct domain, broken down by task, contextual, and managerial performance. SJT construct domains had criterion-related validities that were consistent with our hypotheses for trends in their magnitudes within each criterion type, although in many cases confidence intervals overlapped. First, for contextual performance, we predicted that SJTs assessing interpersonal skills (Hypothesis 1), teamwork skills (Hypothesis 2), and leadership skills (Hypothesis 3) would have higher validities than heterogeneous SJTs. The validities were in the expected direction for SJTs assessing interpersonal skills ($M_\rho = .21$), teamwork skills ($M_\rho = .35$), and leadership skills ($M_\rho = .24$) when compared with heterogeneous composites ($M_\rho = .19$), suggesting support for Hypotheses 1, 2, and 3; although the k for

interpersonal skills was only 3. For task performance, we hypothesized that SJTs assessing job knowledge and skills would have higher validities than heterogeneous composite SJTs; unfortunately, we could not estimate this relationship because there was only one primary study available for analysis ($r = .39$). Finally, for managerial performance, we hypothesized that SJTs assessing leadership (Hypothesis 5) and interpersonal skills (Hypothesis 6) would have higher validities than heterogeneous composite SJTs. SJTs assessing leadership ($M_\rho = .29$) and interpersonal skills ($M_\rho = .36$) were more strongly related to managerial performance than were heterogeneous composites ($M_\rho = .12$), suggesting support for Hypotheses 5 and 6, although the k for interpersonal skills was only 2.

Criterion-related validity by test format. Table 5 presents the results of the moderator analyses of test format. Consistent with Hypothesis 7, video-based SJTs tended to have stronger relationships with job performance than paper-and-pencil SJTs. Further, although the ks for video-based SJTs were relatively small, for two out of the three dimension comparisons the confidence intervals for the video-based and paper-and-pencil formats did not overlap. Our estimates for video-based SJTs measuring interpersonal skills ($M_\rho = .47$) were higher than those for paper-and-pencil SJTs measuring interpersonal skills ($M_\rho = .27$); however the k for the video-based estimate was limited to 2. We also obtained higher estimates for video-based tests measuring leadership ($M_\rho = .33$) than for paper-and-pencil tests measuring leadership ($M_\rho = .27$), although the confidence intervals overlapped and the video-based estimate was based on only four studies. Finally, video-based heterogeneous composites ($M_\rho = .36$) had higher criterion-related validity than paper-and-pencil heterogeneous composites ($M_\rho = .25$).

Discussion

Although SJTs are widely used in employee selection and commonly researched in academe, little is known about the specific constructs assessed by SJTs because researchers and authors typically report results and frame their data more in method terms than in construct terms. Therefore, the primary objectives of this study were to (a) discuss the advantages of attending to and reporting SJT construct-level versus method-level results; (b) develop a typology of constructs that have been assessed by SJTs in the extant literature; and (c) undertake an initial examination of the criterion-related and incremental validity of the identified constructs and to investigate moderators of these validities. We view our efforts as contributing to an initial classification and description of the constructs assessed by the SJT method in the extant literature. We do, however, recognize that as with any first attempt to provide structure to a

nebulous body of literature (e.g., Arthur et al., 2003; Huffcutt et al., 2001), our undertaking will likely be refined and expanded as future research is conducted.

With regard to our first objective, we believe a fundamental issue limiting advancements in understanding and using SJTs for selection is the common failure to disentangle the effects of the measurement method (i.e., the SJT) from the constructs measured by the test. This has significant implications for the use of test scores. For instance, it limits the ability to compare different predictors that are confounded by method and/or construct variance (i.e., comparing apples and oranges). Understanding why a given test predicts performance is important to both researchers and practitioners for measurement, theory testing, or establishing the job-relevancy of the selection tool. Further, the failure to attend to constructs limits the generalizability of SJTs. For example, if one researcher reports a criterion-related validity coefficient of .19 for an SJT in a textile company, the only information gained is that the SJT predicts performance for that job in that company. Without any information about the constructs measured, it remains unclear why the particular test is valid or whether it would be valid in another job or industry. Identification of the construct(s) measured by the SJT, however, offers a point of comparison that would enable practitioners to transport SJTs across contexts.

As part of our construct-based approach, we identified the constructs measured by SJTs in the extant literature (relying on studies that reported construct information). The underlying rationale behind this objective was that a construct typology would provide a common and systematic framework for understanding and applying SJT constructs. We found that a substantial number (33%) of the SJTs in the literature did not report the constructs measured, did not provide enough information to determine the constructs measured, or provided only a composite score, which collapsed across multiple constructs. Nevertheless, our analyses revealed that SJTs are in some cases developed to assess specific constructs, most often leadership skills (38%) and interpersonal skills (13%). Less frequently, SJT studies reported assessing teamwork skills (4%), personality tendencies (10%), and job knowledge (3%). Plausibly, SJTs are often used to measure leadership skills and interpersonal skills because they offer a convenient method for sampling applicants' performance on complex tasks that are otherwise expensive, time consuming, or difficult to assess. In particular, SJTs are well suited to measure behaviors elicited by complex interpersonal and administrative situations, as they often contain ambient details that create rich representations of contextual features. Moreover, although other simulation-based predictor methods offer similar benefits (i.e., work-samples, assessment centers, situational interviews), SJTs typically have a much lower cost of administration and scoring (e.g., Motowidlo et al., 1990; Weekley & Jones, 1999).

Our final objective was to meta-analytically assess the criterion-related validity of each construct domain. Teamwork skills, leadership skills, and interpersonal skills all exhibited relatively high validities for job performance. The criterion-related validities we obtained for Conscientiousness and job knowledge were relatively lower. In addition, we performed two sets of moderator analyses to highlight the benefits of taking a construct-based approach in SJT research. Analyses of the relationships between each predictor construct domain and narrow job performance facets provided support for our typological framework by demonstrating a pattern of relationships (i.e., differential validities) that was consistent with expectations derived from content-based matching of predictor and criterion constructs. Nevertheless, we offer two caveats. First, several of our estimates may be unstable because they were based on small k (e.g., the relationship between teamwork skills and task performance was based on $k = 3$). Second, and likely a result of the first caveat, a few results were not in the direction that might be predicted from the literature. For example, teamwork skills had higher validities for task performance than for contextual performance, but teamwork skills logically seem to be more important for contextual performance. Nevertheless, overall, our findings are of practical significance in that appropriate matching between the content domains of predictors and criteria can strengthen the criterion-related validity of SJTs. Hence, the construct-based approach also could be advantageously applied to the performance domain, in which researchers historically report results using overall or composite job performance (cf. Campbell, 1990).

We have argued that the construct-based approach is a powerful tool for isolating variance due to constructs from method variance. Specifically, the second set of moderator analyses were intended to illustrate the role of method characteristics in determining which constructs might be measured by different formats. We expected that the type of SJT format would moderate SJT criterion-related validities, perhaps by influencing which constructs were measured and how well they were measured. In line with expectations, we found that for each construct domain, video-based SJTs were more strongly correlated with performance than paper-and-pencil SJTs. This information is a significant contribution to the literature because we were able to cross methods and constructs in similar fashion to the suggestions of Campbell and Fiske (1959) for multitrait, multi method matrices.

Why SJT Research Has Been Method-focused Rather Than Construct Focused

There are several possible explanations for why researchers have neglected to specify the constructs measured by SJTs. First, although SJTs

are commonly considered methods, they are often treated as if they are actually measuring a single construct (e.g., situational judgment, practical intelligence). Indeed, even when researchers identified the KSAOs being measured, many studies in our sample reported a single composite score labeled "situational judgment" (e.g., Chan & Schmitt, 2002; Motowidlo et al., 1990; Smith & McDaniel, 1998; Swander, 2000; Weekley & Jones, 1997, 1999). Second, it may simply be difficult to create SJTs that can be scored at the construct level, given current developmental paradigms (Ployhart & Ryan, 2000a). As Ployhart and Ryan suggested, refinements to typical critical incident-based development procedures may be effective; specifically, researchers could delineate the constructs to be assessed *a priori*, conceptualize how the constructs should manifest in work situations, and write response options that correspond to the range of a *single* behavior (i.e., high or low on "demonstrating effort") rather than multiple types of behaviors.

Finally, perhaps one of the most important reasons that SJT research has not focused on construct-level information is that I-O psychologists have only recently begun developing and implementing a construct-oriented paradigm for selection research. As noted by Schmitt and Chan (2006), it remains a problem that "Our field as a whole . . . is more apt to discuss the validity of methods rather than the validity of measurement of constructs" (p. 136). Indeed, although the idea of construct validity has been around for decades (e.g., Campbell & Fiske, 1959), the emphasis on constructs in much of the personnel selection research has only recently gained in importance (e.g., Anastasi & Urbina, 1997; Arthur & Villado, 2008; Binning & Barrett, 1989; Huffcutt et al., 2001; Messick, 1995; Roth et al., 2008). As a result, the importance of constructs may have been less salient to researchers for much of the older SJT literature. On the other hand, recent studies are not immune to the problem; we identified several recently published studies that neglected to report construct-level information.

Where Do We Go From Here? Recommendations for SJT Research and Practice

Although other researchers have pointed out that SJTs are methods of measurement and should therefore attend to construct-level information (e.g., McDaniel et al., 2001; Schmitt & Chan, 2006), the results of this study suggest a need for further development of a construct-oriented paradigm in SJT research. Many of the limitations found in our meta-analysis illustrate the state of the literature at present and therefore highlight gaps that could benefit from a construct-based approach. As such, based on our observations, we next offer recommendations for

research and for practice, with the recognition that the two are not mutually exclusive.

Research recommendations. First, SJT researchers should report detailed construct information. A relatively large number of studies (33%) failed to report the constructs measured or reported composite method-level information (in the heterogeneous category). As a result, one limitation of this research was small sample sizes for a number of construct domains. We urge researchers to maintain the distinction between methods (e.g., SJTs) and constructs (e.g., leadership skills) by reporting information about the specific constructs measured by SJTs as well as reliability estimates, means, standard deviations, group differences, and intercorrelations with other constructs. Providing this information would allow for more meaningful comparisons of subgroup differences (e.g., race, sex), validity (e.g., construct, criterion-related, incremental validity), or test-taker reactions to different measurement methods or different predictor constructs. In this vein, we believe that the typology developed in this study will facilitate the reporting of constructs by providing researchers and practitioners with a common framework to communicate research findings and validity evidence of constructs (cf. Fleishman & Quintance, 1984; Hough & Ones, 2001).

Furthermore, refinements to construct validation procedures typically used for SJTs would be helpful. For instance, without evidence of convergent or discriminant validity, we were unable to determine the full extent to which the SJTs reported in the literature actually measured the constructs they were purported to measure. Of course, this is a criticism leveled at any meta-analysis of predictive validity in which meta-analysts rely on the primary authors' conclusion that the tests being analyzed actually measured the intended constructs. As primary researchers begin to provide more information about both the constructs measured and the extent to which the SJT displayed convergent or discriminant validity with other measures, this criticism can be addressed. In particular, future meta-analytic research could combine the approach taken in this study (i.e., coding constructs based on labels and content) with the approach taken by McDaniel and colleagues (i.e., correlating SJTs with measures of constructs) to obtain multiple sources of construct validity evidence.

In addition, researchers should utilize SJT construct information to hold constructs constant, in order to identify and investigate various features of SJT methodology that impact relevant outcomes. For example, one could compare different dimensions of stimulus material (e.g., paper-and-pencil, computerized), different response modalities (e.g., written, oral), or different scoring strategies (e.g., empirical keying, subject matter experts). Such comparisons are only meaningful if the construct is held constant across different methodological dimensions.

Moreover, it is also the case that predictor methods could influence or constrain the constructs measured (e.g., Arthur & Villado, 2008; Messick, 1995). For example, the measurement of applied social skills might be more easily done with video-based testing than paper-and-pencil testing because video-based tests provide more details and contextual information (e.g., nonverbal behaviors; environmental cues) that are important to social skills. Likewise, SJTs are well suited to measure dimensions of contextual job knowledge, which would be applied to practical problems that are ill-defined, contain incomplete information, or have different solutions. This type of knowledge might be contrasted to job knowledge that relies on facts and procedures that are well defined. In fact, Schmitt and Chan (2006) have already posited that SJTs place some constraints on the range of constructs measured and it would be helpful to identify such boundary conditions. As part of this effort, we would encourage researchers to examine the extent to which there is a strong method factor or higher-order construct measured by SJTs (e.g., judgment or practical intelligence; Schmitt & Chan, 2006).

In addition, whenever possible, SJT researchers should conduct predictive validation studies. Another limitation of this meta-analysis is that we were unable to correct for range restriction because most studies in our dataset were either conducted concurrently or failed to provide sufficient information to make this correction. As a result, our estimates are conservative. Furthermore, given that SJTs are job-centered tests often developed to assess job-relevant behaviors and validated by checking them against the criteria of behaviors actually performed on the job, then the concurrent validation studies we reviewed may be considered estimates of convergent validity. Concurrent, cross-sectional studies are suggestive but cannot adequately evaluate substantive theoretical links between SJT and criterion constructs. Therefore, we would recommend the use of longitudinal, predictive criterion-validation designs.

Practice recommendations. Practitioners can benefit from the construct-based approach by identifying the focal construct(s) of interest before choosing a selection methodology with which to measure the construct(s). In practice, job analysis determines which constructs are to be measured, but practitioners have some latitude in determining which method of measurement to use. Our meta-analytic validity estimates can be useful for seeking methods to measure specific KSAOs. For a given predictor construct, practitioners may consult our study to compare the expected validity with other methods such as interviews (Huffcutt et al., 2001) or higher fidelity simulations (e.g., assessment centers; Arthur et al., 2003). For example, the results of this study indicated that the criterion-related validity for SJTs measuring teamwork skills ($M_\rho = .38$) was slightly higher than the criterion-related validity for the consideration and

awareness of others dimension of assessment centers ($M_\rho = .33$) obtained by Arthur et al. (2003).

In addition, practitioners should consider the criterion carefully when choosing predictor construct(s) to measure using SJTs. Our results demonstrated the potential for nontrivial increases in validity when SJT predictor constructs were matched conceptually with narrowly defined relevant criteria. Our data also suggested that one realm where SJTs often are appropriately matched between predictor construct and criterion is for the prediction of managerial performance. Our results showed that 17 of the 38 data points for managerial performance were SJTs measuring leadership. In most cases SJTs measuring specific constructs had stronger validities than heterogeneous composites. Therefore, in the interest of maximizing predictive power, SJT test developers should consider the criterion of interest when selecting a specific construct to measure using a given SJT.

Conclusions

In conclusion, we have highlighted the importance of a construct-based focus in SJT research. We urge researchers to present results at the construct level when possible (Arthur & Villado, 2008). Such information, as noted by Huffcutt et al. (2001), Arthur et al. (2003), and Roth et al. (2008) in their similar request with regard to interviews, assessment centers, and work samples, will provide future researchers and practitioners with better conceptual, theoretical, and practical understanding of SJTs.

REFERENCES

*indicates studies included in the meta-analysis and typology.

†indicates studies included in the typology only.

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

Anastasi A, Urbina S. (1997). *Psychological testing*. Upper Saddle River, NJ: Prentice Hall.

*Anonymous. (1954). Validity information exchange (No. 7-065). PERSONNEL PSYCHOLOGY, *7*, 201.

Arthur W, Jr., Day EA, McNelly TL, Edens PS. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. PERSONNEL PSYCHOLOGY, *56*, 125–154.

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

*Banki S, Latham GP. (2008, August). *The validity of the situational interview and situational judgment test in Iran*. Paper presented at the 67th annual meeting of the Academy of Management, Anaheim, CA.

Bartram D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185–1203.

Bass BM, Barrett GV. (1972). *Man, work and organizations: An introduction to industrial and organizational psychology*. Oxford, UK: Allyn & Bacon.

*Bass BM, Karstendick B, McCullough, G, Pruitt RC. (1954). Validity information exchange (No 7-024). PERSONNEL PSYCHOLOGY, *7*, 159–60.

*Bergman ME, Donovan MA, Drasgow F, Overton RC. (2001a). *Assessing contextual performance: Preliminary tests of a new framework*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

*Bergman ME, Drasgow F, Donovan MA, Juraska SE. (2001b). *Scoring situational judgment tests*. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

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

Borman W, Brush D. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance, 6*, 1–21.

*Borman WC, Hanson MA, Oppler SH, Pulakos ED, White LA. (1993). Role of early supervisory experience in supervisor performance. *Journal of Applied Psychology, 78*, 443–449.

Borman WC, Motowidlo SJ. (1993). Expanding the criterion domain to include elements of contextual performance. In Schmitt N, Borman WC (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco: Jossey-Bass.

Borman WC, Motowidlo SJ. (1997). Task performance and contextual performance: The meaning for personnel selection. *Human Performance, 10*, 99–109.

Borman WC, White LA, Dorsey DW. (1995). Effects of rate task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology, 80*, 168–177.

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

*Bruce MM, Friesen EP. (1956). Validity information exchange (No. 9-35). PERSONNEL PSYCHOLOGY, *9*, 380.

*Bruce MM, Learner DB. (1958). A supervisory practices test. PERSONNEL PSYCHOLOGY, *11*, 207–216.

Burke MJ, Landis RS. (2003). Methodological and conceptual challenges in conducting and interpreting meta-analyses. In Murphy KR (Ed.), *Validity generalization: A critical review* (pp. 287–309). Mahwah, NJ: Erlbaum.

Campbell JP. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In Dunnette MD, Hough LM (Eds.), *Handbook of industrial and organizational psychology*, 2nd ed., (Vol. 1, pp. 39–74). Palo Alto, CA: Consulting Psychologists Press.

Campbell DT, Fiske DW. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Campbell JP, McCloy RA, Oppler SH, Sager CE. (1993). A theory of performance. In Schmitt N, Borman WC (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco: Jossey Bass.

*Canter RR. (1951). A human relations training program. *Journal of Applied Psychology, 35*, 38–45.

*Carter G. (1952). Measurement of supervisory ability. *Journal of Applied Psychology, 36*, 393–395.

Chan D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology, 82*, 311–320.

*Chan D. (2002, April). *Situational effectiveness x proactive personality interaction on job performance*. Paper presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Canada.

Chan KY, Drasgow F. (2001). Toward a theory of individual differences and leadership: Understanding the motivation to lead. *Journal of Applied Psychology, 86*, 481–498.

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

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

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

*Clevenger J, Pereira GM, Wiechmann D, Schmitt N, Schmidt-Harvey V. *The situational judgment inventory as a measure of contextual job knowledge*. Unpublished manuscript.

*Colonia-Willner R. (1998). Practical intelligence at work: Relationship between aging and cognitive efficiency among managers in a bank environment. *Psychology and Aging, 13*, 45–57.

Connelly MS, Gilbert JA, Zaccaro SJ, Threlfall KV, Marks MA, Mumford MD. (2000). Exploring the relationship of leadership skills and knowledge to leader performance. *Leadership Quarterly, 11*, 65–87.

Conway J. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology, 84*, 3–13.

*Cucina JM, Vasilopoulos NL, Leaman JA. (2003, April). *The bandwidth-fidelity dilemma and situational judgment test validity*. Poster presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

*Decker RL. (1956). An item analysis of how supervise? Using both internal and external criteria. *Journal of Applied Psychology, 40*, 406–411.

*Dicken CF, Black JD. (1965). Predictive validity of psychometric evaluations of supervisors. *Journal of Applied Psychology, 49*, 34–47.

*Dulsky SG, Krout MH. (1950). Predicting promotion potential on the basis of psychological tests. PERSONNEL PSYCHOLOGY, *3*, 345–351.

Edwards BD, Arthur W, Jr. (2007). An examination of factors contributing to a reduction in race–based subgroup differences on a constructed response paper–and–pencil test of scholastic achievement. *Journal of Applied Psychology, 92*, 794–801.

*Edwards WR, Schleicher DJ. (2004). On selecting psychology graduate students: Validity evidence for a test of tacit knowledge. *Journal of Educational Psychology, 96*, 592–602.

*Elias DA, Shoenfelt EL. (2001, April). *Use of a situational judgment test to measure teamwork components and their relationship to overall teamwork performance*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Ellis APJ, Bell BS, Ployhart RE, Hollenbeck JR, Ilgen DR. (2005). An evaluation of generic teamwork skills training with action teams: Effects on cognitive and skill-based outcomes. PERSONNEL PSYCHOLOGY, *58*, 641–672.

Erez A, Bloom MC, Wells MT. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. PERSONNEL PSYCHOLOGY, *49*, 275–306.

Ferris GF, Witt LA, Hochwarter WA. (2001). Interaction of social skill and general mental ability on job performance and salary. *Journal of Applied Psychology, 86*, 1075–1082.

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

File QW, Remmers HH. (1971). *How supervise? Manual 1971 revision*. Cleveland, OH: Psychological Corporation.

Fleishman EA, Quaintance MK. (1984). *Taxonomies of human performance: The description of human tasks*. Orlando, FL: Academic Press, Inc.

*Forhand GA, Guetzkow H. (1961). The administrative judgment test as related to descriptions of executive judgment behaviors. *Journal of Applied Psychology, 45*, 257–261.

*Funke U, Schuler H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment, 6*, 115–123.

*Hanson MA. (1994). Development and construct validation of a situational judgment test of supervisory effectiveness for first-line supervisors in the U.S. Army. (Doctoral dissertation, University of Minnesota, 1994). *Dissertation Abstracts International, 56*(2-B), 1138.

*Hilton AC, Bolin SF, Parker JW, Jr., Taylor EK, Walker WB. (1955). The validity of personnel assessments by professional psychologists. *Journal of Applied Psychology, 39*, 287–293.

Hogan J, Holland B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*, 100–112.

*Holmes FJ. (1950). Validity of tests for insurance office personnel. PERSONNEL PSYCHOLOGY, *3*, 57–69.

Hough LM, Ones DS. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In Anderson N, Ones DS, Sinangil HK, Viswesvaran C (Eds.), *Handbook of industrial, work, and organizational psychology* (Vol. 1, pp. 233–377). London: Sage.

*Howard A, Choi M. (2000). How do you assess a manager's decision-making abilities? The use of situational inventories. *International Journal of Selection and Assessment, 8*, 85–88.

Huffcutt AI, Conway JM, Roth PL, Stone NJ. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 80*, 897–913.

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

Hunter JE, Schmidt FL. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hurtz GM, Donovan JJ. (2000). Personality and job performance: The big five revisited. *Journal of Applied Psychology, 85*, 869–879.

*Johnson RJ. (1954). Validity information exchange (No. 7-090). PERSONNEL PSYCHOLOGY, *7*, 567.

*Jones C, Decotis TA. (1986). Video-assisted selection of hospitality employees. *The Cornell H.R.A. Quarterly, 27*, 68–73.

*Jurgensen CE. (1959). Supervisory practices test. In Burros OK (Ed.), *The fifth mental measurements yearbook* (pp. 946–947). Highland Park, NJ: Gryphon Press.

Kanfer R, Ackerman PL. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*, 657–690.

Katz RL. (1955). Skills of an effective administrator. *Harvard Business Review, 33*, 33–42.

*Kerr MR *Tacit Knowledge as a predictor of managerial success: A field study*. Royal Canadian Mounted Police. Unpublished manuscript.

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

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

Lipsey MW, Wilson DB. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

*Lobenz RE, Morris SB. (1999, April). *Is tacit knowledge distinct from* g*, personality, and social knowledge?* Paper presented at the 14th Annual Convention of the Society for Industrial and Organizational Psychology, Atlanta, GA.

*MacLane CN, Barton MG, Holloway-Lundy AE, Nickels BJ. (2001, April). *Keeping score: Empirical vs. expert weights on situational judgment responses*. Paper presented at the 16th Annual Convention of the Society for Industrial and Organizational Psychology, San Diego, CA.

*McClough AC, Rogelberg SG. (2003). Selection in teams: An exploration of the teamwork knowledge, skills, and ability test. *International Journal of Selection and Assessment, 11*, 56–66.

*McCormick EJ, Middaugh RW. (1956). The development of a tailor-made scoring key for the how supervise? Test. PERSONNEL PSYCHOLOGY, *9*, 27–37.

*McCullough, G, Pruitt RC. (1954). Validity information exchange (No 7-024). PERSONNEL PSYCHOLOGY, *7*, 159–60.

McDaniel MA, Hartman NS, Whetzel DL, Grubb WL. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. PERSONNEL PSYCHOLOGY, *60*, 63–91.

McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.

McDaniel MA, Nguyen NT. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*, 103–113.

McDaniel MA, Whetzel DL, Hartman NS, Nguyen NT, Grubb WL. (2006). Situational judgment tests: Validity and an integrative model. In Weekley J, Ployhart RE (Eds.), *Situational judgment tests* (pp. 183–203). Mahwah, NJ: Erlbaum.

*McDaniel MA, Yost AP, Ludwick MH, Hense RL, Hartman NS. (2004, April). *Incremental validity of a situational judgment test*. Paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

*McElreath J, Vasilopoulos NL. (2002, April). *Situational judgment: Are most and least likely responses the same?* Poster presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Can.

Messick S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

*Meyer HH. (1951). Factors related to success in the human relations aspect of work-group leadership. In Conrad HS (Ed.), *Psychological Monographs: General and Applied* (Vol. 65, pp. 1–12).

*Meyer HH. (1956). An evaluation of a supervisory selection program. PERSONNEL PSYCHOLOGY, *9*, 499–513.

Mohammed S, Mathieu J, Bartlett A. (2002). Technical-administrative task performance, leadership task performance, and contextual performance: Considering the influence of team- and task-related composition variables. *Journal of Organizational Behavior, 23*, 795–814.

Moon H. (2001). The two faces of conscientiousness: Duty and achievement striving in escalation of commitment dilemmas. *Journal of Applied Psychology, 86*, 533–540.

*Morgeson FP, Reider MH, Campion MA. (2005). Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. PERSONNEL PSYCHOLOGY, *58*, 583–611.

*Motowidlo SJ, Dunnette MD, Carter GW. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640–647.

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

Motowidlo S, Van Scotter J. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79*, 475–480.

*Mowry HW. (1957). A measure of supervisor quality. *Journal of Applied Psychology, 41*, 405–408.

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

Mumford M, Zaccaro S, Connelly M, Marks M. (2000). Leadership skills: Conclusions and future directions. *Leadership Quarterly, 11*, 155–170.

*Nguyen NT. (2004). *Response instructions and construct validity of a situational judgment test*. Proceedings of the 11th Annual Meeting of the American Society of Business and Behavioral Sciences, Las Vegas, NV.

*O'Connell MS, Doverspike D, Norris-Watts C, Hattrup K. (2001). Predictors of organizational citizenship behavior among Mexican retail salespeople. *The International Journal of Organizational Analysis, 9*, 272–280.

*O'Connell MS, McDaniel MA, Grubb WL, III, Hartman NS, Lawrence A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment, 15*, 19–29.

*Olson-Buchanan JB, Drasgow F, Moberg PJ, Mead AD, Keenan PA, Donovan MA. (1998). Interactive video assessment of conflict resolution skills. PERSONNEL PSYCHOLOGY, *51*, 1–24.

Overton RC. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3*, 354–379.

*Parry ME. (1968). Ability of psychologists to estimate validities of personnel tests. PERSONNEL PSYCHOLOGY, *21*, 139–147.

*Paulin C, Hanson MA. (2001, April). *Comparing the validity of rationally derived and empirically derived scoring keys for a situational judgment inventory*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Paunonen S, Rothstein M, Jackson D. (1999). Narrow reasoning about the use of broad personality measures for personnel selection. *Journal of Organizational Behavior, 20*, 389–405.

*Pereira GM, Harvey VS. (1999, April). *Situational judgment tests: Do they measure ability, personality or both?* Paper presented at the 14th Annual Conference of the Society for Industrial & Organizational Psychology, Atlanta, GA.

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

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

Ployhart R. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management, 32*(6), 868–897.

*Ployhart RE, Porr WB, Ryan AM. *Developing situational judgment tests in a service context: Exploring an alternative methodology*. Unpublished manuscript.

†Ployhart RE, Ryan AM. (2000a). A construct-oriented approach for developing situational judgment tests in a service context. Unpublished manuscript.

Ployhart RE, Ryan AM. (2000b). *Integrating personality tests with situational judgment tests for the prediction of customer service performance*. Symposium presented at the annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

*Ployhart RE, Weekley JA, Holtz BC, Kemp C. (2003). Web-based and pencil and paper testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? PERSONNEL PSYCHOLOGY, *56*, 733–752.

*Porr WB, Heffner TS. *The incremental validity of situational judgment tests for performance prediction*. Unpublished manuscript.

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

Raju NS, Burke MJ, Normand J, Langlois GM. (1991). A new meta-analysis approach. *Journal of Applied Psychology, 76*, 432–446.

Roth P, Bobko P, McFarland L, Buster M. (2008). Work sample tests in personnel selection: A meta-analysis of black-white differences in overall and exercise scores. PERSONNEL PSYCHOLOGY, *61*, 637–661.

Rotundo M, Sackett P. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach. *Journal of Applied Psychology, 87*, 66–80.

*Rusmore JT. (1958). A note on the "test of practical judgment." PERSONNEL PSYCHOLOGY, *11*, 37.

*Sacco JM, Scheu CR, Ryan AM, Schmitt N, Schmidt DB, Rogg KL. (2000, April). *Reading level and verbal test scores as predictors of subgroup differences and validities of situational judgment tests*. Paper presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Sackett PR, Schmitt N, Ellingson JE, Kabin MB. (2001). High stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist, 56*, 302–318.

*Sartain AQ. (1946). Relation between scores on certain standard tests and supervisory success in an aircraft factory. *Journal of Applied Psychology, 29*, 328–332.

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

Schmidt FL, Hunter JE, Outerbridge AN. (1986). Impact of job experience, work sample, and supervisory ratings of job performance. *Journal of Applied Psychology, 71*, 432–439.

Schmidt F, Viswesvaran C, Ones D. (2000). Reliability is not validity and validity is not reliability. PERSONNEL PSYCHOLOGY, *53*, 901–912.

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

Schmitt N, Clause CS, Pulakos ED. (1996). Subgroup differences associated with different measures of some common job-relevant constructs. *International Review of Industrial and Organizational Psychology, 77*, 116–139.

Scullen SE, Mount MK, Judge TA. (2003). Evidence of the construct validity of developmental ratings of managerial performance. *Journal of Applied Psychology, 88*, 50–66.

*Smirdele D, Perry BA, Cronshaw SF. (1994). Evaluation of video-based assessment in transit operator selection. *Journal of Business and Psychology, 6*, 156–193.

*Smith KC, McDaniel MA. (1998, April). *Criterion and construct validity evidence for a situational judgment measure*. Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Smither JW, Reilly RR, Millsap RE, Pearlman K, Stoffey RW. (1993). Applicant reactions to selection procedures. PERSONNEL PSYCHOLOGY, *46*, 49–76.

*Spitzer ME, McNamara WJ. (1964). A managerial selection study. PERSONNEL PSYCHOLOGY, 19–40.

Sternberg RJ, Wagner RK. (1993). The *g*-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science, 2*, 1–5.

*Sternberg RJ, Wagner RK, Okagaski L. (1993). Practical intelligence: The nature and role of tacit knowledge in work and at school. In Puckett JM, Reese HW (Eds.), *Mechanisms of everyday cognition* (pp. 205–227). Hillsdale, NJ: Erlbaum.

*Stevens MJ, Campion MA. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management, 25*, 207–228.

*Swander CJ. (2000). *Video-based situational judgment test characteristics: Multidimensionality at the item level and impact of situational variables*. Unpublished dissertation.

*Swander CJ. (2001, April). *Exploring the criterion validity of two alternate forms of a situational judgment test*. Paper presented at the 16th Annual Conference of the Society for Industrial & Organizational Psychology: San Diego, CA.

Tett R, Jackson D, Rothstein M, Reddon J. (1999). Meta-analysis of bidirectional relations in personality-job performance research. *Human Performance, 12*, 1–29.

*Thumin FJ, Page DS. (1966). A comparative study of two test of supervisory knowledge. *Psychological Reports, 18*, 535–538.

*Timmreck CW. (1981). Moderating effect of tasks on the validity of selection tests (Doctoral dissertation, University of Houston, 1981). *Dissertation Abstracts International*, 42(3-B), 1221.

Van Scotter JR, Motowidlo SJ. (1996). Evidence for two factors of contextual performance: Job dedication and interpersonal facilitation. *Journal of Applied Psychology, 81*, 525–531.

Viswesvaran C. (2001). Learning from the past and mapping a new terrain assessment of individual performance. In Anderson N, Ones DS, Sinangil HK, Viswesvaran C (Eds.) *Handbook of industrial, work and organizational psychology: Personnel psychology* (Vol. 1) London: Sage.

Viswesvaran C, Ones DS. (2000). Perspectives on models of job performance. *International Journal of Selection and Assessment, 8*, 216–226.

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

*Wagner RK, Sternberg RJ. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49*, 436–458.

*Waugh G. (2002, April). *NCO21 Situational judgment test: Selecting response options and items for a situational judgment test*. Paper presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto.

*Weekley JA, Jones C. (1997). Video-based situational testing. PERSONNEL PSYCHOLOGY, *50*, 25–49.
*Weekley JA, Jones C. (1999). Further studies of situational tests. PERSONNEL PSYCHOLOGY, *52*, 679–700.
*Weekley JA, Ployhart RE. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance, 18*, 81–104.
*Weekley JA, Ployhart RE, Harold CM. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement, and subgroup differences. *Human Performance, 17*, 433–461.
*Weitz J, Nuckols RC. (1953). A validation study of how supervise? *Journal of Applied Psychology, 36*, 301–303.
Wernimont PF, Campbell JP. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372–376.
Whetzel D, McDaniel M, Nguyen N. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance, 21*, 291–309.
*Wiener DN. (1961). Evaluation of selection procedures for a management development program. *Journal of Consulting Psychology, 8*, 121–128.
Witt LA, Ferris GR. (2003). Social skill as a moderator of the conscientiousness-performance relationship: Convergent results across four studies. *Journal of Applied Psychology, 88*, 809–820.
*Wolf MB, McClellan. (2008, April). *Do respondents perceive a difference between SJT response instructions?* Poster presented at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.