PERSONNEL PSYCHOLOGY
2007, 60, 63–91

SITUATIONAL JUDGMENT TESTS, RESPONSE INSTRUCTIONS, AND VALIDITY: A META-ANALYSIS

MICHAEL A. McDANIEL
Virginia Commonwealth University

NATHAN S. HARTMAN
John Carroll University

DEBORAH L. WHETZEL
U.S. Postal Service

W. LEE GRUBB III
East Carolina University

Situational judgment tests (SJTs) are personnel selection instruments that present job applicants with work-related situations and possible responses to the situations. There are typically 2 types of instructions: behavioral tendency and knowledge. Behavioral tendency instructions ask respondents to identify how they would likely behave in a given situation. Knowledge instructions ask respondents to evaluate the effectiveness of possible responses to a given situation. Results showed that response instructions influenced the constructs measured by the tests. Tests with knowledge instructions had higher correlations with cognitive ability. Tests with behavioral tendency instructions showed higher correlations with personality constructs. Results also showed that response instructions had little moderating effect on criterion-related validity. Supplemental analyses showed that the moderating effect of response instructions on construct validity was not due to systematic differences in item content. SJTs have incremental validity over cognitive ability, the Big 5, and over a composite of cognitive ability and the Big 5.

Research on SJTs for employee selection has increased dramatically in recent years (Weekley & Ployhart, 2006). SJTs present applicants with work-related situations and possible responses to the situations. The criterion-related validity of SJTs has been evaluated in several primary studies (Chan & Schmitt, 2002; Motowidlo, Dunnette, & Carter, 1990;

This paper has benefited from comments by Paul Sackett, Robert Ployhart, David Chan, and Sheldon Zedeck. The authors appreciate the data received from many colleagues and are particularly appreciative of the large amount of data obtained from Jeff Weekley. The authors appreciate the assistance of colleagues who completed the survey on response instructions.

Correspondence and requests for reprints should be addressed to Michael McDaniel, Virginia Commonwealth University, P.O. Box 844000 Richmond, VA 23284-4000; mamcdani@vcu.edu.

Smith & McDaniel, 1998) and in a meta-analysis (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). The construct validity of SJTs also has been assessed in several primary studies (Vasilopoulos, Reilly, & Leaman, 2000; Weekley & Jones, 1997, 1999). Researchers contend that SJTs predict performance because they measure job knowledge (Motowidlo, Borman, & Schmit, 1997), practical intelligence (Sternberg, Wagner, & Okagaki, 1993), or general cognitive ability (McDaniel et al., 2001). Meta-analyses of the construct validity data show that SJTs are correlated with Conscientiousness, Emotional Stability, Agreeableness (McDaniel & Nguyen, 2001), and cognitive ability (McDaniel et al., 2001). However, the magnitude of the cognitive and noncognitive correlates of SJTs, as well as their criterion-related validities, vary substantially, suggesting the existence of one or more moderators. The current study evaluates whether SJT response instructions operate as a moderator of the construct and criterion-related validities of SJTs.

McDaniel and Nguyen (2001) identified two categories of SJT response instructions: knowledge and behavioral tendency. Knowledge response instructions ask respondents to select the correct or best possible response or to rate the effectiveness of responses. Behavioral tendency response instructions ask the respondent to select the response that represents what the respondent would likely do or to rate the likelihood that they would perform an action. McDaniel and Nguyen (2001) hypothesized that response instructions may influence construct and criterion-related validity.

The distinction between the two instruction types is relevant to the concept of typical versus maximal performance, which was first addressed by Cronbach (1949, 1984) . In maximal performance tests, one assesses how a respondent performs when doing their best. From a maximal performance test, one can make inferences about ability. Cronbach (1984) offered personality and interest tests as measures of typical performance. Subsequent research expanded the construct to job performance (DuBois, Sackett, Zedeck, & Fogli, 1993; Ployhart, Lim, & Chan, 2001; Sackett, Zedeck, & Fogli, 1988). In the job performance literature, maximal performance is viewed as more heavily dependent on cognitive skills than on typical performance. Predictors such as cognitive ability, job knowledge tests, and work sample tests assess maximal performance in that respondents are motivated to display their knowledge or abilities accurately. SJTs with knowledge instructions are also maximal performance measures because respondents make judgments about what constitutes effective, or maximal, performance. All maximal performance measures should have substantial cognitive correlates.

Other predictors, such as personality tests and SJTs with behavioral tendency instructions, ask respondents to report how they typically behave.

Relative to measures of maximal performance, these predictors can be expected to have smaller cognitive correlates and larger noncognitive correlates.

The distinction between maximal versus typical predictors is more than a cognitive–noncognitive distinction. The two types of predictors are differentially susceptible to distortions associated with self-deception and impression management (Palhaus, 1984). When using a predictor that requests self-reports of typical behavior, respondents with tendencies toward self-deception might report that they typically behave in an agreeable way at work even when their typical behavior is widely known to be abrasive. Likewise, by consciously engaging in impression management, respondents who typically behave in an unethical manner would respond that they behave ethically. In contrast, self-report predictors of knowledge are not subject to distortion by self-deception or impression management. Thus, if asked for the cube root of 27, mathematically challenged respondents might guess, but respondents' tendency toward self-deception or impression management does not distort their response.

The present study examined response instructions as a moderator of the construct and criterion-related validity of SJTs through a meta-analysis. We also present analyses evaluating alternative explanations to the response instruction findings. Specifically, we consider whether SJT content, a large sample outlier, and data source might explain the effect we attribute to response instructions.

Hypotheses

Construct validity. The type of response instruction is expected to affect the degree to which various constructs are correlated with the SJT. Specifically, SJTs that use knowledge instructions should have higher positive correlations with cognitive ability than SJTs that use behavioral tendency response instructions for two reasons. First, both assessments of knowledge and cognitive ability are requests for maximal performance. In SJTs with knowledge instructions, respondents are asked to judge the effectiveness of various actions and thus are judgments of what constitutes maximal performance. In cognitive ability tests, respondents are asked to identify the best (the maximally correct) answer to an item. Second, both knowledge and cognitive ability are cognitive constructs. Conversely, SJTs that use behavioral tendency instructions should have higher positive correlations with personality tests for three reasons. First, both assessments are requests for typical behavior. Second, both assessments can be expected to tap noncognitive constructs. Third, both assessments can be influenced by self-deception and impression management errors. Based on this discussion, we offer two hypotheses:

Hypothesis 1: SJTs with knowledge instructions will have higher correlations with cognitive ability than SJTs with behavioral tendency instructions.

Hypothesis 2: SJTs with behavioral tendency instructions will have higher correlations with personality tests than SJTs with knowledge instructions.

Criterion-related validity. There are four reasons why SJTs with knowledge response instructions should yield higher validities for predicting job performance than SJTs with behavioral tendency response instructions. First, knowledge measures are excellent predictors of job performance (Dye, Reck, & McDaniel, 1993; Schmidt & Hunter, 1998). Second, if our construct hypotheses are correct, a knowledge instruction measure will be more highly correlated with cognitive ability. Because cognitive ability tests (Schmidt & Hunter, 1998) have substantial criterion-related validity, it is reasonable to suggest that SJTs with high cognitive loadings will have higher validity than SJTs with lower cognitive loadings. Third, the validity of measures that tap multiple constructs, such as SJTs and interviews (Huffcutt, Roth, & McDaniel, 1996), are likely to increase as their cognitive load increases. Fourth, knowledge instruction SJTs are not subject to self-deception and impression management errors because they are not self-reports of typical behavior. Therefore, we offer the following hypothesis:

Hypothesis 3: SJTs with knowledge instructions will have higher criterion-related validity than SJTs with behavioral tendency instructions.

Incremental validity. Most personnel selection systems rely on more than one selection procedure. Thus, the incremental validity of the SJT is an important topic. Often selection systems include measures of cognitive ability due to their high validity for all jobs (Schmidt & Hunter, 1998) and the value of a SJT increases if it has incremental validity above and beyond cognitive ability. The incremental validity of SJTs over measures of general cognitive ability has been a topic of several studies (Clevenger, Pereira, Wiechmann, Schmitt, Schmidt Harvey, 2001; Chan & Schmitt, 2002; McDaniel et al., 2001; Weekley & Jones, 1997, 1999). The current paper will examine whether response instructions moderate incremental validity. The paper will also extend the incremental validity analyses to include the Big Five.

The incremental validity of a test is the extent to which the test can explain unique variance in job performance not explained by other measures in the battery. To the extent that SJTs are correlated with cognitive ability, the potential of SJTs to predict unique variance beyond cognitive ability is limited. Likewise, to the extent that SJTs are correlated with the Big Five, the potential of SJTs to predict unique variance beyond the Big Five is limited. However, several studies have reported incremental validity of

a SJT over a battery containing cognitive ability and personality (Chan & Schmitt, 2002; Clevenger et al., 2001; O'Connell, Hartman, McDaniel, Grubb, & Lawrence, in press; Weekley & Ployhart 2005). If our construct hypotheses are correct, SJTs with behavioral tendency instructions should be less highly correlated with cognitive ability than SJTs with knowledge instructions. The opposite should be true with personality measures. Therefore, we offer the following two hypotheses:

Hypothesis 4: SJTs with behavioral tendency instructions will have larger incremental validity over cognitive ability than SJTs with knowledge instructions.

Hypothesis 5: SJTs with knowledge instructions will have larger incremental validity over the Big Five than SJTs with behavioral tendency instructions.

Concerning the incremental validity of a SJT over a battery of a cognitive test and the Big Five, there is no compelling reason why response instructions would moderate incremental validity of an SJT. The cognitive correlations of SJTs with knowledge instruction will impair the ability of such a SJT to add incremental validity over cognitive ability. The personality correlations of SJTs with behavioral tendency instructions will impair the ability of such SJTs to add incremental validity over personality. Thus, the incremental validity of SJTs, regardless of instruction type, is impaired by other predictors sharing the same construct space. Therefore, we offer the following hypothesis:

Hypothesis 6: SJTs will have incremental variance over a composite of cognitive ability and the Big Five, but the relationship will not be meaningfully moderated by response instructions.

Method

Literature Search

The data set from the McDaniel et al. (2001) meta-analysis provided this study with a large amount of its criterion-related validity data and construct validity data concerning cognitive ability. The McDaniel and Nguyen (2001) data set examining construct validity of the Big Five provided the majority of the data for the Big Five construct validity analysis. We also obtained a large number of additional studies, both published and unpublished, conducted since 2001. These new studies added criterion-related validity coefficients as well as construct validity data. The reference lists developed from these studies were updated to represent a comprehensive list of studies using SJTs in the personnel selection literature. We solicited additional studies from researchers and practitioners working in this

area and reviewed recent journals and the programs of recent conferences for additional studies.

Analysis

In their 2001 meta-analysis of the criterion-related validity of SJTs, McDaniel et al. (2001) corrected the variance estimate for differences across studies in the reliability of SJTs. The available reliability estimates were all coefficient alpha. We argue that this was an inappropriate estimate of the reliability of SJTs because such tests are seldom homogeneous. Specifically, there are no published factor analyses of SJTs in personnel selection with interpretable factors (McDaniel & Whetzel, 2005). Therefore, we do not correct the variance of the correlations for differences across studies due to the reliability of SJTs. Note that this decision has no impact on the estimated mean validities because means are not corrected for measurement error in meta-analyses of criterion-related validities.

The psychometric artifact-distribution meta-analytic method was used in this study (Hunter & Schmidt, 1990, 2004). For the construct validity analyses, the estimated population mean correlation was corrected for measurement error in the construct (the measure of cognitive ability or personality) but not for measurement error in the SJT. The estimated population variance was corrected for sampling error and for differences across studies in measurement error in the construct.

Range restriction can be a problem in interpreting the incremental validity results. For example, if the incumbents in the validity studies have been selected on cognitive ability, cognitive ability would have very restricted variance permitting an SJT to show incremental validity. However, no range restriction corrections were conducted because we have little data on the degree of range restriction in the SJTs in our samples.

For the criterion-related validity analyses, the estimated population mean correlation was corrected for measurement error in the criterion. The estimated population variance was corrected for sampling error and for differences across studies in measurement error in the criterion. Again, no range restriction corrections were conducted and no corrections were made for differences across studies in the reliability of SJTs.

Reliability

The reliability artifact distributions for the job performance criterion and for cognitive ability were drawn from McDaniel et al. (2001). The reliability distribution for the personality constructs were drawn from studies in the meta-analysis and are presented in Table 1.

TABLE 1
Big Five Reliability Distribution

Reliability	Frequency
0.63	2
0.66	2
0.67	1
0.69	1
0.70	1
0.72	1
0.73	1
0.74	2
0.76	6
0.77	3
0.78	3
0.79	2
0.80	3
0.81	3
0.82	2
0.83	1
0.84	3
0.85	1
0.86	1
0.87	1
0.88	2

Response Instruction Taxonomy

A taxonomy of response instruction types was developed (Table 2). The taxonomy is hierarchical. At the highest level of the taxonomy, we distinguish between knowledge and behavioral tendency instructions. At the lowest level of the taxonomy, we show variations in knowledge instructions and variations in behavioral tendency instructions.

Decision Rules

Analysis of the criterion-related validity data generally followed the four main decision rules used in the original McDaniel et al. (2001) meta-analysis. First, only studies whose participants were employees or applicants were used. Second, SJTs were required to be in the paper-and-pencil format. Other formats, such as interview (e.g., Latham, Saari, Pursell, & Campion, 1980) and video-based SJTs (e.g., Dalessio, 1994; Jones & DeCotiis, 1986), were not included. The third rule gave priority to supervisor ratings of job performance over other available measures. The fourth and final rule defined boundary conditions on the job performance measure. Surrogate performance measures such as years of school completed,

TABLE 2
Response Instruction Taxonomy

Response instruction	Illustrative study
Behavioral tendency instructions	
1. Would do	Bruce (1953)
2. Most & least likely	Pulakos & Schmitt (1996)
3. Rate and rank what you would most likely do	Jagmin (1985)
4. Rate the tendency to perform on a Likert scale	Doherty (personal communication, July 7, 2005)
Knowledge instructions	
5. Should do	Hanson (1994)
6. Best response	Greenberg (1963)
7. Best & worst	Clevenger & Haaland (2000)
8. Best & second best	Richardson, Bellows, & Henry Co. (undated)
8. Best & second best	
9. Rate effectiveness	Chan & Schmitt (1997)
10. Best, second, & third best	Cardell (1942)
11. Level of importance	Corts (1980)

hierarchical level in the organization, number of employees supervised, years of management experience, ratings of potential, and salary were not used. Thus, data from Ployhart and Ehrhart (2003) were not included because the respondents were not employees or applicants, and the criterion was grade point average. In addition, for the criterion-related validity data, we permitted only one correlation per sample per response instruction to be used.

The rules concerning construct-related data generally followed the decision rules used in McDaniel and Nguyen (2001). McDaniel and Nguyen (2001) included data from any personnel selection SJT that reported a correlation with a test that could be classified as a Big Five measure. No restrictions were placed on the sample members (the sample members could be students, applicants, or employees). If a sample reported more than two correlations between a SJT and a Big Five measure, McDaniel and Nguyen (2001) included both correlations. Thus, if a sample provided a separate correlation between a SJT and each of two measures of Conscientiousness, both correlations were included in the analysis of the SJT with Conscientiousness. However, in the present study, for the construct validity data, we permitted only one correlation per construct per sample per response instruction to be used.

In two samples, the respondents were administered the same SJT twice, once with a knowledge instruction and once with a behavioral tendency instruction. These samples contributed two correlations per analysis. For the 118 criterion-related validity analyses, 116 of the correlations are from

independent samples. The remaining two are from a single sample (Doherty, personal communication, July 7, 2005, Sample 3) where in one sample, the same SJT was administered with both knowledge instructions and behavioral tendency instructions, and the validity was reported separately by response instruction. For the correlations between a SJT and a construct, all but two of the correlations in each distribution came from a separate sample. The remaining two were from Nguyen, Biderman, & McDaniel (2005), where the same SJT was administered with both knowledge instructions and behavioral tendency instructions to the same sample.

Another improvement over previous research involves the exclusion of various studies. We made a few changes to the decision rules used by McDaniel et al. (2001) and McDaniel and Nguyen (2001). First, we did not include studies using the *How Supervise?* test (File, 1943) in either the criterion-related and construct validity analysis. The *How Supervise?* test presents respondents with single statements about supervisory practices and asks whether they agree with the statement. *How Supervise?* items lack both a situation and a series of responses to the situations. We also excluded studies in which SJT response instructions could not be coded due to lack of information or due to the response instructions not fitting into the taxonomy (Jones, Dwight, & Nouryan, 1999).

Publication bias analyses. McDaniel, Rothstein and Whetzel (2006) and others (Duval, 2005; McDaniel, Hurtz & Donovan, 2006; McDaniel, McKay & Rothstein, 2006; Whetzel, 2006) have raised the issue of publication bias in meta-analyses within industrial and organizational psychology. Publication bias exists to the extent that not all studies are available to the meta-analytic researcher, and the missing studies have systematically different results from those that are available. Because publication bias has been found in some industrial and organizational research areas, and because publication bias analyses are required in meta-analyses published in *Psychological Bulletin* (Cooper, 2003), we applied these methods to our data.

Results

Our results are divided into four sections. First, we present the meta-analysis results addressing the response instruction moderator (Hypotheses 1, 2, and 3). Second, we present meta-analysis results for three alternative hypotheses to the response instruction moderator findings. Third, we describe the method and results of a primary study evaluating an alternative hypothesis to the response instruction moderator findings. Finally, we present the results of the incremental validity analyses addressing Hypotheses 4, 5, and 6.

TABLE 3
Meta-Analytic Results: Construct and Criterion-Related Validity

	Observed distribution				Population distribution			
Distribution	N	No. of rs	Mean r	σ	ρ	σ_ρ	% of σ^2_ρ due to artifacts	80% CI
Agreeableness	25,473	51	.22	.15	.25	.17	8	.03 to .47
Knowledge	17,115	34	.17	.10	.19	.11	16	.05 to .34
Behavioral tendency	8,358	17	.33	.18	.37	.20	6	.12 to .62
Conscientiousness	31,277	53	.23	.13	.27	.14	10	.09 to .44
Knowledge	23,043	38	.21	.10	.24	.10	15	.11 to .37
Behavioral tendency	8,234	15	.30	.16	.34	.18	6	.11 to .57
Emotional stability	19,325	49	.19	.17	.22	.18	6	−.02 to .45
Knowledge	11,067	33	.10	.09	.12	.08	34	.01 to .22
Behavioral tendency	8,258	16	.31	.18	.35	.20	6	.10 to .60
Extraversion	11,351	25	.13	.09	.14	.09	25	.03 to .26
Knowledge	9,533	14	.14	.08	.15	.08	23	.06 to .25
Behavioral tendency	1,818	11	.07	.13	.08	.12	35	−.07 to .23
Openness to Experience	4,515	19	.11	.08	.13	.05	66	.06 to .19
Knowledge	2,921	11	.12	.08	.14	.06	56	.06 to .22
Behavioral tendency	1,594	8	.09	.07	.11	.01	98	.09 to .12
Cognitive ability	30,859	95	.29	.18	.32	.19	7	.08 to .57
Knowledge	24,656	69	.32	.17	.35	.19	6	.11 to .60
Behavioral tendency	6,203	26	.17	.13	.19	.13	20	.02 to .36
Job performance	24,756	118	.20	.10	.26	.10	38	.13 to .39
Knowledge	22,050	96	.20	.10	.26	.10	38	.13 to .38
Behavioral tendency	2,706	22	.20	.13	.26	.13	36	.08 to .43

Notes. Estimated mean population correlations for construct validity analyses are corrected for measurement error in the personality and cognitive ability measures. Estimated mean population correlations for criterion-related validity are corrected for measurement error in job performance. Estimated population variance estimates for all analyses are corrected for differences across studies in measurement error in the Big Five (for the Big Five construct analyses), in cognitive ability (for the cognitive ability constructs analyses), in job performance (for the criterion-related validity analyses).

Meta-Analytic Results for Response Instruction Moderator

Table 3 presents the construct and criterion-related validity results. These results address Hypothesis 1 (correlations with cognitive ability), Hypothesis 2 (correlations with the Big Five), and Hypothesis 3 (criterion-related validity). The first column identifies the distribution of validities analyzed. Total sample size across studies, the total number of correlation coefficients, and the mean and standard deviation of the observed distribution are presented in Columns 2 through 5. Columns 6 and 7 contain estimates of the population mean correlations and standard deviations. The percentage of variance in the observed distribution corrected for sampling error and reliability differences across studies and the 80% credibility

interval for the true validities are presented in the remaining two columns. Because no corrections for range restriction were conducted, all reported populations correlations are likely to be downwardly biased (underestimates of the actual population value). For the analyses addressing cognitive ability correlates, the population correlation distribution is corrected for measurement error in the cognitive ability measures. Because no corrections in the population variance were made for differences across studies in range restriction or differences across studies in measurement error in the SJTs, the reported population variances are likely to be upwardly biased (overestimates of the actual population variance).

Results relevant to Hypothesis 1 concerning the correlation between SJTs and cognitive ability are shown in Table 3. The estimated population correlation is .32 for all SJTs, .35 for knowledge instruction SJTs, and .19 for behavioral tendency SJTs. Thus, Hypothesis 1 is supported. We note that both the wide 80% credibility intervals and the low percentage of population distribution variance due to artifacts suggest substantial variation in the population distributions. Although SJTs on average show correlations with general mental ability and the magnitude of the correlation is moderated by response instructions, there is substantial variability across studies.

Concerning Hypothesis 2, the estimated mean population correlations between SJTs and the Big Five were .25 for Agreeableness, .27 for Conscientiousness, .22 for Emotional Stability, .14 for Extraversion, and .13 for Openness to Experience. For three of the Big Five, the correlations between the SJT and the personality trait were higher for the behavioral tendency than for the knowledge instruction set: Agreeableness (.37 vs .19), Conscientiousness (.34 vs. 24), and Emotional Stability (.35 vs .12). Thus, Hypothesis 2 is supported for the personality traits of Agreeableness, Conscientiousness, and Emotional Stability. Hypothesis 2 is not supported for Extraversion and Openness to Experience. We note that the distributions (Extraversion and Openness to Experience) that did not support Hypothesis 2 had the fewest number of coefficients.

Concerning criterion-related validity, the estimated population correlation of .26 is somewhat lower than the .34 coefficient reported in McDaniel et al. (2001). This difference is due to the addition of newly found studies (primarily unpublished) that, on average, have lower criterion-related validities than the mean of studies reported in McDaniel et al. (2001). The validities in this study were the same for both instruction types ($r =$ .26), thus, Hypothesis 3 is not supported. The lower 90th percentile values for the validities, used for inferences regarding validity generalizations, are .13 for all SJTs, .13 for SJTs with knowledge instructions, and .08 for SJTs with behavioral tendency instructions. These non-zero, positive correlations indicate that the validity of SJTs generalize.

TABLE 4
Meta-Analytic Results Relevant to Alternative Hypotheses: Outliers and Data Sources

	Observed distribution				Population distribution			
Distribution	N	No. of rs	Mean r	σ	p	σ_p	% of σ_p^2 due to artifacts	80% CI
Large sample outlier analyses SJT with cognitive ability								
SJT (knowledge instructions) with cognitive ability. All coefficients excluding Pereira & Schmidt Harvey's (1999) Study 2	19,070	68	.37	.17	.41	.18	9	.19 to .64
Criterion-related validity by data source								
Test vendor manuals	1,108	14	.28	.15	.36	.14	45	.18 to .54
Government technical reports	3,342	23	.30	.10	.38	.07	55	.29 to .48
Proprietary technical reports	10,524	28	.16	.06	.20	.02	78	.17 to .23
Journal articles	4,453	31	.22	.12	.28	.12	37	.12 to .44
Journal articles – Knowledge instructions	3,591	23	.21	.12	.27	.12	37	.11 to .42
Journal articles – Behavioral tendency instruction	862	8	.25	.13	.33	.13	38	.16 to .49
Book chapter	151	4	.42	.14	.55	.04	81	.50 to .60
Conference papers	3,692	15	.16	.07	.21	.05	70	.15 to .27
Conference papers – Knowledge instructions	2,170	7	.16	.04	.20	.00	166	.20 to .20
Conference papers – Behavioral tendency instructions	1,522	8	.16	.10	.21	.09	.09	.09 to .33
Dissertations and master's theses	1,486	3	.24	.06	.31	.05	37	.25 to .38

Note. Table 3's note applicable to Table 4.

Meta-Analytic Results of Alternative Hypotheses to the Response Instruction Moderator

Tables 4 and 5 offer meta-analyses addressing three alternative explanations to our conclusions regarding the moderating effect of response instruction on construct and criterion-related validity. The first alternative

TABLE 5

Meta-Analytic Results Relevant to Alternative Hypotheses: Test Content is Held Constant

	Observed distribution				Population distribution			
Distribution	N	No. of *r*s	Mean *r*	σ	ρ	σ_p	% of σ_p^2 due to artifacts	80% CI
Analyses where test content is held constant								
Agreeableness	1,465	8	.15	.04	.17	.00	100	.17 to .17
Knowledge	763	4	.12	.04	.14	.00	100	.14 to .14
Behavioral tendency	702	4	.17	.04	.20	.00	100	.20 to .20
Conscientiousness	1,465	8	.24	.09	.27	.07	55	.18 to .36
Knowledge	763	4	.19	.10	.21	.07	52	.12 to .31
Behavioral tendency	702	4	.29	.05	.33	.00	100	.33 to .33
Emotional stability	1,465	8	.06	.08	.07	.04	82	.02 to .12
Knowledge	763	4	.02	.03	.02	.00	100	.02 to .02
Behavioral tendency	702	4	.11	.09	.13	.06	63	.04 to .21
Extroversion	1,465	8	.04	.06	.04	.00	100	.04 to .04
Knowledge	763	4	.02	.04	.02	.00	100	.02 to .02
Behavioral tendency	702	4	.06	.07	.07	.00	100	.07 to .07
Openness to experience	1,465	8	.07	.09	.08	.06	62	.00 to .16
Knowledge	763	4	.05	.07	.05	.01	99	.04 to .07
Behavioral tendency	702	4	.09	.10	.10	.08	49	−.01 to .21
Cognitive Ability	1,497	8	.20	.13	.22	.13	26	.06 to .39
Knowledge	737	4	.25	.15	.28	.15	19	.08 to .47
Behavioral tendency	760	4	.15	.09	.17	.07	58	.08 to .25
Job Performance	631	6	.15	.08	.20	.00	100	.20 to .20
Knowledge	341	3	.20	.07	.26	.00	100	.26 to .26
Behavioral tendency	290	3	.09	.04	.12	.00	100	.12 to .12

Note. Table 3's note is applicable to Table 5.

hypothesis concerns large sample outliers and holds that some studies with large samples distort the findings. Pereira and Schmidt Harvey (1999, Study 2) reported a very low correlation between an SJT with knowledge instructions and cognitive ability ($r = .14$) based on a sample of 5,586. Table 4 shows that when this correlation is dropped from the analysis, the population correlation between knowledge instruction SJTs and cognitive ability increased from .35 (Table 3) to .41 (Table 4). Thus, the large sample outlier results mitigated support for Hypothesis 1 and discredits the first alternative hypothesis.

The second alternative hypothesis holds that publication bias distorts the conclusions regarding the response instruction moderator. Table 4 shows that criterion-related validity varies by the source of the data. This analysis was conducted because recent research has shown data source differences can be symptomatic of publication bias (McDaniel, McKay &

Rothstein, 2006; McDaniel, Rothstein & Whetzel, 2006). We conducted trim and fill publication bias analyses (Duval & Tweedie, 2000a, 2000b) on validity distributions subset by source of data. Publication bias was not a compelling explanation for validity differences by data source (results available from the first author). Also, within data source there were too few validity coefficients for behavioral tendency instruction SJTs to provide a meaningful replication by the response moderator analyses within data source. Thus, although there are some differences in validity by publication source, there are insufficient data to determine whether a publication source moderator has any effect on our conclusions regarding a response instruction moderator.

The third alternative hypothesis to our conclusion regarding the response instruction moderator concerns the content of the SJTs. A meta-analysis, such as the current study, which compares the criterion-related and construct validity of SJTs with different response instructions, cannot definitively state that any observed differences are due to the response instructions because the content of the SJTs is not held constant. A rival hypothesis is that SJTs with knowledge instructions have systematically different content than SJTs with behavioral tendency instructions and that it is the content difference and not the response instruction difference that causes any observed moderating effect.

Table 5 addresses the third alternative hypothesis by summarizing the construct and criterion-related validity data for studies in which a SJT was administered twice, once with knowledge instructions and once with behavioral tendency instructions. These types of studies are ideal for evaluating a response instruction moderator because the content of the SJTs is held constant. Eight studies were available for the construct hypotheses (Hypotheses 1 and 2), and six studies were available for the criterion-related validity hypothesis (Hypothesis 3). The analyses for cognitive ability correlates support Hypothesis 1 because the SJTs with knowledge instructions have substantially larger correlations with cognitive ability than the same SJTs administered with behavioral tendency instructions (.28 vs. .17). Related to Hypothesis 2, the SJTs with behavioral tendency instructions have larger correlations with the Big Five than the same SJTs administered with knowledge behavioral tendency instructions for Agreeableness (.20 vs. .14), for Conscientiousness (.33 vs. .21), for Emotional Stability (.13 vs. .02), for Extraversion (.07 vs. .02), and for Openness to Experience (.10 vs. .05).

These supplemental construct analyses are supportive of Hypotheses 1 and 2. In contrast to analyses in Table 3 for the full data set in which Hypothesis 2 was confirmed for only three of the Big Five, Table 5 shows that when SJT content is held constant, Hypothesis 2 is confirmed for all of the Big Five constructs. However, when comparing Table 5 with

Table 3, the magnitude of the differences between the two response instructions is smaller when the SJT content is held constant for Agreeableness (.06 difference vs. .18 difference) and Emotional Stability (.11 difference vs. .23 difference). Because there are only eight studies in which the SJT content was held constant, one should be cautious in drawing conclusions. With these caveats in mind, we conclude that the construct validity response instruction moderator is not actually due to a content artifact. We also recommend that this conclusion be reevaluated as more studies controlling for SJT content become available.

The criterion-related validity results in Table 5 differ from those in Table 3. Table 3 showed no evidence for a response instruction moderator of criterion-related validity, both estimated populations validities were .26. In Table 5, where SJT content was held constant, a response instruction moderator favoring knowledge instruction SJTs is evident. The estimated population criterion-related validity falls from .26 to .12 when the content of the SJT is held constant. However, the validity of .12 is based on only three correlations with a total sample size of 290. One can either base one's conclusions on a large amount of data (Table 3, 118 samples) or on a very small of data from the "best" studies (Table 5, 3 samples). We chose to rely on the large amount of data (Table 3) and conclude that knowledge instruction and behavioral tendency instructions SJTs yield the same criterion-related validity (.26). We recommend that this conclusion be reevaluated as more studies controlling SJT content become available.

Method and Results for a Primary Study Addressing an Alternative Hypothesis to the Response Instruction Moderator

If the content differences across SJTs are responsible for the construct validity differences we attribute to response instruction, then those content differences should be readily observable. Specifically, those with expertise in SJTs should be able to identify the response instructions associated with an item based on reviewing its content. To test this alternative hypothesis, we created a Web-administered survey to identify the extent to which SJT subject matter experts could correctly identify whether a situational judgment item was from a test using knowledge instructions or a test using behavioral tendency instructions. We selected 30 items from SJTs, 15 were from SJTs that used knowledge response instructions and 15 were from SJTs that used behavioral tendency response instructions. If needed, the items were edited to remove the response instruction. For example, if a situational stem ended with "What would you do?" or "Pick the best response," we deleted the phrase. These items were assembled into a survey. The survey explained the distinction between knowledge and behavioral tendency instructions and asked the respondents to judge

whether each item was drawn from a SJT that used knowledge instructions or one that used behavioral tendency instructions.

Based on an initial administration of the survey, we discovered that a subset of the respondents could score well above chance by adopting a strategy of assigning items written with second person pronouns (Your boss is angry with you) to the behavioral tendency category and items written with third person pronouns (An employee's boss is angry with him) to the knowledge category. Because this pronoun clue was not relevant to substantive differences in content, we edited all items into the third person.

Respondents had experience with SJTs. Through e-mail, we contacted the authors of SJT research presented in conference papers, book chapters, and journal articles. We also contacted individuals at consulting firms who were known to build SJTs. Because we were concerned that we might be oversampling researchers, we also submitted an invitation to respond on the e-mail list server of the International Association of Public Personnel Management Assessment Council (IPMAAC). Twenty-three SJT subject matter experts completed the survey where all items were in the third person. The mean number correct on the survey was 17.1 (15 items would be correct with random responding). Thus, 57.1% of the items were correctly identified. We conclude that, in the absence of pronoun clues, SJT subject matter experts were unable to accurately identify the response instructions of an item based on its content at a level much greater than chance. Thus, the alternative content hypothesis as a source of the construct validity differences is not supported by these results.

Incremental Validity Results

We estimated the incremental validity of SJTs over cognitive ability, over the Big Five, and over a composite of cognitive ability and the Big Five. These results are relevant to Hypotheses 4, 5, and 6. We also examined the extent to which cognitive ability and the Big Five add incremental validity to SJTs. To build correlation matrices needed for these analyses, we needed to use other data to estimate the criterion-related validity of cognitive ability and the Big Five, and the intercorrelations among all measures. We conducted the incremental validity analyses by running hierarchical linear regressions using a correlation matrix of all variables.

For the correlations between cognitive ability and the other variables we drew data from this and other studies. We replicated the McDaniel et al. (2001) analysis using .25 as the observed validity for cognitive ability. We reported correlations between cognitive ability and SJTs with knowledge instructions with (Table 3, $r = .32$) and without the Pereira et al. (1999) outlier (Table 4, $r = .37$). This yielded two combinations

of cognitive ability correlates and performance correlates for SJTs with knowledge instructions. To obtain estimates of the correlation between cognitive ability and the Big Five, we conducted a meta-analysis of the correlations between cognitive ability and each of the Big Five based on data obtained from seven studies (Grubb, 2003; Leaman, Vasilopoulos, & Usala, 1996; Lobsenz & Morris, 1999; McDaniel, Yost, Ludwick, Hense, & Hartman, 2004; Nguyen, 2001; O'Connell, McDaniel, Grubb, Hartman, & Lawrence, 2002; Peeters & Lievens, 2005). The mean observed correlations between cognitive ability and the Big Five are .02 for Agreeableness, .05 for Conscientiousness, .03 for Emotional Stability, .06 for Extraversion, and .18 for Openness to Experience.

For analyses involving the Big Five, we needed estimates of the criterion-related validity of the Big Five and the intercorrelations among the Big Five. The criterion-related validities of the Big Five were taken from Hurtz & Donovan (2000, Table 1 observed validities, page 873). The Big Five intercorrelations were taken from Ones, Viswesvaran, & Reiss (1996, Table 6, page 669). Because the Ones et al. (1996) correlations were corrected for unreliability, we disattenuated them by multiplying the correlations by the square root of .75. That mean reliability was also drawn from Ones et al. (1996). All incremental validity analyses are based on observed correlations that are not corrected for measurement error or range restriction.

In the top section of Table 6, we show the correlation between cognitive ability (g) and the SJT used in each of the three analyses. We also show the criterion-related validity values used for g, the SJT, and the Big Five. The validity for the Big Five was estimated by entering the Big Five in a regression to predict job performance and the validity listed (.16) is the multiple R from that regression. Next, we examined the validity of various composites of g, SJT, and the Big Five. These validities are multiple Rs from regressions. The lower section of Table 6 shows incremental validities of various predictors over other predictors. Consider the first row of the incremental validity results in which the incremental validity of the SJT over cognitive ability (g) is listed as .03. That number was calculated by subtracting the validity of g alone ($r = .25$) from the multiple R where cognitive ability and SJT were optimally weighted in a regression to predict job performance. The multiple R, which is shown in the top section of Table 6, is .28. Thus, by adding a SJT to a predictor battery already containing g, the SJT incremented the uncorrected validity by .03. Three scenarios are presented. Two are for knowledge instruction SJTs, where the correlation between g and the SJT differ, and one is for behavioral tendency instruction SJTs.

In all three incremental validity scenarios, SJTs provide incremental validity over cognitive ability ranging from .03 to .05. The largest

TABLE 6

Criterion-Related and Incremental Validities for Single Predictors and Composites. Validities Are Not Corrected

Observed validity of individual predictors and composites								
Situational judgment instructions	SJT with *g*	Criterion-related validities						
		g	SJT	Big Five	*g* + SJT	Big Five + SJT	*g* + Big Five	*g* + Big Five + SJT
1. Knowledge	0.32[1]	0.25	0.20[2]	0.16	0.28	0.23	0.29	0.31
2. Knowledge	0.37[3]	0.25	0.20[2]	0.16	0.28	0.23	0.29	0.30
3. Behavioral tendency	0.17[4]	0.25	0.20[2]	0.16	0.30	0.22	0.29	0.31

Incremental validities								
	Increments over one predictor					Increments over two predictors		
Situational judgment instructions	SJT increment over *g*	Big Five increment over *g*	*g* increment over SJT	SJT increment over Big Five	Big Five increment over SJT	SJT over *g* + Big Five	Big Five over *g* + & SJT	*g* over Big Five + SJT
1. Knowledge	0.03	0.04	0.08	0.07	0.03	0.02	0.03	0.08
2. Knowledge	0.03	0.04	0.08	0.07	0.03	0.01	0.02	0.07
3. Behavioral tendency	0.05	0.04	0.10	0.06	0.02	0.02	0.01	0.09

Notes.

1. Situational judgment test correlation with cognitive ability estimate: Table 3. All tests with knowledge instructions.
2. Situational judgment test correlation with job performance estimate: Table 3. Criterion-related validity is the same for both behavioral tendency and knowledge instructions.
3. Situational judgment test correlation with cognitive ability estimate: Table 3. All tests with knowledge instructions except Pereira and Schmidt Harvey's (1999) Study 2.
4. Situational judgment test correlation with cognitive ability estimate: Table 3. All tests with behavioral tendency instructions.

There are no range restriction or measurement error corrections applied to the results in this table.

incremental validity is for SJTs with behavioral tendency instructions (.05 vs. .03). Such SJTs have the lowest correlations with *g* and thus have a higher probability of predicting over and above *g*. Although the direction of the moderating effect is consistent with Hypothesis 4, the magnitude of the moderating effect is very small and may not be meaningful.

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

In all three scenarios, SJTs offer incremental validity over a composite of *g* and the Big Five with incremental values ranging from .01 to .02. The response instruction moderator does not appear to meaningfully moderate the incremental validity thus supporting Hypothesis 6. We note that these observed incremental values are small, but few predictors offer incremental prediction over an optimally weighted composite of six variables (*g* and the Big Five).

Discussion

This paper makes several contributions to an understanding of the validity of SJTs. The first contribution is the documentation of a response instruction moderator on construct validity. Response instructions have a clear moderating effect on the correlation between *g* and SJTs. Knowledge instruction SJTs have substantially higher correlations with *g* than behavioral tendency instruction SJTs. This is true across the analysis of all correlations, when an outlier is removed, and when the analysis is restricted to those studies where the content of SJTs is held constant. Response instructions also moderate the correlations between SJT and personality. SJTs with behavioral tendency instructions have higher correlations with the Big Five than SJTs with knowledge instructions. The moderating effect is most clear for Agreeableness, Conscientious, and Emotional Stability where the moderating effect is shown for the analysis of all correlations and for analyses in which SJT content is held constant. The moderating effect for Extraversion and Openness to Experience is less clear.

The second contribution is extending theory by applying the typical versus maximal assessment distinction to understanding the SJT response instruction moderating effect. SJTs with knowledge instructions are maximal performance assessments, saturated with cognitive variance, and likely unaffected by individual differences in tendencies toward

self-deception and impression management. In contrast, SJTs with behavioral tendency instructions are typical performance assessments, saturated with noncognitive variance and subject to error associated with tendencies toward self-deception and impression management.

A third contribution is the demonstration that SJTs are unique among all other personnel assessment types in that they can serve as either assessments of typical performance or of maximal performance. We know of no other assessment method in which the construct validity of the test varies substantially as a function of the response instructions.

A fourth contribution of this research is knowledge of the extent to which response instructions moderate criterion-related validity. The bulk of our data suggest that there is no difference in the criterion-related validity of SJTs that vary in response instructions. However, the limited data on criterion-related validities when SJT content is held constant suggest that knowledge instructions SJTs have substantially higher validity than SJTs with behavioral tendency instructions. We encourage additional research on the criterion-related validity of SJTs where content is held constant. However, until such research proves otherwise, the most compelling conclusion is that response instructions do not moderate criterion-related validity.

There are at least three ways to explain the lack of response instruction moderation of criterion-related validity. First, it is possible that many respondents ignore the behavioral tendency instructions seeking reports of typical performance and instead respond to the items as if they were given knowledge instructions (e.g., respond with what they believe is the correct answer, as opposed to what they might typically do). This issue has been raised by Weekly, Ployhart, and Holtz (2006). The criterion-related validity analyses in our study are based on 114 concurrent studies and 4 predictive studies. If, in concurrent studies, respondents are ignoring behavioral tendency instructions and instead are responding as if receiving knowledge instructions, we can expect even more of this behavior in applicant samples.

A second possibility is that there are aspects of job performance that can be predicted by either personality or cognitive ability. A knowledge instruction SJT could predict X amount of criterion variance primarily through the test's cognitive loading and secondarily through the test's personality loading. A behavioral tendency instruction SJT could predict the same X amount of criterion variance primarily through the test's personality loading and secondarily through its cognitive loading. Although the tests have different weightings of cognitive and personality variance, they could account for the same amount of criterion variance.

The third possibility is that response instructions simply do not moderate the criterion-related validity of SJTs. Although the response

instructions may indicate a difference in the construct measured, the overall validity of SJTs is unchanged as a result of response instruction.

The fifth contribution of the research is that it can provide guidance on expected validities from composites of cognitive ability, personality, and SJTs. SJTs provide meaningful incremental prediction over the Big Five (observed validity increments of .06 to .07), modest incremental prediction over cognitive ability (observed validity increments of .03 to .05), and a small degree of incremental validity over a composite of cognitive ability and the Big Five (observed validity increments of .01 to .02). Our results suggest that the incremental validity of SJTs is not meaningfully moderated by response instructions. Our results suggest that to maximize criterion-related validity, one should always use a cognitive ability test. If one were to add an additional test to supplement cognitive ability, a Big Five or a SJT would provide about the same amount of incremental prediction. If the existing battery included a cognitive ability test and a Big Five test, the addition of a SJT can be expected to provide a small amount of incremental validity (observed validity increment .01 to .02). If the existing battery included a cognitive ability test and a SJT, the addition of a Big Five test can be expected to provide a small amount of incremental validity (observed validity increment .01 to .03).

We caution the reader not to over interpret the incremental validities. Cognitive ability measures might have lower or higher validities than the .25 value assumed in these analyses. Cognitive ability validity varies with job complexity (Hunter & Hunter, 1984), the degree of range restriction, and measurement error. Likewise, personality test validities may vary based on the measure used and also are influenced by range restriction and measurement error. Finally SJT correlations with job performance, cognitive ability, and the Big Five vary widely. One could clearly construct scenarios where SJTs could contribute substantially to a predictor composite or offer near zero incremental validity. We used mean values in our analyses to offer what we believe might be typical validities. We also note that the validities are uncorrected for measurement error and range restriction.

Boundary Conditions for Validity Inferences

To aid in drawing validity inferences from a meta-analysis, the *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003) encourage that the boundary conditions of the meta-analysis be specified with respect to the content, structure, and scoring of the tests within the meta-analysis. We offer these boundary conditions. Conclusions concerning the validity of a SJT could be informed based on this meta-analysis if the test had job-related content, is used for the job or job family for which

it was designed, was a written test (as opposed to video) with either knowledge or behavioral tendency instructions, contained items with a situation as the stem and possible responses to the situation as item options, and where the test is scored objectively using either a rationally developed or an empirically developed key. If a test falls within these boundaries, this meta-analysis could be one source of information concerning validity.

Limitations of the Study and Calls for Additional Research

The first limitation is that for most of the data, the meta-analysis examining response instructions could not control for content differences across the SJTs. Thus, it is possible that at least some of the differences attributed to the response instruction moderator are actually due to unknown content differences that co-vary with response instructions. However, we refuted the credibility of the alternative content hypothesis with the results in Table 5 that controlled for the content of the SJT. In addition, we refuted the alternative content hypothesis by showing that individuals with SJT expertise could not meaningfully differentiate the response instructions used with the items. However, we encourage that our conclusion be further evaluated with research in which the content of SJTs is held constant.

The second limitation of this study is that almost all validity studies on SJTs are based on concurrent designs where the respondents are incumbents. Job applicants are more likely than incumbents to try to respond so as to improve their score. When job applicants are administered a SJT with behavioral tendency instructions, they may choose to respond with what they believe is the best action to take rather than what they would most likely do. This would make the SJT function as a knowledge instruction SJT. If many applicants are responding in this manner, the differences observed in this study between response instructions may become smaller. Thus, we encourage that more validity studies of SJTs use applicant data and that the response instruction moderator be evaluated in predictive studies as more such studies accumulate.

We declined to make corrections for measurement error in the SJTs because we argued that most situational judgments tests are heterogeneous, yet most reported reliabilities for such tests are coefficient alphas. We encourage researchers to consider better ways to estimate the reliability of SJTs. More reasonable estimates of SJT reliability would be derived from test–retest or parallel form reliability analyses.

We also encourage research on the specification of content assessed in SJTs and the relation between content and validity. It is reasonable to expect that some content will yield different criterion-related and construct validity than other content. We encourage research in test development technology so that SJTs can be written to achieve prespecified correlations with other measures.

The data source moderator needs additional attention. A reasonable cause of data source difference, publication bias, was not shown to explain the data source differences. Future research should reexamine the data source moderator as more data become available.

Finally, we encourage research on different ways of operationalizing knowledge and behavioral tendency instructions. In the literature, as shown in Table 2, we found four ways of operationalizing behavioral tendency instructions and seven ways of operationalizing knowledge instructions.

Conclusion

The use of SJTs in personnel selection has gained increasing interest. The study extends theory in personnel selection by applying the typical versus maximal performance distinction to predictors. This distinction provides a credible explanation for the moderating effect of response instructions on construct validity. The study also informs practice in the use of SJTs by providing estimates of the criterion-related validity of SJTs and by exploring the incremental validity of cognitive ability, personality, and SJTs. Finally, it guides practice by showing that the cognitive and noncognitive correlates of SJTs can be altered through response instructions. To date, SJTs are the only personnel selection instruments that display this characteristic. Although these analyses should be repeated as additional data cumulate, our results are currently the best estimates of the validity of SJTs. These results can guide the development of future SJTs and provide evidence for the validity of tests that fall in the boundary of the studies examined in this research.

REFERENCES

*Beaty JC Jr, Howard MJ. (2001, April). *Constructs measured in situational judgment tests designed to assess management potential and customer service potential*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

*Bess TL. (2001). *Exploring the dimensionality of situational judgment: Task and contextual knowledge*. Unpublished master's thesis from Virginia Polytechnic Institute and State University. Supplemented by information from Morell Mullians to Michael McDaniel on July 25, 2006.

*Bosshardt M, Cochran. (1996). *Proprietary document*. Personal communication to Michael McDaniel.

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

*Bruce MM. (1965). *Examiner's manual, business judgment test, revised. 1-12*.

*Bruce MM. (1974). *Examiner's manual supervisory practices test revised*. Larchmont, NY: Martin M. Bruce, PhD.

*Bruce MM, Friesen EP. (1956). Validity information exchange. PERSONNEL PSYCHOLOGY, *9*, 380.

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

*Campbell JP, Dunnette MD, Lawler EE III, Weick KE Jr. (1970). *Research results: Actuarial studies of managerial effectiveness, managerial behavior, performance, and effectiveness* (pp. 164–198, 500–528). New York: McGraw-Hill.

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

Cardall AJ. (1942). *Preliminary manual for the test of practical judgment*. Chicago: Science Research Associates.

*Carrington DH. (1949). Note on the Cardall practical judgment test. *Journal of Psychology, 33*, 29–30.

*Carter GW, Johnson JW (Eds.). (2002). *Institute report #408*. Minneapolis, MN: Personnel Decisions Research Institutes, Inc.

*Chan D. (2002, April). *Situational judgment 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 D. (2006). Interactive effects of situational judgment effectiveness and proactive personality on work perceptions and work outcomes. *Journal of Applied Psychology, 91*, 475–481.

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

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 JP, Haaland DE. (2000, April). *The relationship between job knowledge and situational judgment test performance*. Paper presented at the 15th Annual Conference of the Society for Industrial and Organization Psychology, Inc. New Orleans, LA.

Cooper HM. (2003). Editorial. *Psychological Bulletin, 129*, 3–9.

*Corts DB. (1980). *Development of a procedure for examining trades and labor applicants for promotion to first-line supervisor*. Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center, Research Branch.

Cronbach LJ. (1949). *Essentials of psychological testing*. New York: Harper & Row.

Cronbach LJ. (1984). *Essentials of psychological testing*. (4th ed.) New York: Harper & Row.

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

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

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

DuBois CLZ, Sackett PR, Zedeck S, Fogli L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and white-black differences. *Journal of Applied Psychology, 78*, 205–211.

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

Duval SJ. (2005). The "trim and fill" method. In Rothstein H, Sutton AJ, Borenstein M (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (pp. 127–144). Chichester, UK: Wiley.

Duval SJ, Tweedie RL. (2000a). A non-parametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89–98.

Duval SJ, Tweedie RL. (2000b). Trim and fill: A simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 276–284.

Dye DA, Reck M, McDaniel MA. (1993). Moderators of the validity of written job knowledge measures. *International Journal of Selection Assessment, 1*, 153–157.

File QW. (1943). *How Supervise? (Questionnaire Form B)*. New York: The Psychological Corporation.

*Gekoski N, Schwarty SL. (1960). *SRA Supervisory Index*. Chicago: Science Research Associates.

*Greenberg SH. (1963). *Supervisory judgment test manual (Technical Series No. 35)*. Washington, DC: Personnel Measurement Research and Development Center, U.S. Civil Service Commission, Bureau of Programs and Standards, Standards Division.

*Grubb WL. (2003). *Situational judgment and emotional intelligence tests: Constructs and faking*. Dissertation completed at Virginia Commonwealth University, Richmond. ISBN 0-496-55186-8.

*Hanson MA. (1994). *Development and construct validation of a situational judgment test of supervisory effectiveness for first-line supervisors in the U.S. Army*. Unpublished doctoral dissertation, University of Minnesota.

*Hartman NS, Grubb WL III. (2005, November). *Situational judgment tests and validity: It's a matter of instruction*. Paper presented at the Annual Meeting of the Southern Management Association, Charleston, SC.

*Hill AM. (1950). *An evaluation of the Cardall Test of Practical Judgment in industrial supervisory selection*. Unpublished master's thesis, University of Toronto, Toronto.

*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.

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

*Houston JS, Schneider RJ. (1997). *Institute Report #292*. Minneapolis, MN: Personnel Decisions Research Institutes.

Huffcutt JM, Roth PL, McDaniel MA. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology, 81*, 459–473.

*Hunter DR. (2002). *Measuring general aviation pilot decision-making using a situational judgment technique*. Washington, DC: Federal Aviation Administration.

Hunter JE, Hunter RF. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Hunter JE, Schmidt FL. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hunter JE, Schmidt FL. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd edition). Newbury Park, CA: Sage.

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

*Jagmin N. (1985). *Individual differences in perceptual/cognitive constructions of job-relevant situations as a predictor of assessment center success*. College Park, MD: University of Maryland.

Jones C, DeCotiis TA. (1986). Video-assisted selection of hospitality employees. *The Cornell H.R.A. Quarterly,* 67–73.

*Jones MW, Dwight SA, Nouryan TR. (1999, April). *Exploration of the construct validity of a situational judgment test used for managerial assessment*. Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

*Jurgensen CE. (1959). Supervisory Practices Test. In Buros OK (Ed.), *The fifth mental measurements yearbook—Test & reviews: Vocations—specific vocations* (pp. 946–947). Highland Park, NJ: Gryphon.

*Kang M. (2005). *Comparison of validities of situational judgment tests according to instruction types and scoring alternatives*. Unpublished master's thesis. Sungkyunkwan University.

*Kirkpatrick DL, Planty E. (1960). *Supervisory inventory on human relations*. Chicago: Science Research Associates.

Latham GP, Saari LM, Pursell ED, Campion MA. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422–427.

*Leeds JP, Griffith R, Frei RL. (2003). The development and validation of a situational judgment test for security officers. *Security Journal, 16*, 63–78.

*Leaman JA, Vasilopoulos NL. (1997). *Development and validation of the U.S. immigration officer applicant assessment*. Washington, DC: U.S. Immigration and Naturalization Service.

*Leaman JA, Vasilopoulos NL. (1998). *Development and validation of the detention enforcement officer applicant assessment*. Washington, DC: U.S. Immigration and Naturalization Service.

*Leaman JA, Vasilopoulos NL, Usala PD. (1996, August). *Beyond integrity testing: Screening border patrol applicants for counterproductive behaviors*. Paper presented at the 104th Annual Convention of the American Psychological Association, Washington, DC.

*Lobsenz RE, Morris SB. (1999, April). *Is tacit knowledge distinct from* g*, personality, and social knowledge?* Paper presented at the 14th Annual Conference 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 Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

McDaniel MA, Hurtz GM, Donovan JJ. (2006, May). *An evaluation of publication bias in Big Five validity data*. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

McDaniel MA, McKay P, Rothstein H. (2006, May). *Publication bias and racial effects on job performance: The elephant in the room*. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. (2001). Predicting job performance using situational judgment tests: 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, Rothstein H, Whetzel DL. (2006). Publication bias: A case study of four test vendors. PERSONNEL PSYCHOLOGY, *59*, 927–953.

McDaniel MA, Whetzel DL. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33*, 515–525.

*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?* Paper presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, ON.

*Meyer HH. (1951). Factors related to success in the human relations aspect of work-group leadership. *Psychological Monographs: General and Applied, 65*, 1–29.

*Millard KA. (1952). Is How Supervise? an intelligence test? *Journal of Applied Psychology, 36*, 221–225.

*Moss FA. (1926). Do you know how to get along with people? *Scientific American, 135*, 26–27.

*Motowidlo S. (1991). *The situational inventory: A low fidelity for employee selection*. Gainesville: University of Florida.

Motowidlo S, Borman W, Schmit M. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10*, 71–83.

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

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

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

*Mullins ME, Schmitt N. (1998, April). *Situational judgment testing: Will the real constructs please present themselves*. Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

*Nguyen NT. (2001). *Faking in situational judgment tests: An empirical investigation of the Work Judgment Survey*. Dissertation completed at Virginia Commonwealth University. ISBN 0-493-38637-8.

*Nguyen NT. (2004, April). *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.

*Nguyen NT, Biderman MD, McDaniel MA. (2005). Effects of response instructions on faking in a situational judgment test. *International Journal of Selection and Assessment, 13*, 250–260.

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

*O'Connell MS, McDaniel MA, Grubb WL III, Hartman NS, Lawrence A. (2002, April). *Incremental validity of situational judgment tests for task and contextual performance*. Paper presented at the 17th Annual Conference of the Society of Industrial Organizational Psychology, Toronto, Ontario.

*Ones DS, Viswesvaran C, Reiss AD. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660–679.

*Oswald FL, Schmitt N, Kim HB, Ramsay LJ, Gillespie MA. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89*, 187–207.

Palhaus DL. (1984). Two component models of social desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609.

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

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

*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 RE, Lim B, Chan K. (2001). Exploring relations between typical and maximum performance ratings and the five factor model of personality. PERSONNEL PSYCHOLOGY, *54*, 809–843.

Ployhart RE, Ehrhart MG. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11*, 1–16.

*Ployhart RE, Ryan AM. (2000, April). *Integrating personality tests with situational judgment tests for the prediction of customer service performance*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

*Potosky D, Bobko P. (2004). Selection testing via the internet: Practical considerations and exploratory empirical findings. PERSONNEL PSYCHOLOGY, *57,* 103–1034.

*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.

*Richardson, Bellows & Henry Company, Inc. (1949). *Test of supervisory judgment: Form S*. Washington, DC: Richardson, Bellows, and Henry.

*Richardson, Bellows & Henry Company, Inc. (1963). *Male reliability and validity studies, Test of supervisory judgment: Form T*. Washington, DC: Richardson, Bellows, and Henry.

*Richardson, Bellows & Henry Company, Inc. (1988). *The supervisor profile record*. Minneapolis, MN: National Computer Systems.

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

Sackett PR, Zedeck S, Fogli L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73*, 482–486.

*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 FL, Hunter JE. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

*Schippmann JS, Prien EP. (1985). The Ghiselli self-description inventory: A psychometric appraisal. *Psychological Reports, 57*, 1171–1177.

*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.

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

*Spitzer ME, McNamara WJ. (1964). A managerial selection study. PERSONNEL PSYCHOLOGY, *17*, 19–40.

*Sternberg RJ, Wagner RK, Okagaki L. (1993). *Practical intelligence: The nature and role of tacit knowledge in work and at school*. New Haven, CT: Yale University.

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

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

*Timmereck CW. (1981). *Moderating effect of tasks on the validity of selection tests*. Unpublished Doctoral Dissertation, University of Houston, Houston, TX.

*Van Iddekinge CH, Dager CE, Le H. (2005). Cross-instrument analyses. In Knapp DJ, Sager CE, Tremble TT (Eds.), *Development of experimental army enlisted personnel selection and classification tests and job performance. Technical Report 1168*. Arlington, TX: United States Army Research Institute for the Behavioral and Social Sciences.

*Vasilopoulos NL, Reilly RR, Leaman JA. (2000). The influence of job familiarity and impression management on self-report measure scale scores and response latencies. *Journal of Applied Psychology, 85*, 50–64.

*Vasilopoulos NL, Cucina JM, Hayes TL, McElreath JA. (2005, April). *Effect of situational judgment test response instructions on validity*. Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles.

*Wagner RK, Sternberg RJ. (1991). *The common sense manager*. San Antonio: Psychological Corporation.

*Watley DJ, Martin HT. (1962). Prediction of academic success in a college of business administration. *Personnel and Guidance Journal,* 147–154.

*Weekley JA. (2005). Personal communication from Jeff Weekley to Michael A. McDaniel, June 22, 2005.

*Weekley JA. (2006). Personal communication to Michael McDaniel in June and July 2006.

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. (2006). *Situational judgment tests: Theory, management, and application*. Mahwah, NJ: Erlbaum.

*Weekley JA, Ployhart RE. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance, 18*, 81–104.

Weekley JA, Ployhart RE, Holtz BC. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In Weekley JA, Ployhart RR (Eds.), *Situational judgment tests: Theory, management, and application*. Mahwah, NJ: Erlbaum.

Whetzel DL. (2006, May). *Publication bias the validity of customer service measures.* Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Inc. Dallas, TX.

*Wickert FR. (1952). Relation between How Supervise? intelligence and education for a group of supervisory candidates in industry. *Journal of Applied Psychology, 36*, 301–303.