Journal of Applied Psychology
2001, Vol. 86, No. 4, 730–740

0021-9010/01/$5.00 DOI: 10.1037//0021-9010.86.4.730

Use of Situational Judgment Tests to Predict Job Performance: A Clarification of the Literature

Michael A. McDaniel
Virginia Commonwealth University

Frederick P. Morgeson
Michigan State University

Elizabeth Bruhn Finnegan
Ameren Services

Michael A. Campion
Purdue University

Eric P. Braverman
AT&T

Although situational judgment tests have a long history in the psychological assessment literature and continue to be frequently used in employment contexts, there has been virtually no summarization of this literature. The purpose of this article is to review the history of such tests and present the results of a meta-analysis on criterion-related and construct validity. On the basis of 102 coefficients and 10,640 people, situational judgment tests showed useful levels of validity (ρ = .34) that were generalizable. A review of 79 correlations between situational judgment tests and general cognitive ability involving 16,984 people indicated that situational judgment tests typically evidence relationships with cognitive ability (ρ = .46). On the basis of the literature review and meta-analytic findings, implications for the continued use of situational judgment tests are discussed, particularly in terms of recent investigations into tacit knowledge.

Tests assessing an individual's judgment concerning work-related situations have had a long history in the psychological assessment literature. Despite this extensive use of such situational judgment tests, however, there does not exist a thorough summary of research, and few conclusions can be drawn concerning the validity of such tests. This gap in the literature is all the more striking given the increased use of such tests over the past 10 years (e.g., Chan & Schmitt, 1997; Hanson & Borman, 1989; Motowidlo, Dunnette, & Carter, 1990; Olson-Buchanan et al., 1998; Smith & McDaniel, 1998).

The present article endeavors to address the gaps in the situational judgment test literature in three ways. First, we examine the situational judgment test literature to provide a more detailed description of the nature of these tests, to understand their essential elements, and to review the relevant findings. Second, we present the results of a meta-analysis that statistically cumulates empirical findings on situational judgment tests across studies. We examine the criterion-related validity of these tests and their relationship with cognitive ability. Third, on the basis of the literature review and meta-analytic findings, we discuss implications for the continued use of situational judgment tests, particularly in terms of recent research into tacit knowledge (e.g., Sternberg, Wagner, Williams, & Horvath, 1995).

In this study, we broadly defined situational judgment tests as any paper-and-pencil test designed to measure judgment in work settings. Some of these tests can be classified as situational, in that a scenario is described and the respondent must identify an appropriate response from a list of alternatives. The following item from an Army judgment test is illustrative:

> A man on a very urgent mission during a battle finds he must cross a stream about 40 feet wide. A blizzard has been blowing and the stream has frozen over. However, because of the snow, he does not know how thick the ice is. He sees two planks about 10 feet long near the point where he wishes to cross. He also knows where there is a bridge about 2 miles downstream. Under the circumstances he should:
> A. Walk to the bridge and cross it.
> B. Run rapidly across on the ice.
> C. Break a hole in the ice near the edge of the stream to see how deep the stream is.
> D. Cross with the aid of the planks, pushing one ahead of the other and walking on them.
> E. Creep slowly across the ice. (Northrop, 1989, p. 190)

Other measures do not present situations but rather require the respondents to indicate their level of agreement with statements concerning the appropriateness of various work-related behaviors. For example, a respondent might be asked to agree or disagree with the following statement: "An effective supervisor will closely monitor the work of a new employee." Our definition of situational

Michael A. McDaniel, Department of Management, Virginia Commonwealth University; Frederick P. Morgeson, Department of Management, Michigan State University; Elizabeth Bruhn Finnegan, Ameren Services, St. Louis, Missouri; Michael A. Campion, Krannert School of Management, Purdue University; Eric P. Braverman, AT&T, Morristown, New Jersey.

Correspondence concerning this article should be addressed to Michael A. McDaniel, Department of Management, School of Business, Virginia Commonwealth University, Richmond, Virginia 23284-4000. Electronic mail may be sent to mamcdani@vcu.edu.

judgment tests did not include verbal or logical reasoning tests, which are sometimes labeled *judgment tests.*

Review of Research on Situational Judgment Tests

The use of situational judgment tests dates back to the 1920s. One of the first widely used and evaluated tests that attempted to measure judgment was the George Washington Social Intelligence Test. One of the subtests, called Judgment in Social Situations, included many items in work situations. This subtest required "keen judgment, and a deep appreciation of human motives, to answer correctly" (Moss, 1926, p. 26). Several solutions to each situation were offered in a multiple-choice format, only one of which was correct. In an early review of empirical studies, Thorndike and Stein (1937) criticized the test. They noted that correlations between the test and other tests of presumed social attributes were very low and that the test was highly correlated with tests of abstract intelligence. They concluded that the test did not tap a distinct social intelligence construct and that it was best classified as a measure of general intelligence. In later reviews, Thorndike (1941) and Taylor (1949a) came to similar conclusions.

During World War II, Army psychologists attempted to assess the judgment of soldiers (Northrop, 1989). These judgment tests consisted of scenarios with a number of alternative responses. Solutions rested on the person's ability to draw on his or her common sense, experience, and general knowledge, rather than logical reasoning. In reviewing such tests, Northrop stated that although a judgment factor had originally been isolated by early Army research, tests that attempted to measure judgment were factorially complex. She concluded that judgment is a *g*-saturated and multifaceted cognitive attribute that is not factorially pure and does not consistently load on any specific cognitive factor.

Starting in the 1940s, a number of situational judgment tests were developed to measure supervisory potential. These tests included the Practical Judgment Test (Cardall, 1942), How Supervise? (File, 1945; File & Remmers, 1948), the Supervisory Practices Test (Bruce & Learner, 1958), the Business Judgment Test (Bruce, 1965), the Supervisory Judgment Test (Greenberg, 1963), and the Supervisory Inventory on Human Relations (Kirkpatrick & Planty, 1960). Because these tests are quite similar in form, only a few are discussed here.

The Practical Judgment Test consisted of multiple-choice items describing everyday business and social situations (Cardall, 1942). Items tapping special knowledge or factual information were explicitly discarded during test construction. Respondents indicated the best action to take in each situation. Cardall asserted that practical judgment is a unique factor that is statistically independent from factors such as intelligence and academic background. Several researchers, however, offered evidence indicating that the Practical Judgment Test correlated significantly with tests of general intelligence (e.g., Carrington, 1949; Taylor, 1949b).

The Supervisory Practices Test was designed to measure the ability of a supervisor to function effectively in situations that required decisions involving people (Bruce, 1974; Oliverio, 1959). The items required the respondent to make judgments on what to do in typical work-related situations that a supervisor might face. Scores on the test successfully distinguished between groups of supervisors and nonsupervisors (Bruce & Learner, 1958). In addition, it was found that scores correlated with tests of mental abilities (r = .18–.35).

Along with the Supervisory Judgment Test, another of the most widely used and researched of these tests is How Supervise? It was designed to measure a supervisor's knowledge and insight concerning human relations in industry (File & Remmers, 1948, 1971). The items were divided into sections. The supervisory practices and the company policies sections concerned specific situation-based actions that the supervisor endorsed as desirable to initiate. The supervisory opinions section concerned problem situations that the supervisor must face when dealing with workers, and each item consisted of a short phrase to which a respondent indicated agreement, disagreement, or uncertainty. Criterion-related validity has been mixed. For example, the 1971 revision of the manual listed nine studies that found positive relationships between How Supervise? scores and performance ratings. The manual also reported seven studies that found no such relationship. In narrative reviews, both Mandell (1953) and Rosen (1961) concluded that the positive evidence outweighed the negative evidence for the validity of the test.

It is interesting that File and Remmers (1971) reported a large number of studies that found significant correlations between How Supervise? and tests of general cognitive abilities. For example, Millard (1952) examined the relationship between scores on How Supervise? and scores on a test of intelligence. He found correlations of .71 and .62 in two groups of supervisors and a nonsignificant correlation in a third group. He found similar but smaller relationships with educational level. From these findings, Millard concluded that How Supervise? was essentially an intelligence test.

In the late 1950s and the early 1960s, situational judgment tests were also used by large organizations as part of selection test batteries to predict managerial success. For example, the Standard Oil Company of New Jersey designed a program of research called the Early Identification of Management Potential to identify employees who had the potential to be successful in management (Campbell, Dunnette, Lawler, & Weick, 1970). One of the measures used in the test battery was the Management Judgment Test. This particular measure contained items that described problem situations and presented test takers with several possible choices of action or decision. The respondent was required to indicate which action he or she believed was the most appropriate.

There recently has been renewed interest in the use of situational judgment measures for predicting job success. For example, the U.S. Office of Personnel Management designed Test 905 to assess the human relations capacity and potential of applicants for promotion to first-line federal trades and labor supervisory positions (Corts, 1980). The multiple-choice items on Test 905 consist of problem situations that involve interaction with workers. Respondents indicate which solutions, from a number of alternatives, would best solve the problem. A small concurrent validity study reported by Corts found small relationships (rs = .06–.21) between scores on Test 905 and various job performance criteria.

Motowidlo et al. (1990) examined the use of a situational judgment test, which they referred to as a low-fidelity simulation, for selecting entry-level managers. This measure, like other judgment measures, presented applicants with verbal descriptions of work situations and several alternative courses of action for each situation. Examinees chose both the response they would most

likely make and the response they would least likely make in each situation. In validation studies with samples of managers from seven different companies, correlations between the test and various job performance criteria ranged from the .20s to the .40s. The test was generally independent of cognitive aptitude test scores ($r < .10$), but low to moderate correlations were observed in some samples for grade point average ($r < .30$) and class rank ($r < .17$). Given that incumbents had been preselected partly on the basis of their aptitude test scores and academic achievement, there was likely restriction of range on these variables. Thus, any conclusions that this measure was independent from general cognitive ability should be viewed cautiously.

Wagner and Sternberg (1991) published a test called the Tacit Knowledge Inventory for Managers (TKIM). This measure is based on their theory of tacit knowledge, or "practical know-how that usually is not openly expressed or stated and which must be acquired in the absence of direct instruction" (Wagner, 1987, p. 1236). The purpose of this measure is to identify individuals whose tacit knowledge indicates the potential for successful performance in managerial or executive careers (Wagner & Sternberg, 1991). The TKIM presents scenarios that require respondents to choose a course of action from a list of alternatives. These scenarios differ from those of previously mentioned tests in that the TKIM scenarios are considerably longer and more detailed. Wagner and Sternberg (1991) reported five studies examining the criterion-related validity of tacit knowledge measures in academic and business settings, although no validity was presented for the TKIM itself. They found moderate correlations between their measure and a variety of criteria, some of which would be considered job performance measures. They also reported and claimed (e.g., Sternberg et al., 1995) that these measures were unrelated to measures of general cognitive ability. However, this conclusion should be tempered by the fact that their convenience samples (e.g., Yale University undergraduate students) were likely to evidence substantial range restriction on measures of general cognitive ability, thus attenuating observed relationships on the restricted predictor.

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

Summary of Research on Situational Judgment Tests

This prior research into situational judgment tests reveals several important points. First, it appears that these tests reflect a measurement method that can be used to assess a variety of constructs. In this respect, situational judgment tests are similar to other selection techniques such as employment interviews or assessment centers, in that the tests are measurement methods that may assess a variety of constructs.

Second, these tests tend to be quite similar in format. That is, they are typically in a paper-and-pencil format, present hypothetical problem situations that occur at work, and require the respondent to exercise judgment in picking or evaluating alternative courses of action. In this respect, they are what Wernimont and Campbell (1968) referred to as samples of likely job performance rather than signs of possible job performance. Third, these tests have demonstrated at least moderate validity. Finally, these studies evidence differing degrees of correlation with general cognitive ability, supporting the contention that situational judgment tests are a measurement method with different tests assessing different constructs or the same construct to varying degrees. Some show strong correlations with g, whereas others show correlations of lower magnitude.

Meta-Analysis of Situational Judgment Tests

Although narrative reviews help convey the qualitative differences and similarities among studies, they are necessarily subjective in their conclusions. In addition, there is considerable variance in the findings among studies that may be due to statistical artifacts such as sampling variation. As a result, we conducted a meta-analysis to provide a more objective and precise summary of the literature. The meta-analysis attempted to answer three questions: (a) What is the best estimate of the validity of situational work judgment tests? (b) What is the best estimate of the correlation with general cognitive ability? and (c) Are there any important moderators of these correlations?

Four possible moderators of the relationship between situational judgment tests and job performance and cognitive ability were considered. Two of these moderators represent points of variability in how the construct of situational judgment is operationalized in the published literature and are commonly investigated in other selection contexts. The third moderator directly assesses how the g loading of a test affects validity. The fourth moderator concerns the potential moderating effect of whether the validity study has a predictive or a concurrent design.

The first moderator concerned whether a job analysis was used to develop the test. That is, some tests appear to be based on a detailed job analysis, whereas others appear to be based on the intuition of the test developer or a small sample of subject matter experts. For example, the items for the Supervisory Judgment Test and How Supervise? were prepared by small groups of experts. In contrast, the items for the Situational Judgment Test (Hanson & Borman, 1989) and Motowidlo et al.'s (1990) low-fidelity simulation were based on detailed critical-incident job analyses. Three additional arguments support job analysis as a potential moderator. First, job analysis is important to job relatedness; thus, the potential link to validity seems logical. Second, just because a commercially available test is used in a study rather than a new test developed, this does not mean a job analysis is irrelevant. A job analysis should be conducted to pick the commercial test to use. Third, previous meta-analyses have found job analysis to be an important moderator of validity in such areas as personality tests (Tett, Jackson, & Rothstein, 1991) and employment interviewing (Wiesner & Cronshaw, 1988). Thus, we anticipated that tests developed from a job analysis would have higher correlations with job performance.

The second moderator explored the amount of detail in the work-related situational questions. Some tests use longer, more detailed questions, whereas others use brief, nonspecific questions. In our review of the tests, it appeared that longer and more detailed

questions tended to be much more nuanced and to provide far more work-specific information. Thus, the detail in the question may enhance the importance of work-specific knowledge. Questions with greater detail make it likely that individuals with greater work-specific knowledge will perform better. In contrast, question length may increase the *g* loading by increasing the reading-ability requirements. In fact, increasing the latter may increase validity by itself. Finally, examining different types of questions as a moderator of validity has a substantial precedent. For example, several authors have examined question type as a moderator of the validity of interview questions (e.g., Campion, Campion, & Hudson, 1994; Huffcutt & Roth, 1998; McDaniel, Whetzel, Schmidt, & Maurer, 1994). This evidence suggests that tests with detailed questions will have higher correlations with job performance.

The third moderator explored was whether the *g* loading of a test influences validity. The hypothesis was that the validity of situational judgment tests is due, in part, to their correlation with general mental ability. From an operational viewpoint, if the validity of a situational judgment test is primarily due to the test measuring general cognitive ability, one could use an existing mental ability test instead of investing considerable resources in developing a situational judgment test. We expected that tests with larger relationships with cognitive ability would also have higher criterion-related validity.

The fourth moderator concerned whether the criterion-related validity study was a predictive or a concurrent design. This moderator was suggested for analysis by a reviewer. It was explored for the criterion-related validity analyses but not for the analyses concerning the relation between situational judgment test scores and general cognitive ability because many of the correlations between situational judgment tests and general cognitive ability were not drawn from studies involving criterion-related validity analyses.

Method

Sample of Studies

An extensive literature search identified data from 39 different situational judgment tests that were useful for these analyses. We searched PsycLIT for the years 1887–2000. We also searched existing relevant papers for references to papers presented or published earlier. Data were obtained on measures that are, or have been, commercially marketed, including the Business Judgment Test, How Supervise?, the Test of Practical Judgment, the Supervisory Index, the Supervisory Inventory on Human Relations, the Social Judgment Test for Supervisors, the Supervisor's Opinionaire, the Supervisory Practices Test, the Supervisory Problems Test, the Supervisory Profile Record, the TKIM, the Teamwork–KSA Test, and the Test of Supervisory Judgment. We also located data on four measures used by the federal government. The remaining measures were neither government-owned nor commercially marketed and primarily were developed by consulting firms for specific clients.

We obtained 102 correlation coefficients between situational judgment tests and job performance. These coefficients were based on data from 10,640 participants. We obtained 80 correlation coefficients, based on 22,580 individuals, between judgment tests and measures of general cognitive ability. These two distributions of correlation coefficients were meta-analyzed using procedures described by Hunter and Schmidt (1990).

Analysis Method

The Hunter–Schmidt (Hunter & Schmidt, 1990) psychometric meta-analytic method used in this study is based on the hypothesis that much of the variation in results across studies may be due to statistical and methodological artifacts rather than to substantive differences in underlying population relationships. Many of these artifacts also reduce the correlations below their true (i.e., population) values. This method determines the variance attributable to sampling error and to differences between studies in reliability and range restriction, subtracts that amount from the total amount of variation, and yields estimates of the true variation across studies and of the true average correlation (Hunter & Schmidt, 1990). Artifact distribution meta-analysis, using the interactive method, was used (Hunter & Schmidt, 1990, chap. 4). The mean observed correlation ($\bar{r}$) was used in the sampling error variance formula (Hunter & Schmidt, 1990, pp. 208–210; Schmidt et al., 1993). In our analyses, the estimated population mean reflected corrections for measurement error in the criterion. No corrections for measurement error in the test were made to the mean. The estimated population variance resulted from correcting the observed variance for sampling error and for differences across studies in both the test reliability and the criterion reliability. The computer program used was described by McDaniel (1986a). Additional detail on the program is presented in Appendix B of McDaniel, Schmidt, and Hunter (1988).

Reliability, Range Restriction, and g *Loading Data*

Table 1 shows the reliability artifact distribution for the situational judgment tests. This distribution was empirically derived on the basis of a review of all located studies using situational judgment tests. Scant information was available on the criterion reliability for the reported coefficients. Therefore, for the analysis of correlations between situational judgment measures and job performance, the job performance criterion reliability from Pearlman (1979) was used. The use of a criterion reliability distribution with a mean value of .60 may be conservative (i.e., underestimates the true validity of the predictors) because Rothstein (1990) found that across 9,975 employees and across all time periods of supervisory exposure to employees, the mean interrater reliability for one rater was .48. Because data were not available concerning the amount of range restriction, we did not use any corrections for range restriction. Notwithstanding this lack of corrections for range restriction, it is highly likely that there was at least some range restriction in these data. Thus, the results to be

Table 1
Reliability Artifact Distribution for Situational Judgment Tests

Reliability	Frequency in distribution
.43	1
.62	1
.64	2
.67	2
.69	2
.70	2
.72	1
.74	3
.77	1
.79	1
.80	3
.81	1
.82	3
.83	1
.84	1
.85	1
.86	1
.87	1
.88	2
.89	1
.91	1
.94	1

reported are likely conservative estimates (i.e., underestimates) of the true validity of the tests.

The meta-analysis of the correlations between situational judgment tests and measures of general cognitive ability also used artifact distribution meta-analysis. The population distribution was estimated by correcting the observed validities for measurement error in both the situational judgment measures and the cognitive ability measures. Specifically, the mean was corrected for measurement error in both the tests and the criteria. The variance was corrected for sampling error and differences across studies in measurement error in both the tests and the criterion. The reliability distribution shown in Table 1 was used for the situational judgment measures. The cognitive test reliability distribution offered by Pearlman (1979) was used to correct for the measurement error in the general cognitive ability measures (average criterion reliability of .80).

We measured the *g* loading of the test by locating and cumulating correlations between *g* measures and situational judgment tests. The sample-size weighted mean correlation was the operational measure of the *g* loading. For the purpose of the meta-analysis, we defined a high *g* loading as those with mean correlations greater than .50, a medium *g* loading as correlations between .35 and .50, and a low *g* loading as correlations less than .35.

Decision Rules

We used five main decision rules. First, participants had to be employees or applicants. Studies of students were excluded to avoid concerns over generalizability to work-related settings. Second, the predictor had to be a paper-and-pencil situational judgment test. Other test types using situational questions, such as interviews (e.g., Latham, Saari, Pursell, & Campion, 1980), were not included. We also did not include data from video-based situational judgment tests (e.g., Dalessio, 1994; C. Jones & DeCotiis, 1986) because there were too few data to allow separate analyses. Third, when possible, the immediate supervisor's rating of job performance was used. When use of this rating was not possible, we tried to use the rater with the most contact with the employees. For example, if results included supervisor ratings and ratings from others, we used the supervisor ratings because they had the most contact with the employees.

Finally, if more than one criterion was reported and one was a performance criterion, we used the performance criterion. Almost all of the studies used supervisory ratings as criteria. We did not include criteria such as years of school completed beyond high school, hierarchical level, number of employees, years of management experience, percentage of time spent in various duties, year of degree, conferences attended, sick leave or vacation leave taken, salary, ratings of potential, or turnover because they are not actual job performance measures.

Results

Interpretation of Tables

Table 2 presents meta-analytic results of the criterion-related validities of situational judgment tests. The first column identifies the distribution of validities analyzed. Columns 2 through 5 present information about the observed distribution, including the total sample size across studies, the number of correlation coefficients, and the mean and standard deviation of the observed distribution. The last four columns of Table 2 present an estimate of the population distribution. We present the estimated population mean (ρ), the estimated population standard deviation (σ_ρ), the percentage of variance in the observed distribution due to artifacts, and the 90% credibility value for the distribution of true validities. The estimated mean population validity was corrected solely for measurement error in the criteria; it was not corrected for measurement error in the predictors. The variance of the estimated population distribution was corrected for sampling error and differences across studies in predictor reliability and criterion reliability. Because neither the mean nor the variance was corrected for range restriction, the mean estimates are likely to underestimate the population validity, and the validity variance estimates are likely to be overestimated.

Table 3 presents the meta-analytic results of the correlations of situational judgment tests with measures of general cognitive ability. These reported statistics are similar to those outlined above. The estimated population distribution was corrected for measurement error. The estimated mean population validity was corrected for measurement error in both the situational judgment tests and the cognitive ability tests. The variance of the estimated population distribution was corrected for sampling error and differences across studies in the reliability of both tests. In addition, the 10th and 90th percentile values of the credibility interval of the population distribution are reported.

Validity of Situational Judgment Tests

The first row of Table 2 shows that the estimated population validity of situational judgment tests was .34. In addition, the 90th percentile credibility value was positive, which indicates that the validities were likely to be positive in nearly all cases. Therefore, although we encourage the replication of our analyses as new data become available, we believe that the data collected to date provide evidence for the generalized validity of situational judgment tests.

Part of the variation in validity across studies, however, appears to have been due to differential test validity. That is, there was nontrivial variability around the estimated population mean correlation, which indicates that some tests demonstrated greater validity than others. When we examined the results for the three tests that had enough coefficients to analyze separately (i.e., at least 10 coefficients), the estimated population means varied from .21 for How Supervise? to .41 for the Supervisory Judgment Test (see Table 2). Just as the content of the employment interview influences the validity of the interview (McDaniel et al., 1994), the content of situational judgment tests may influence their validity. Our moderator analyses attempted to discern the characteristics that cause some situational judgment tests to have higher validity than others.

We next examined the four moderators. Consistent with our predictions, the results highlighted the importance of using job analysis to develop tests, because those tests based on a job analysis had higher validity (.38) than tests not based on a job analysis (.29). Contrary to our predictions, tests with less detailed questions had slightly higher validity (.35) than tests with more detailed questions (.33). The *g* loading of a test did not linearly affect the validity of the test. Finally, the mean validity from studies with predictive designs (.18) was lower than the mean validity from studies with concurrent designs (.35).

Unfortunately, these moderator analyses are not overly informative. The results for the detail of question moderator are limited by the availability of only 10 coefficients and a relatively small sample size for tests with detailed questions. The results for the study design are limited by the fact that there were only six predictive studies with a total sample size of 346. Thus, these

Table 2
Meta-Analytic Results of the Criterion-Related Validity of Situational Judgment Tests

Distribution	Observed distribution				Population distribution			
	N	No. of rs	Mean r	σ	ρ	σ_ρ	% of σ^2 due to artifacts	90% CV
All distributions	10,640	102	.26	.14	.34	.14	45	.16
Test distributions with at least 10 validity coefficients								
How Supervise?	1,242	16	.16	.14	.21	.11	64	.07
RBH Test of Supervisory Judgment	945	11	.25	.13	.33	.09	70	.21
Supervisory Judgment Test	2,983	21	.32	.08	.41	.00	100	.41
Test distribution containing tests with fewer than 10 validity coefficients								
Miscellaneous tests	5,470	54	.25	.16	.33	.16	37	.12
Analyses by job analysis category								
Tests based on job analysis	5,959	36	.29	.09	.38	.06	73	.29
Tests not based on job analysis	3,251	49	.22	.17	.29	.16	46	.08
Unknown use of job analysis	1,430	17	.24	.20	.31	.22	26	.03
Analyses by detail of question category								
Detailed questions	2,218	10	.25	.11	.33	.10	46	.20
Not detailed questions	6,747	71	.27	.14	.35	.12	53	.19
Unknown level of detail	1,675	21	.25	.19	.33	.21	29	.06
Analyses by g loading of test								
High g loading (>.50)	3,269	25	.31	.09	.41	.03	94	.37
Medium g loading (.35–.50)	1,826	22	.14	.13	.18	.10	66	.05
Low g loading (<.35)	3,443	42	.26	.15	.34	.14	49	.16
Unknown g loading	2,102	13	.29	.14	.38	.15	33	.18
Analyses by study design								
Predictive	346	6	.14	.14	.18	.05	91	.12
Concurrent	10,294	96	.27	.14	.35	.14	45	.17

Note. CV = credibility value; RBH = Richardson, Bellows, and Henry.

analyses did little to further elucidate the effects of the proposed moderators.

In summary, situational judgment tests showed generalizable validity for almost all the distributions examined. Despite the generalization of the validity, the tests varied in their population validities. Our moderator analyses showed that tests based on a job analysis were substantially more valid (.38) than tests not based on a job analysis (.29). The results for the question detail moderator, the g loading, and the study design moderators are much less clear. Conclusions on these moderators should await the accumulation of more data and a replication with hierarchical extension of the current meta-analyses.

Relationship Between Situational Judgment Tests and General Cognitive Ability

This distribution of correlations between situational judgment tests included a study (Pereira & Harvey, 1999, Study 2) with a very large sample size (N = 5,586) and a coefficient much smaller (.14) than the average of all the other studies (observed mean r = .36). For all distributions that contained the coefficient, we report the analysis with and without the coefficient. Although there is no correct answer concerning which distribution is the "best" to interpret and discuss, because of the differences produced by Pereira and Harvey, we discuss the results that excluded this coefficient.

The second row of Table 3 shows that situational judgment tests had a mean correlation of .46 with general cognitive ability tests. There was, however, variability around the mean correlation, which indicates some situational judgment tests are more highly correlated with general cognitive ability than others. This finding suggests there are potential moderators of this relationship. As shown in Table 3, situational judgment tests based on a job analysis were more highly related to general cognitive ability (.50) than measures not based on a job analysis (.38). In addition, situational judgment tests with less detailed questions were more highly related to general cognitive ability (.56) than those with more detailed questions (.47).

Given this relationship with g, a reasonable question concerns whether situational judgment measures have incremental validity in predicting job performance beyond the contribution made by general cognitive ability. One can estimate the incremental validity on the basis of the correlations among these two tests and a

Table 3
Meta-Analytic Results of Correlations Between Situational Judgment Tests and General Cognitive Ability Measures

Distribution	Observed distribution				Population distribution			
	N	No. of rs	Mean r	σ	ρ	σ_ρ	% of σ^2 due to artifacts	10th and 90th percentile of CI
All coefficients	22,580	80	.31	.19	.39	.23	9	.09–.69
All coefficients excluding Pereira & Harvey's (1999) Study 2	16,994	79	.36	.19	.46	.23	12	.17–.75
Analyses by job analysis category								
Tests based on job analysis	18,817	49	.31	.19	.40	.24	6	.10–.71
Tests based on job analysis excluding Pereira & Harvey's (1999) Study 2	13,321	48	.39	.19	.50	.23	10	.21–.79
Tests not based on job analysis	3,154	24	.30	.15	.38	.16	31	.18–.58
Not known if test is based on job analysis	609	7	.10	.20	.13	.22	26	−.16–.41
Analyses by detail of question category								
Detailed questions	3,965	12	.37	.09	.47	.09	38	.34–.59
Not detailed questions	8,115	51	.44	.20	.56	.23	14	.26–.86
Not known if test has detailed questions	10,500	17	.18	.12	.23	.14	12	.05–.41
Not known if test has detailed questions excluding Pereira & Harvey's (1999) Study 2	4,914	16	.23	.16	.29	.19	13	.04–.54

Note. CI = credibility interval.

measure of job performance. Such an estimation process clearly depends on the values chosen for the correlations. For our analysis, we used the mean observed correlation for all situational judgment tests with job performance (r = .26) and the mean observed correlation between all g measures and all situational judgment measures, excluding Pereira and Harvey's (1999) Study 2 (r = .36). Identifying the correct correlation between measures of general cognitive ability and job performance is also a matter of judgment because the validity of general cognitive ability measures is moderated by the cognitive complexity demands of the jobs (Gutenberg, Arvey, Osburn, & Jeanneret, 1983; Hunter & Hunter, 1984; McDaniel, 1986b). We chose a value of .25, which was the value obtained in both Hunter's (1983) classic validity generalization study and Gandy's (1986) reanalysis of a modified version of the same data set. On the basis of these values, the estimated validity of a composite of a situational judgment test and a general cognitive ability test was .31, which was higher than either the situational judgment test alone (r = .26) or the general cognitive ability alone (r = .25). When one corrects the correlations for attenuation due to measurement error in the criterion (using an assumed reliability of .60 for the criterion), the validity of the composite of the situational judgment test is .40, which is higher than either the situational judgment test alone (.34) or the general cognitive ability alone (r = .32).

Thus, situational judgment tests may have incremental validity over and above tests of general cognitive ability. However, the validity of both situational judgment tests and cognitive ability tests is affected by moderators, and the intercorrelation of situational judgment tests and general cognitive ability varies across situational judgment tests. Thus, decision makers should consider what validities and test intercorrelations apply to their particular situation.

Discussion

These results provide insight into the nature of situational judgment tests. It is clear that they are good predictors of job performance. The estimated population validity of such measures is .34 across a wide range of measures and samples. This level of validity is comparable to such commonly used selection measures as assessment centers (Gaugler, Rosenthal, Thornton, & Benson, 1987), employment interviews (McDaniel et al., 1994), and biographical data measures (Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990). We argue that the estimated population mean validities are conservative (downwardly biased) because no corrections for range restriction were made. The results also reveal that the validity of these tests is generalizable. As such, recent attempts to predict job performance using situational judgment questions rest on a solid empirical foundation. In addition, tests based on a job analysis evidence higher validity than those not based on a job analysis.

Furthermore, situational judgment tests as a whole show substantial correlations with general cognitive ability (ρ = .46). There is, however, a nontrivial amount of population variance around this mean value. That is, some tests have substantial correlations with tests of general cognitive ability, whereas others have lower correlations. Nevertheless, an examination of the range of credibility values in Table 3 indicates that situational judgment tests typically have moderate correlations with general cognitive ability. For example, 9 of 10 situational judgment tests based on a job analysis had a correlation with general cognitive ability of at least .21 (the 10th percentile point), and the expected value (the mean of the distribution) would be .49. In short, it would be very unusual for a situational judgment test to have a correlation of .00 with a measure of general cognitive ability. Our g-loading moderator analysis did not yield compelling results, and awaiting further data,

we reserve judgment on the extent to which the validity of situational judgment tests is a function of the *g* loading of the tests.

Implications for Tacit Knowledge Research

This article was inspired, in part, by tacit knowledge research and claims that measures of tacit knowledge predict job performance while remaining largely unrelated to traditional measures of general cognitive ability (Sternberg et al., 1995). As we noted, the TKIM has been published to assess tacit knowledge (Wagner & Sternberg, 1991).

Curiously, although tacit knowledge has been conceptualized as "practical know-how that usually is not openly expressed or stated and which must be acquired in the absence of direct instruction" (Wagner, 1987, p. 1236), its operationalization in the Tacit Knowledge Inventory bears less resemblance to its construct definition and more to other situational judgment tests. An examination of the measures themselves reveals many similarities. For example, all of these tests present, in a paper-and-pencil format, situationally based questions. They include items that present hypothetical work-related scenarios that require either a judgment about the appropriateness of an action or a choice of the best action from among a range of options. In addition, they share a number of similarities in content, such that the situations tend to involve things such as teamwork, supervisory practices, and so on.

Comparing some of the test items reveals many of these similarities. Although exact items cannot be presented because of their proprietary nature, the similarities can be described. For example, Item 2 of the Tacit Knowledge Inventory presents a situation that assumes you are a second-level manager. An employee who reports to one of your subordinates wants to talk with you about some problems, but the employee has not talked to the immediate supervisor because of the problem's sensitive nature. You are presented with 10 possible courses of action to judge, including refusing to meet with the employee until he or she meets with the immediate supervisor, meeting with the employee but only if the supervisor is present, meeting with the employee and supervisor separately, and so on. This item is very similar to Item 2 of the Supervisory Practices Test (Bruce, 1974) and Item 23 of the Teamwork–KSA Test (Stevens & Campion, 1999). The Supervisory Practices Test asks what you would do if one of your most competent employees continually complained to other personnel about the lack of advancement. You are presented with three courses of action, including explaining to the employee the harmful effects of complaints on morale, asking the employee if a transfer to another section is desired, or asking the employee to discuss the matter with you. The Teamwork–KSA Test asks what you would do if you were having a problem with another member of your workteam and the problem was not just miscommunication. You are presented with four possible courses of action, including negotiating a solution, using persuasion, and so on.

As another example, Item 4 of the Tacit Knowledge Inventory asks the respondent to rate 10 different strategies for handling the day-to-day work of a business manager. These strategies include thinking in terms of tasks instead of hours worked, being in charge of all phases of every task or project, spending time planning the best way to do a task, and so on. This item is very similar to Item 41 of the Supervisory Practices Test (Bruce, 1974) and Item 21 of the Teamwork–KSA Test (Stevens & Campion, 1999). The Supervisory Practices Test asks how you would do your daily work. Your response options include taking care of the details while asking for planning from your boss, delegating details but planning with your boss, or handling the planning and details by yourself. The Teamwork–KSA Test asks you to indicate which of four response options would most likely help the team do its planning and coordinating. These options include examining past practices as a guide, considering priorities, pacing, sequencing of tasks and activities, and so on.

As these two examples illustrate, there are some obvious similarities between these purportedly different measures. This is not to suggest that all items are isomorphic between different tests, nor that these highlighted items are identical, but rather to show that these different tests are measuring similar subject matter by using a similar methodology. It is clear that, in the absence of comparative data, it is impossible to directly assess how the tacit knowledge measure relates to other situational judgment measures (although conducting the meta-analysis without the tacit knowledge data yielded virtually identical results). There are, however, a number of qualitative similarities that should not be overlooked. Given the prima facie similarities, it is useful to examine the implications of the present research for tacit knowledge.

As the present research demonstrates, the large literature on similar situational judgment tests has shown good criterion-related validity. Thus, the findings of validity for such measures in the tacit knowledge research are hardly a new finding. In contrast to the tacit knowledge research, the large literature on situational judgment tests has found moderate relationships with general cognitive ability. These results, coupled with a critical examination of tacit knowledge research, suggest a number of potential reasons why this is the case. First, tacit knowledge tests are not based on a job analysis and contain detailed questions (Wagner & Sternberg, 1991). The results of the present study suggest that these test characteristics serve to minimize observed relationships with cognitive ability (see Table 3).

Second, many of the validation studies of tacit knowledge tests have used samples with a restricted range of mental ability, such as undergraduate students at elite universities, business graduate students, and academic psychologists (Wagner, 1987; Wagner & Sternberg, 1985, 1991). Samples such as Yale University undergraduates are highly preselected on cognitive ability (through standardized college entrance tests) but may vary on "tacit knowledge." This preselection will attenuate the relationship between tacit knowledge and general cognitive ability and decrease the validity of general cognitive ability when compared with an unrestricted predictor (Nunnally & Bernstein, 1994).

Third, the two field studies reported by Wagner and Sternberg (1991) appear to be composed of small convenience samples (54 and 29 participants). Consequently, the observed results have uncertain stability and generalizability. In addition, tacit knowledge research has typically used unusual performance indices when compared with most validation research in industrial and organizational psychology. Criteria such as the quality of the psychology program, percentage of time spent on research, and whether a manager's organization was a top *Fortune* 500 company are not valid surrogates for job performance measures.

Finally, attempting to specify the constructs measured by the Tacit Knowledge Inventory (Wagner & Sternberg, 1991) is particularly difficult because of the way in which the test is scored.

That is, the test is keyed by having an "expert group" in the organization take the inventory. Scores are then based on differences from these expert answers. As a result, there is no uniform or standardized answer key. Although the constructs measured by any given key are restricted somewhat by the content of the questions, some expert groups might develop a key that is highly *g* loaded. Another expert group might give more points to responses that demonstrate teamwork. A third expert key might primarily reflect knowledge gained through work and life experiences. Because there is little research correlating measures of tacit knowledge with other measures commonly assessed in the wider industrial and organizational psychology literature, such as job knowledge and personality factors, it is difficult to integrate tacit knowledge research into a broader nomological network of constructs.

Conclusion

It is clear that situational judgment tests are good predictors of job performance. Our data suggest that general cognitive ability is a construct partly reflected in such measures, although general cognitive ability does not typically account for all the variance. By investigating four possible moderators, we attempted to clarify the mechanisms through which situational judgment tests predict job performance. Future research should repeat our moderator analyses on larger data sets and attempt to identify other possible moderators to allow a more complete understanding of the nomological network within which these situational judgment measures reside.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Anonymous. (1954). Validity information exchange (No. 7-065). *Personnel Psychology, 7,* 301.

*Bass, B. M., Karstendiek, B., McCullough, G., & Pruitt, R. C. (1954). Validity information exchange (No. 7-024). *Personnel Psychology, 7,* 159–160.

*Bruce, M. M. (1953). The prediction of effectiveness as a factory foreman. *Psychological Monographs, 67*(12, Whole No. 362).

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

*Bruce, M. M. (1974). *Examiner's manual: Supervisory Practices Test* (Rev. ed.). Larchmont, NY: Author.

*Bruce, M. M., & Friesen, E. P. (1956). Validity information exchange (No. 9-35). *Personnel Psychology, 9,* 380.

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

*Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. (1970). *Managerial behavior, performance and effectiveness.* New York: McGraw-Hill.

Campion, M. A., Campion, J. E., & Hudson, J. P. (1994). Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology, 79,* 998–1002.

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

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

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

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

*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.

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

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

*Corts, D. B. (1980). *Development and validation of a test for the ranking of applicants for promotion to first-line federal trades and labor supervisory positions* (Report No. PRR-80-30). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center.

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

*Decker, R. L. (1956). An item analysis of How Supervise? using both internal and external criteria. *Journal of Applied Psychology, 40,* 406–411.

*Dicken, C. F., & Black, J. D. (1965). Predictive validity of psychometric evaluations of supervisors. *Journal of Applied Psychology, 49,* 34–47.

*Dulsky, S. G., & Krout, M. H. (1950). Predicting promotion potential on the basis of psychological tests. *Personnel Psychology, 3,* 345–351.

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

File, Q. W., & Remmers, H. H. (1948). *How Supervise? manual 1948 revision.* New York: Psychological Corporation.

File, Q. W., & Remmers, H. H. (1971). *How Supervise? manual 1971 revision.* Cleveland, OH: Psychological Corporation.

Gandy, J. A. (1986). *Job complexity, aggregated subsamples, and aptitude test validity: Meta-analysis of the GATB data base.* Paper prepared for the U.S. Office of Personnel Management, Office of Personnel Research and Development, Washington, DC.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Benson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511.

*Gekoski, N., & Schwartz, S. L. (1966). *Supervisory Index.* Chicago: Science Research Associates.

*Greenberg, S. H. (1963). *Supervisory Judgment Test manual.* Washington, DC: U.S. Civil Service Commission.

Gutenberg, R. L., Arvey, R. D., Osburn, H. G., & Jeanneret, P. R. (1983). Moderating effects of decision-making/information-processing demands on test validities. *Journal of Applied Psychology, 68,* 602–608.

*Hanson, M. A. (1995). 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.

Hanson, M. A., & Borman, W. C. (1989, April). *Development of a situational judgment test to be used as a job performance measure for first-line supervisors in the U.S. Army.* Paper presented at the 4th Annual Conference of the Society for Industrial and Organizational Psychology, Boston, MA.

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

*Hilton, A. C., Bolin, S. F., Parker, J. W., Taylor, E. K., & Walker, W. B. (1955). The validity of personnel assessments by professional psychologists. *Journal of Applied Psychology, 39,* 287–293.

*Holmes, F. J. (1950). Validity of tests for insurance office personnel. *Personnel Psychology, 3,* 57–69.

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83,* 179–189.

Hunter, J. E. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery* (U.S. Employment Service Test Research Rep. No. 45). Washington, DC: U.S. Department of Labor.

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

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

*Jagmin, N. (1986). Individual differences in perceptual/cognitive constructions of job-relevant situations as a predictor of assessment center success (Doctoral dissertation, University of Maryland, 1985). *Dissertation Abstracts International, 47*(4-B), 1768–1769.

*Johnson, R. J. (1954). Validity information exchange (No. 7-090). *Personnel Psychology, 7,* 567.

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

*Jones, M. W., Dwight, S. A., & Nouryan, T. R. (1999, April). *Exploration of the construct validity of a situational judgment test used for managerial assessment.* Paper presented at the 14th Annual Convention of the Society for Industrial and Organizational Psychology, Atlanta, GA.

*Jurgensen, C. E. (1959). Supervisory Practices Test. In O. K. Burros (Ed.), *The fifth mental measurements yearbook* (pp. 946–947). Highland Park, NJ: Gryphon Press.

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

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65,* 422–427.

*Leaman, J. A., Vasilopoulos, N. L., & Usala, P. D. (1996, August). *Beyond integrity testing: Screening border patrol applicants for counterproductive behaviors.* Paper presented at the 104th Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

*Lobenz, R. E., & Morris, S. B. (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.

Mandell, M. M. (1953). How Supervise? In O. K. Buros (Ed.), *The fourth mental measurements yearbook* (pp. 774–775). Highland Park, NJ: Gryphon Press.

*McCormick, E. J., & Middaugh, R. W. (1956). The development of a tailor-made scoring key for the How Supervise? test. *Personnel Psychology, 9,* 27–37.

McDaniel, M. A. (1986a). Computer programs for calculating meta-analysis statistics. *Educational and Psychological Measurement, 46,* 175–177.

McDaniel, M. A. (1986b). The evaluation of a causal model of job performance: The interrelationships of general mental ability, job experience, and job performance (Doctoral dissertation, George Washington University, 1986). *Dissertation Abstracts International, 47*(3-A), 839–840.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of methods for rating training and experience in personnel selection. *Personnel Psychology, 41,* 283–314.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.

*Meyer, H. H. (1951). Factors related to success in the human relations aspect of work group leadership. *Psychological Monographs, 65*(3, Whole No. 320).

*Meyer, H. H. (1956). An evaluation of a supervisory selection program. *Personnel Psychology, 9,* 499–513.

*Millard, K. A. (1952). Is How Supervise? an intelligence test? *Journal of Applied Psychology, 36,* 221–224.

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

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

*Motowidlo, S. J., & 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, H. W. (1957). A measure of supervisory quality. *Journal of Applied Psychology, 41,* 405–408.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Oliverio, M. E. (1959). Supervisory Practices Test. In O. K. Buros (Ed.), *The fifth mental measurements yearbook* (pp. 954–955). Highland Park, NJ: Gryphon Press.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51,* 1–24.

*Parry, M. E. (1968). Ability of psychologists to estimate validities of personnel tests. *Personnel Psychology, 21,* 139–147.

Pearlman, K. (1979). *The validity of tests used to select clerical personnel: A comprehensive summary and evaluation* (Test Rep. No. TS-79-1). Washington, DC: Office of Personnel Management, Personnel Research and Development Center.

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

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

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

*Ployhart, R. E., & Ryan, A. M. (2000). *A construct-oriented approach for developing situational judgment tests in a service context.* Unpublished manuscript.

*Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9,* 241–258.

*Richardson, Bellows, Henry, & Co., Inc. (1949). *Test of Supervisory Judgment: Form S.* Washington, DC: Richardson, Bellows, and Henry.

*Richardson, Bellows, Henry, & Co., Inc. (1963). *Test of Supervisory Judgment: Form T.* Washington, DC: Richardson, Bellows, and Henry.

*Richardson, Bellows, Henry, & Co., Inc. (1988a). *The Manager Profile Record.* Minneapolis, MN: National Computer Systems.

*Richardson, Bellows, Henry, & Co., Inc. (1988b). *The Supervisor Profile Record.* Minneapolis, MN: National Computer Systems.

Rosen, N. A. (1961). How Supervise?—1943–1960. *Personnel Psychology, 14,* 87–99.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75,* 322–327.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75,* 175–184.

*Rushmore, J. T. (1958). A note on the "Test of Practical Judgment." *Personnel Psychology, 11,* 37.

*Sartian, A. Q. (1946). Relation between scores on certain standard tests

and supervisory success in an aircraft factory. *Journal of Applied Psychology, 29,* 328–332.

*Schippman, J. S., & Prien, E. P. (1985). The Ghiselli Self-Description Inventory: A psychometric appraisal. *Psychological Reports, 57,* 1171–1177.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. A. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78,* 3–12.

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

*Spitzer, M. E., & McNamara, W. J. (1964). A managerial selection study. *Personnel Psychology, 17,* 19–40.

*Sternberg, R. J., Wagner, R. K., & Okagaski, L. (1993). Practical intelligence: The nature and role of tacit knowledge in work and at school. In J. M. Puckett & H. W. Reese (Eds.), *Mechanisms of everyday cognition* (pp. 205–227). Hillsdale, NJ: Erlbaum.

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

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

*Super, D. E. (1949). *Appraising vocational fitness by means of psychological tests.* New York: Harper.

Taylor, H. R. (1949a). Social Intelligence Test: George Washington University series. In O. K. Buros (Ed.), *The third mental measurements yearbook* (pp. 96–97). New Brunswick, NJ: Rutgers University Press.

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

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44,* 703–742.

Thorndike, R. L. (1941). Social Intelligence Test. In O. K. Buros (Ed.), *The 1940 mental measurements yearbook* (p. 1253). Highland Park, NJ: Mental Measurements Yearbook.

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

*Thumin, F. J., & Page, D. S. (1966). A comparative study of two tests of supervisory knowledge. *Psychological Reports, 18,* 535–538.

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

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

*Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 48,* 436–458.

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

*Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52,* 679–700.

*Weitz, J., & Nuckols, R. C. (1953). A validation study of How Supervise? *Journal of Applied Psychology, 36,* 301–303.

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

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

*Wiener, D. N. (1961). Evaluation of selection procedures for a management development program. *Journal of Counseling Psychology, 8,* 121–128.

Wiesner, W. H., & Cronshaw, S. F. (1988). The moderating impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61,* 275–290.

Received November 6, 1997
Revision received August 10, 2000
Accepted August 23, 2000 ■