Situational Judgment Tests: A Review of Practice and Constructs Assessed

Michael A. McDaniel and Nhung T. Nguyen*

In this article, we seek to summarize current practice concerning situational judgment tests in personnel selection. We begin by describing the manner in which situational judgment tests are developed and examining the diverse ways in which situational items are presented and scored. We then offer speculation concerning constructs assessed by situational judgment tests as well as discuss the legal aspect of situational judgment measures. We also review meta-analytic evidence concerning the construct validity of situational judgment tests and offer several new meta-analytic findings. Situational judgment tests are shown to be typically correlated moderately with general mental ability. Their primary personality correlates are emotional stability, conscientiousness, and agreeableness. Situational test scores also tend to increase with increasing years of job experience. The article concludes with a list of areas that need addressed in future research.

Introduction

Situational judgment tests are assessments designed to measure judgment in work settings. All such tests present the respondent with a situation and a list of possible responses to the situation. The respondent is asked to consider the situation and then make judgments concerning possible responses to the situation. Situational judgment tests may be classified as job simulations (Motowidlo, Hanson, and Crafts, 1997; Thornton and Cleveland, 1990). Simulations are based on the assumption that one can predict how well an individual may perform on a job based on how the individual performs on a simulation of the job. Recently, McDaniel, Morgeson, Finnegan, Campion, and Braverman (in press) examined the criterion-related validity of situational judgment tests and found that the tests have substantial validity (ρ = .34) for the prediction of job performance. They also found that the tests typically had a moderate correlation with general cognitive ability (r = .36) but that the magnitude of this correlation varied widely across tests. This article complements the McDaniel *et al.* (in press) effort, by describing the manner in which situational judgment test items are developed and examining the diverse ways in which items are presented and scored. Meta-analytical evidence concerning the construct validity of situational judgment tests is provided and suggestions for future research are discussed.

Development Procedures for Situational Judgment Items

Situational judgement items present work-related situations to respondents and request that the respondent evaluate several possible responses to the situation. Approaches to the development of these measures are described in Motowidlo *et al.* (1997) and various primary studies describing the development of specific measures (Motowidlo, Dunnette and Carter 1990; Smith and McDaniel 1998). Here, we present what we believe to be the most common procedure to develop these items.

To develop the situational judgment items it is common to obtain two sets of data from incumbents or other subject matter experts (Motowidlo *et al.* 1997). In the first wave of data, critical incidents (Anderson and Wilson 1997; Flannagan 1954) are collected from the subject matter experts. The critical incidents are stories about situations encountered on the job. Sometimes the subject matter experts are not provided with any particular guidance on the topic areas to be covered in the critical incidents. In other efforts, the subject matter experts are directed to write items targeted to certain competencies derived from a job analysis (Peterson and Jeanneret 1997). For example, for a customer service job, the respondents might be asked to write critical incidents concerning understanding customers' needs, promoting the product to customers, and seeking a balance between the needs of the customers and the company's interests.

The critical incidents are then reviewed by the test developer with the goal of identifying a set

*Address for correspondence: Michael A. McDaniel, Virginia Commonwealth University, Department of Management, P.O. Box 844000, Richmond, VA 23284-4000 E-mail: mikemcdaniel@vcu.org

of situation descriptions that will serve as the stems for the situational judgment items. Practice in this area varies widely. One issue that always needs addressed is what to do with very similar critical incidents. Motowidlo *et al.* (1997) suggested grouping the incidents into similar content areas and then selecting representative scenarios from each content area. For example, in critical incidents written for customer service jobs, it is common to obtain many incidents concerning handling difficult customers. One could review the incidents to identify all the difficult customer incidents and then select from the set of incidents those that appear to be representative of all incidents concerning difficult customers. Such a procedure allows one to tap a range of scenarios in a content domain without including nearly duplicate scenarios.

Another issue that always needs addressed is the editing of the critical incidents into situational stems. Typically the critical incidents are longer than the desired length of a situational stem. One seeks to edit the incidents into stems of similar length having a similar format. The situations may also be edited to make them more applicable to the full range of duties in the job. For example, if the situation references difficulty in mastering a specific piece of software used by some but not all individuals in the job, one might want to describe the software more generically so that the situation is applicable to all individuals in the job.

Situations may also be excluded from further consideration if the content of the situation raises legal concerns or perceptions of legal concerns. For example, some situations may reference physical activities such as driving a vehicle and such items may be viewed as unfair to individuals who cannot drive due to a disability. Other situations may be excluded because the topic of the situation (e.g., violence in the workplace) may be deemed inappropriate to present to job applicants.

The reviewed and edited situations are assembled into a survey, which is administered to a second set of individuals. The survey requests that the respondent identify one or more responses to a situation. Most test developers have the respondent identify what the respondent would most likely do or what the respondent believes is the best thing to do in a situation. If the respondent is asked to identify more than one response to a situation, the respondent might be asked to identify both the best response and a second response, which is reasonable but not optimal. The test developer then reviews all the offered responses to each situation and prepares an edited list of potential responses to each situation. Responses are edited to remove duplicate responses and to increase the comprehensibility of the response. Some responses may be discarded because the response may be deemed inappropriate to present to job applicants (e.g., respond to an interpersonal conflict by assaulting the offending employee). Typically, the test developer wants to have multiple responses to each situation and to have the responses span a range of effectiveness. The individuals who complete this survey may be subject matter experts or they may be relatively new and inexperienced employees. Subject matter experts are useful because they should be able to identify the best responses to the situation and based on their experience can generate some common responses that are less than optimal. Inexperienced employees are useful because they will offer responses with a wide range of effectiveness. We know of no data indicating which type of respondents is optimal.

In the article so far, we have addressed how items can be built but have provided little detail concerning the variety of ways that the situational items can be formatted and presented. Below we describe the characteristics of situational judgment items and then we will discuss methods of scoring the items.

Characteristics of Situational Judgment Item Stems and Responses

Situational judgment items can be divided into the item stem and item responses. The item stem is the portion of the item which presents the situation to the respondent. The item responses consist of a list of possible responses to the situation that are presented to the respondent for their evaluation. Both the item stems and the item responses can be categorized in a variety of ways.

Item stems can be distinguished along five characteristics. First, item stems can vary in their fidelity. Fidelity refers to the extent to which the format of the stem is consistent with how the situation would be encountered in a work setting. Some tests present the stems in the form of a short video that conveys the situation to the respondent. In other tests, the stems describe a situation in written format. It is reasonably argued (Lievens, Coetsier, and Decaesteker, 2000; Motowidlo *et al.* 1997) that video presentation of situations have a higher fidelity than the stems described in written form. Second, stems can vary in their length. Some tests present very short descriptions of situations (see *How Supervise?*, File and Remmers 1971). Other tests present very detailed descriptions of situations (see *Tacit Knowledge Inventory*, Wagner and Sternberg 1991). Third, situational stems

vary in their complexity. Some stems present simple scenarios (one has difficulty with a new assignment and needs instructions). Other stems present more complex scenarios (one has multiple supervisors, who are not cooperating with each other and providing conflicting instructions concerning which of your assignments has the highest priority).

We suspect that the complexity of the situation is related to the length of the scenario in that more words are required typically to describe complex situations than less complex situations. Fourth, situational stems may vary in their comprehensibility. It is harder to understand the meaning and import of some situations than other situations. Sacco, Schell, Ryan, Schmitt, Schmidt, and Rogg (2000) examined the comprehensibility of item stems using readability formulas. It is a reasonable conjecture that the length, complexity, and comprehensibility of the situation are interrelated and may drive the cognitive loading of the situational stems. Fifth, some tests present situational stems within other stems. For example, some situational items produced by the consulting firm Aon (Clevenger and Halland 2000; Parker, Golden, Russell and Redmond 2000) present an overall situation followed by subordinate situations. Responses need to be evaluated for the subordinate situations. Most situational judgment tests appear to have items where the situations are not tied to each other hierarchically.

Unlike the stems of situational items that can vary on a number of characteristics, the responses of situational items do not tend to vary much across situational judgment tests in that they are usually presented in a written format even if the items stems are presented through video vignettes (Lievens *et al.* 2000; Motowidlo *et al.* 1997). However, there is substantial variation in how respondents are instructed to evaluate the potential responses to a situation.

Some tests ask the respondent to identify the response they would most likely perform. A variant of this approach is to ask the respondent to identify the response they would most likely perform and those they would least likely perform (Dalessio 1994; Smith and McDaniel 1998). This variant of the instructions gives one twice as many responses to potentially score although the two responses to a given item are not independent. In using this most likely/least likely instruction set, the senior author found a 0.5 standard deviation difference between an applicant sample and a sample of individuals who took the test to assist in identifying training needs. The applicant sample scored higher. We speculate that this difference is due to the instruction set permitting faking by those applicants who wish to fake. For example, when faced with a situational stem concerning returning manuscripts to a journal editor, although one might generally be tardy in returning a manuscript to a journal editor, one could obtain a higher score on the situational judgment item by asserting that one would always be on time in returning manuscripts.

Whether it is an attempt to make the test more faking resistant or other reasons, many tests instruct the respondent to identify the best response to a situation. The variant of this instruction set is to ask the respondent to identify both the best and the worst response to the situation. We suspect that this instruction set makes the situational judgment test more faking resistant than the instruction set asking for the most likely response (or the most and least likely response). In a small sample study, the senior author compared three groups of respondents. One group completed the situational judgment test with instructions to identify the most and least likely responses. A second group received the same instructions but was asked to fake on the test to look good and was offered a monetary incentive to score well. A third group completed the test with instructions to identify the best and worst responses for each scenario. The faking group and the group with instructions to identify the best and worst responses had nearly identical mean scores, which were 0.5 standard deviation above the group who provided their most and least likely responses. We offer this as evidence that instructions asking for the respondent's most and least likely response permits score inflation through faking and that the instructions asking for the best and worst response results in a faking-resistant test in that scores cannot be improved when the applicants are motivated to fake. We are not arguing that this latter instruction set is totally immune to faking. Certainly, someone with the answer key could score very well on the test regardless of the instruction set. Likewise, effective coaching is likely to improve test scores regardless of instructions. Although no approach to reducing faking is likely to be entirely effective, the search for methods of reducing faking in situational judgment tests is an important one. For example, Reynolds, Sydell, Scott and Winter (2000) found higher validities for less fakable situational judgment items.

We suspect that the difference between a most likely/least likely response instruction set and a best/worst response instruction set results in different constructs being measured. In the best/worst instructions, the measures are more clearly tapping knowledge of how to respond. This would be true for both respondents who are answering honestly and for those respondents who are faking to look good. For the most likely/least likely response instruction set, honest

respondents are reporting their behavioral tendencies and respondents seeking to fake good are reporting their knowledge. We speculate that if all respondents were answering honestly, that the most likely/least likely response instruction set would yield higher validities than the best/worst response instruction set because the behavioral tendencies assessed by the former instruction set should better predict future behavior than the knowledge assessed by the latter instruction set. However, when applicants are faking to look good, both instruction sets should be measuring knowledge and the resulting validities should be the same for both response instruction sets.

Although asking respondents to identify the most and least likely alternative or the best and worst alternative can yield valid and useful instruments, it makes the item analysis of the test difficult. The process of asking the respondent to identify the best response from a list of options (or the best and worst, or most and least likely from a list of options) makes the item responses partially ipsative (Hicks 1970) which presents a host of problems for item and reliability analysis.

Perhaps to avoid the problem of partial ipsativity produced by the response instructions discussed above, some tests ask the respondent to rate the effectiveness of each response. The effectiveness of one response is not dependent on the effectiveness of another response so there is no ipsativity. In addition, instead of having one or two scores per item as is the case with instructions asking for the most favorable or the most and least favorable, one has as many scoreable items as there are responses. When respondents rate the effectiveness of each response, one can perform a variety of item analyses that would not make sense with the partially ipsative scoring procedures. Nonetheless, the interpretation of statistical analyses such as exploratory factor analysis, can be somewhat muddled, however, because the effectiveness of the response is a function of the scenario with which the response is tied. Thus the construct loading of two identical responses can be very different depending on the scenario with which the responses are associated. For example, asking one's supervisor for help with a problem might be associated with high cognitive ability if the scenario concerns difficulty with an assignment whereas it might be associated with low cognitive ability if the scenario concerns being sexually harassed by the same supervisor.

Scoring of Situational Judgment Tests

Just as there are a variety of ways to present situational judgment stems and responses, there are a variety of ways to score these items. Typically, the situational judgment answer key is developed judgmentally using a pool of individuals purported to be subject matter experts or excellent employees. These individuals make judgments concerning the effectiveness of the various item responses and these judgments are pooled subsequently either using consensus or actuarial methods. Responses where the experts fail to show substantial agreement concerning the effectiveness of the response should be dropped (Motowidlo *et al.* 1997). The second scoring option involves collecting responses to the surveys and using central tendency statistics to determine which responses are effective and which are less effective. The third scoring option is to employ empirical-scoring approaches similar to those used in developing scoring keys for biodata (Mumford and Whetzel 1997). Research in this area would benefit from examining the literature of rationale versus empirical scoring of biodata inventories. The one study we found seeking to compare judgmental versus empirical scoring of the same instrument had an insufficient sample size to address the question adequately (Parker *et al.* 2000). What is sometimes found in the biodata literature but yet to be seen in the situational judgment literature is an effort to build construct homogeneous keys (Mumford and Whetzel 1997), for example, scoring a situational judgment test to yield a scale score for conscientiousness or for general cognitive ability. Such keys would be difficult to build for situational judgment items because situational items tend to be construct heterogeneous. For example, the selection of a given response as effective might reflect both the conscientiousness and general cognitive ability of the respondent.

What Do Situational Judgment Tests Measure?

There is substantial debate concerning what situational judgment tests measure. Although not acknowledging that their tacit knowledge measures assess situational judgment, Sternberg and Wagner (1993) contended that their tacit knowledge measures assess 'practical know-how that usually is not openly expressed or stated and which must be acquired in the absence of direct instruction' (Wagner 1987: 1236). Schmidt and Hunter (1993) responded that there is nothing tacit about tacit knowledge and argue that situational judgment tests are simply measures of job knowledge.

Most of the debate concerning what situational judgment tests measure have an implicit assumption that there is a single situational judgment construct and the studies seek to understand this unitary construct. We

think this reasoning is misguided. We concur with several recent authors (Chan and Schmitt 1997; McDaniel *et al*., in press; Weekly and Jones 1999) who argue that situational judgment tests are measurement methods. Like other measurement methods, such as the employment interview or job knowledge testing, situational judgment tests can be built to measure a variety of constructs. To assess interpersonal constructs, one can build a test with many interpersonal situations. Alternatively, one can build a test where conscientiousness is a major determinant of individual differences in item responding. Or one can build a test which is primarily a measure of cognitive ability. There are, however, limits to what a situational judgment test can or cannot measure.

We suggest that it is reasonable for any measure assessing judgment to have some correlation with general cognitive ability. McDaniel *et al*. (in press) found a mean observed correlation of .36. The population level correlation, which was corrected for measurement error in both the situational judgment test and the measure of general cognitive ability, was .46 with a credibility interval of .17 to .75. Thus it appears that there are boundaries concerning the extent to which a situational judgment test will correlate with general cognitive ability such that it would be unlikely that a situational judgment test can be entirely unrelated to general cognitive ability. However, it appears that one can build a situational judgment test where most of the test's reliable variance taps general cognitive ability.

There is much less data on the non-ability test correlates of situational judgment tests. As with general cognitive ability and situational judgment tests, one should expect substantial systematic variability in the extent to which the non-ability test correlates with situational judgment. However, one would also expect some non-ability to have consistently non-zero correlations with situational judgment tests.

We suggest that measures of job knowledge, usually operationalized as measures of job experience, should have positive correlates with situational judgment measures (Clevenger and Haaland, 2000). Larger correlations can be expected where the sample has variance in experience and where job experience and not other potential sources of job knowledge such as formal education are the primary determinants of job knowledge. Schmidt, Hunter, and Outerbridge (1986) and McDaniel, Schmidt, and Hunter (1988) have argued that individual differences in experience in the early years of job experience have greater relations to job knowledge than individual differences in experience in the later years of job experience. Thus, for example, it is argued that the difference in job knowledge between a person with one year of experience and six years of experience is much larger than the difference in job knowledge between a person with 11 years of job experience and 16 years of job experience. This occurs because one learns most of the job knowledge needed for job performance in the early years of experience and each additional year of job experience contributes less and less job knowledge. Following this reasoning, we speculate that the correlations between job experience and situational judgment tests will be larger when the sample is composed of relatively inexperienced individuals and will be smaller when the sample is composed of those with substantial amounts of job experience.

There are a small number of studies reporting correlations between situational judgment tests, job experience, and personality measures, on which we conducted a 'bare-bones' meta-analysis (Hunter and Schmidt 1990). In such a meta-analysis, sampling error correlations are the sole correction made. The meta-analysis results of these situational judgment tests correlates are shown in Table 1.

As shown in Table 1, job experience was found to have a small positive correlation with situational judgment tests among all the correlates examined (mean $r = .05$, $k = 18$, $N = 7{,}762$). Given the large sample size and the negative correlation between job experience and situational judgment measures in one study included in the analysis (Clevenger and Haaland, 2000) (see the Appendix for a list of studies grouped by construct and magnitude of relationship with situational judgment measures), we reported two separate results, one with and without that study. When Clevenger and Haaland (2000) were excluded from the analysis, the corrected mean observed correlation was found to be (mean $r = .07$, $k = 17$, $N = 6{,}260$). Although this effect size is small, it provides evidence that situational judgment tests are measures of job knowledge (Schmidt and Hunter 1993). Thus, situational judgment tests may owe some of their criterion-related validity due to their assessment of job knowledge (Dye, Reck and McDaniel 1993). We speculate that the correlations between situational judgment tests with job knowledge are probably larger than that of job experience because job experience is a less than perfect measure of job knowledge. McDaniel *et al*. (1988) argued that job experience is asymptotically related to job knowledge and we speculate that personal, occupational, and organizational influences would likely be responsible for how much job knowledge one gains in a fixed period of experience. Thus, as noted earlier, we would expect correlates with job experience to be stronger when the applicant pool has relatively

Table 1. Meta-analysis results for correlates of situational judgment tests

Scale	*k*	N	$\bar{r}$	σ_r	σ_{res}	% var. explained	95% CI
Agreeableness							
All correlations	12	12,855	.25	.18	.18	2.50	−.10 to .60
Without Leaman and Vasilopoulos (1998)	11	8,483	.13	.07	.06	2.40	.00 to .25
Conscientiousness							
All correlations	13	13,600	.26	.14	.14	4.50	.00 to .52
Without Leaman and Vasilopoulos (1998)	12	9,228	.17	.08	.08	17.60	.03 to .32
Emotional stability	11	7,482	.31	.20	.19	3.10	−.07 to .69
Extroversion	8	2,555	.06	.12	.10	24.12	−.12 to .21
Openness	3	814	.09	.02	.00	698.0	.09 to .09
Experience							
All correlations	18	7,762	.05	13	.12	14.16	−.18 to .28
Without Clevenger and Haaland (2000)	17	6,260	.07	.12	.11	17.75	−.14 to .29

Note: k = Number of studies included in the analysis, N = cases summed across studies, r = mean observed effect size, σ_r = standard deviation of the mean observed effect size, σ_{res} = residual standard deviation of the observed mean effect size, % var. explained = percentage of observed variance explained by sampling error, 95% CI = observed *r*'s 95% credibility interval.

low levels of experience as we discussed earlier in the paper.

Within the Big Five personality dimensions, emotional stability was found to have the highest correlation with situational judgment tests (mean r = .31, k = 11, N = 7482). Agreeableness and conscientiousness were also found to have non-trivial correlations with situational judgment measures (mean r = .25 and .26, respectively). This finding provides indirect evidence that situational judgment tests predict job performance because the above three personality dimensions were shown to be valid predictors of job performance across job domains (Barrick and Mount 1991). The fact that agreeableness was found to have a smaller effect size than that of conscientiousness and emotional stability with situational judgment tests is consistent with earlier research showing agreeableness having less consistent correlation with job performance than conscientiousness and emotional stability (Barrick and Mount 1991; Salgado 1998). We reported two separate results for agreeableness and conscientiousness because one of the studies included in the meta-analysis (Leaman and Vasilopoulos 1998) reported substantially large correlations between agreeableness as well as conscientiousness and situational judgment tests (r = .49 and r = .43, N = 4372, respectively). For agreeableness, the large sample size of this study and the large magnitude relationship reported was responsible for boosting the overall mean corrected correlation when it was included in the analysis (mean r = .25, k = 12, N = 12855) as compared to when it was not (mean r = .13, k = 11, N = 8483). For conscientiousness, the mean corrected observed effect size with the Leaman and Vasilopoulos' (1998) study was (mean r = .26, k = 13, N = 13600) and pulled down to (mean r = .17, k = 12, N = 9228) when the above study was excluded from the analysis.

It is important to note that the credibility intervals of the observed mean correlates with situational judgment reported here are wide and that sampling error did not account for much of the observed variance. The large credibility intervals provide evidence for the existence of moderators, which can best be represented by the diversity of the situational judgment tests. The small number of studies included in the analysis prevented us from conducting a moderator sub-setting analysis because of low power (Hunter and Schmidt 1990). However, our present meta-analytic findings reinforced our contention that situational judgment tests are best viewed as measurement methods with which one can assess a wide variety of content or constructs. As more data accumulate, future research should examine the moderating effect of the characteristics of the situational judgment measure such as the test development method used in examining the correlates of situational judgment tests. Also, our results reported here should be considered as lower bound because we did not correct for any measurement errors (e.g., reliability of the situational judgment test and the personality variables, range restriction, etc.). Thus, users of situational judgment tests would

be advised that the population correlates of situational judgment tests would be larger than what is shown in this article.

From the findings of this meta-analysis, we have evidence to believe that situational judgment tests appear to capture assorted construct variance with general mental ability being the largest correlate (McDaniel *et al.*, in press) and emotional stability, conscientiousness, and agreeableness being the personality correlates of largest magnitude.

Legal Issues/Concerns in Situational Judgment Tests

Although evidence exists showing that situational judgment tests have less racial adverse impact than do cognitive ability tests (Chan and Schmitt 1997; Motowidlo and Tippins 1993; Weekly and Jones 1999), in an initial case, a situational judgment test was considered lacking the legal standard of content validity by the US courts. In the sole court case (Jerome Green vs. Washington State Patrol 1997), it was ruled that the situational judgment test did not meet the requirements for content validity. Users of situational judgment tests in the United States would be advised to gather criterion-related validity evidence to document the validity of the measures.

Weekly and Jones (1999) reported that females scored higher than males on situational judgment measures, which might be considered ground for discrimination based on sex by the US courts. To date, we know of no court cases concerning sex discrimination in situational judgment tests. However, the reported sex differences in situational judgment measures deserve more research for us to have a better understanding of the nature of male-female differences in situational judgment tests.

Directions for Future Research

Although situational judgment tests have been in existence for decades (McDaniel *et al.*, in press; Motowidlo *et al.* 1997), research on the measures, other than traditional validity studies, has largely been a product of the last ten years. Here, we offer our thoughts on useful avenues for research concerning situational judgment tests.

Development of Methodologies for Targeting Specific Constructs

Situational judgment tests are inherently construct heterogenous methods. One's judgement of the effectiveness of a potential response to a work situation is likely the result of many individual difference variables including cognitive ability, job knowledge, and personality factors (Motowidlo, Borman and Schmit 1997). Job analyses frequently drive the content of the situations included in a test by identifying competencies needed for the jobs. The competencies are often expressed in terms of job duties (e.g., ability to promote products to a customer) and the successful performance of these job duties is likely the result of many individual difference variables. The inability to target specific individual difference constructs is a major limitation of the current technologies for developing situational judgment measures. Ployhart and Ryan (2000) identified several problems that arise when one does not know the constructs assessed by a situational judgment measure. These include the difficulty in professionally and legally defending the test, the limited ability of such tests to further understanding the predictor-criterion relationship and the inability to improve tests based on knowledge of their content. New technologies need to be developed for better-targeted situational judgment tests to assess constructs of interest. Weekly and Jones (1999) and Ployhart and Ryan (2000) have offered some suggestions in this area.

Determine How Item Characteristics Influence Validity

We have documented that situational judgment items can vary widely across tests. Very little is known concerning the relationships between these item characteristics and the validity of the items. We see various camps of researchers and practitioners, some preferring item types of one format and other preferring other types. More systematic research is needed to determine what item characteristics influence validity.

Determine the Extent to which Situational Tests can be Faking Resistant

In the non-cognitive literature, there is currently a rancorous debate concerning the extent to which personality and other non-cognitive tests are faked by applicants desiring to look better than they are and the extent to which such faking harms the usefulness of such tests (Douglas, McDaniel, and Snell 1996; Ones and Viswesvaran 1998; Ones, Viswesvaran, and Reiss 1996; Snell, Sydell, and Lueke 1999). In this article we speculated concerning whether certain response instructions can result in faking-resistant measures. To the extent that one can assess specific constructs using situational judgment tests, it may be possible to build faking-resistant measures of non-cognitive constructs.

Determine Item Characteristics that Influence Adverse Impact

It is often found that situational judgment tests show less racial adverse impact than do cognitive tests (Chan and Schmitt 1997; Motowidlo and Tippins 1993; Weekly and Jones 1999). Some of the adverse impact can be attributed to the extent to which situational judgment tests measure general cognitive ability and some of the reduced adverse impact of situational judgment tests relative to cognitive tests can be related to the systematic non-cognitive variance in situational judgment measures. However, there are a variety of possible reasons why situational judgment tests have less adverse impact and their exploration could further our understanding of ways to increase or maintain validity while reducing race-based adverse impact. Weekly and Jones (1999) have summarized the literature on sex-based adverse impact for situational judgment measures. Females often score better than males on situational judgment tests. Although the sex effect size is usually small, it raises questions concerning the source of the sex difference and its meaning for understanding male-female differences and the construct validity of situational judgment tests. Thus, sex differences should also be examined in future research.

Conclusion

Although situational judgment tests have been used for many decades (McDaniel *et al*., in press; Motowidlo *et al*. 1997), our knowledge base is relatively small. We know that the tests come in many formats but we know little about the extent to which formats influence validity and adverse impact. We know that these tests generally correlate with general cognitive ability and other factors but know little concerning how to build the tests to assess the constructs we wish to measure. Thus, there are a host of research and practical issues to be addressed. Given the growing interest in these measures, we anticipate an increased amount of research concerning situational judgment tests in the years ahead.

Acknowledgements

This article has benefited from the feedback provided by Filip Lievens and the Editor.

Appendix: List of studies included in the meta-analysis grouped by construct

Author	N	r
Agreeableness		
Leaman and Vasilopoulos (1998)	4372	0.49
Watley and Martin (1962)	62	0.23
Mullins and Schmitt (1998)	348	0.22
Ployhart and Ryan (2000)	208	0.21
Pereira and Harvey (1999)	233	0.16
Pereira and Harvey (1999)	5586	0.16
Leaman *et al.* (1996)	134	0.13
Smith and McDaniel (1998)	168	0.1
Jones, Dwight and Nouryan (1999)	298	0.07
Lobsenz and Morris (1999)	100	0.05
Bruce (1965)	62	0.04
Clevenger, Jockin and Morris (1994)	1284	−0.03
Conscientiousness		
Leaman and Vasilopoulos (1998)	4372	0.43
Smith and McDaniel (1998)	168	0.32
Mullins and Schmitt (1998)	348	0.26
Ployhart and Ryan (2000)	208	0.22
Pereira and Harvey (1999)	5586	0.22
Pereira and Harvey (1999)	233	0.19
Bosshardt and Cochran (1996)	284	0.11
Clevenger, Jockin and Morris (1994)	1284	0.09
Leaman *et al.* (1996)	134	0.07

Author	N	r
Bosshardt and Cochran (1996)	284	0.04
Jones, Dwight and Nouryan (1999)	298	0.02
Schippmann and Prien (1985)	301	−0.05
Lobsenz and Morris (1999)	100	−0.1
Emotional Stability		
Leaman and Vasilopoulos (1998)	4372	0.47
Smith and McDaniel (1998)	168	0.22
Pereira and Harvey 1999)	233	0.19
Ployhart and Ryan (2000)	208	0.16
Mullins and Schmitt (1998)	348	0.16
Carrington (1949)	313	0.15
Leaman *et al.* (1996)	134	0.12
Jones, Dwight and Nouryan (1999)	298	0.04
Clevenger, Jockin and Morris (1994)	1284	0.03
Watley and Martin (1962)	62	−0.03
Bruce (1965)	62	−0.25
Extroversion		
Bruce (1965)	62	0.36
Leaman and Vasilopoulos (1997)	176	0.3
Mullins and Schmitt (1998)	348	0.2
Pereira and Harvey (1999)	233	0.08
Lobsenz and Morris (1999)	100	0.03
Clevenger, Jockin and Morris (1994)	1284	0.03
Jones, Dwight and Nouryan (1999)	298	−0.01
Smith and McDaniel (1998)	168	−0.09
Watley and Martin (1962)	62	−0.3
Job experience		
Wagner and Sternberg (1991)	64	0.3
Smith and McDaniel (1998)	212	0.27
Weekley and Jones (1999)	844	0.26
Thumin and Page (1966)	55	0.22
Wagner and Sternberg (1985)	54	0.21
Weekley and Jones (1997) - Sample 2 b	198	0.2
Weekley and Jones (1997) - Sample 1 b	684	0.16
Weekley and Jones (1999) - Sample 2	1040	0.16
Weekley and Jones (1997) - Sample 2 a	412	0.14
Weekley and Jones (1997) - Sample 1 a	787	0.13
Mullins and Schmitt (1998)	348	0.1
Bosshardt and Cochran (1996)	284	0.09
Weekley and Jones (1997) - Sample 2 a	412	0.09
Thumin and Page (1966)	55	0.06
Jones, Dwight and Nouryan (1999)	298	0.02
Weekley and Jones (1997) - Sample 2 b	198	0
Clevenger and Haaland (2000)	1502	−0.04
Patton (1954)	315	−0.23
Openness		
Mullins and Schmitt (1998)	348	0.11
Smith and McDaniel (1998)	168	0.1
Jones, Dwight and Nouryan (1999)	298	0.06

References

Note: * indicates studies included in the meta-analysis.

Anderson, L. and Wilson, S. (1997) Critical incident technique. In D.L. Whetzel and G.R. Wheaton (eds.), *Applied Measurement Methods in Industrial Psychology*. Palo Alto, CA: Davies-Black.

Barrick, M.R. and Mount, M.K. (1991) The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, **44**, 1–26.

*Bosshardt, M.J. and Cochran, C.C. (1996) *Development and Validation of a Selection System for Financial Advisors* (Technical Report No 276). Minneapolis: Personnel Decision Research Institute, Inc.

*Bruce, M.M. (1965) *Examiner's Manual Business Judgment test*. Author.

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

Chan, D. and 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. and Haaland, D.E. (2000) Examining the relationship between job knowledge and situational judgment performance. Paper presented at the 15th annual conference of the Society of Industrial and Organizational Psychology, New Orleans, April.

*Clevenger, J.P., Jockin, T., Morris, S. and Anselmi, J. (1999) 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 of Industrial and Organizational Psychology, Atlanta, April.

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

Douglas, E.F., McDaniel, M.A. and Snell, A.F. (1996) The validity of non-cognitive measures decays when applicants fake. *Proceedings of the Academy of Management*, August.

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

File, Q.W. and Remmers, H.H. (1971) *How Supervise? Manual 1971 Revision*. Cleveland, OH: The Psychological Corporation.

Flanagan, J.C. (1954) The critical incident technique. *Psychological Bulletin*, **41**, 237–358.

Hicks, L.E. (1970) Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, **74**, 167–184.

Hunter, J.E. and Schmidt, F.L. (1990) *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage Publications.

*Jones, M.W., Dwight, S.A. and Nouryan, T.R. (1999) 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, April.

*Leaman, J.A. and Vasilopoulos, N.L. (1998) *Development and Validation of the detention enforcement officer applicant assessment.* Report 911. Research and Development Division, Office of Human Resources and Development. Washington, DC: US Immigration and Naturalization Service.

*Leaman, J.A., Vasilopoulos, N.L. and Usala, P.D. (1996) Beyond integrity testing: Screening Border Patrol applicants for counterproductive behaviors. Paper presented at the 104th annual convention of the American Psychological Association.

Lievens, F., Coestsier, P., Decaesteker, C. (2000) Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. Manuscript under review.

*Lobsenz, R.E. and Morris, S.B. (April, 1999). *Is tacit knowledge distinct from g, personality, and social knowledge?* Poster presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta: GA.

McDaniel, M.A., Schmidt, F.L. and Hunter, J.E. (1988) Job experience correlates of job performance. *Journal of Applied Psychology*, **73**, 327–330.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A. and Braverman, E.P. (in press) Predicting job performance from common sense. *Journal of Applied Psychology*.

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

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

Motowidlo, S.J., Hanson, M.A. and Crafts, J.L. (1997) Low-fidelity simulations. In D.L. Whetzel and G.R. Wheaton (eds.), *Applied Measurement Methods in Industrial Psychology*. Palo Alto, CA: Davies-Black.

Motowidlo, S.J. and 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.

*Mullins, M.E. and Schmitt, N. (1998) 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, Texas.

Mumford, M.D. and Whetzel, D.L. (1997) Background data. In D.L. Whetzel and G.R. Wheaton (eds.), *Applied Measurement Methods in Industrial Psychology*. Palo Alto, CA: Davies-Black.

Ones, D.S. and Viswesvaran, C. (1998) The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, **11**, 145–169.

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

Parker, C.W., Golden III, J.H., Russell, D.P. and Redmond, M.R. (2000) The development of a construct-related scoring key of a situational judgment inventory for enhancing criterion-related validity. Paper presented at the 15th

annual conference of the Society of Industrial and Organizational Psychology, New Orleans, April.
Patton, W.M. (1954) Studies in industrial empathy: A study of supervisory empathy in the textile industry. *Journal of Applied Psychology*, **38**, 285–288.
*Pereira, G.M. and Harvey, V.S. (1999) 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, April.
Peterson, N. and Jeanneret, R.P. (1997) Job analysis: Overview and description of deductive methods. In D.L. Whetzel and G.R. Wheaton (eds.), *Applied Measurement Methods in Industrial Psychology*. Palo Alto, CA: Davies-Black.
*Ployhart, R.E. and Ryan, A.M. (2000) A construct-oriented approach for developing situational judgment tests in a service context. Manuscript submitted for publication.
Reynolds, D.H., Sydell, E.J., Scott, D.R. and Winter, J.L. (2000) Factors affecting situational judgment test characteristics. Paper presented at the 15th annual conference of the Society of Industrial and Organizational Psychology, New Orleans, April.
Sacco, J.M., Scheu, C.R., Ryan, A.M., Schmitt, N., Schmidt, D.B. and Rogg, K.L. (2000) *Reading level and verbal test scores as predictions of subgroup differences and validities of situational judgment tests*. Paper presented at the 15th Annual Meeting of the Society of Industrial and Organizational Psychology, New Orleans: LA.
Salgado, J.F. (1998) Big Five personality dimensions and job performance in army and civil occupations: A European perspective. *Human Performance*, **11**, 271–288.
*Schippmann, J.S. and Prien, E.P. (1985) The Ghiselli self-description inventory: A psychometric appraisal. *Psychological Reports*, **57**, 1171–1177.
Schmidt, F.L. and Hunter, J.E. (1993) Tacit knowledge, practical intelligence, general mental ability and job knowledge. *Current Directions in Psychological Science*, **2**, 8–9.
Schmidt, F.L., Hunter, J.E. and Outerbridge, A.N. (1986) Impact of job experience and ability on job knowledge, work sample performance and supervisory ratings of job performance. *Journal of Applied Psychology*, **71**, 432–439.
*Smith, K.C. and McDaniel, M.A. (1998) 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., Texas, April.
Snell, A.F., Sydell, E.J. and Lueke, S.B. (1999) Towards a theory of applicant faking: Integrating studies of perception. *Human Resource Management Review*, **9**, 219–242.
Sternberg, R.J. and Wagner, R.K. (1993) The g-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science*, **2**, 1–12.
Thornton, G.C. and Cleveland, J.N. (1990) Developing managerial talent through simulation. *American Psychologist*, **45**, 190–199.
*Thumin, F.J. and Page, D.S. (1966) A comparative study of two tests of supervisory knowledge. *Psychological Reports*, **18**, 535–538.
Wagner, R.K. (1987) Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology*, **52**, 1236–1247.
*Wagner, R.K. and Sternberg, R.J. (1985) Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology*, **49**, 436–458.
*Wagner, R.K. and Sternberg, R.J. (1991) *Tacit Knowledge Inventory for Managers: User Manual*. San Antonio, TX: The Psychological Corporation.
*Watley, D.J. and Martin, H.T. (1962) Prediction of academic success in a college of business administration. *Personnel and Guidance Journal*, **41**, 147–154.
*Weekley, J.A. and Jones, C. (1999) Further studies of situational tests. *Personnel Psychology*, **52**, 679–700.