Incremental Validity of Situational Judgment Tests for Task and Contextual Job Performance

Matthew S. O'Connell*, Nathan S. Hartman, Michael A. McDaniel***, Walter Lee Grubb III*** and Amie Lawrence***

*Select International Inc., 5700 Corporate Dr., Suite 250, Pittsburgh, PA 15237, USA. moconnell@selectintl.com
**John Carroll University, 20700 North Park Blvd, University Heights, OH 44118, USA
***Virginia Commonwealth University, 1015 Floyd Ave., Box 844000, Richmond, VA 23284, USA

This paper has three goals. First, it responds to calls for additional research on subgroup differences in situational judgment tests. Second, it expands the cumulative knowledge on the incremental validity of situational judgment tests beyond cognitive ability and personality. Third, it examines the validity and incremental validity of various predictors for both task and contextual performance.

1. Introduction

Situational judgment tests (SJT) are simulations requiring the respondent to exercise judgment when responding to hypothetical problem situations that occur in work settings. The use of SJTs dates back to the 1920s (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Procedures for developing this type of test item are discussed in several studies (Motowidlo, Dunnette, & Carter, 1990; McDaniel & Nguyen, 2001; Smith & McDaniel, 1998). The following item from a World War II era Army judgment test is illustrative of a SJT presented in written form (Northrop, 1989, p. 190):

> 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.

SJTs have become popular measures for gathering respondent's knowledge of how to handle particular situations and/or their behavioral tendencies in these situations. The popularity of these instruments has led to research investigating the nature of the construct(s) measured by these items.

1.1. What are SJTs measuring?

Some authors of SJTs have asserted that their tests measure a single construct such as practical judgment (Cardall, 1942), managerial success (Campbell, Dunnette, Lawler, & Weick, 1970), and tacit knowledge

9600 Garsington Road, Oxford, OX4 2DQ, UK and 350 Main St., Malden, MA, 02148, USA

(Sternberg & Wagner, 1993). On the other hand, three recent meta-analyses (McDaniel *et al*., 2001; McDaniel & Nguyen, 2001; Nguyen & McDaniel, 2001) have shown that situational judgment measures typically measure several well-established constructs including cognitive ability, conscientiousness, emotional stability, and agreeableness. In addition to these constructs, job knowledge appears to have a relationship with SJTs as suggested by the modest correlations between job experience and SJTs. Although evidence supports the consistent relationship between these constructs and SJTs, the magnitude of correlations across studies varies substantially, even after correcting for artifacts. For example, although most SJTs have moderate correlations with cognitive ability, any given test might show a very large correlation with cognitive ability, or a very low correlation with cognitive ability. As SJTs measure a variety of constructs and different tests assess these constructs to varying degrees, we join others in arguing that SJTs are best viewed as measurement methods and not measures of a single construct (Chan & Schmitt, 1997; Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; McDaniel & Nguyen, 2001; McDaniel *et al*., 2001; Nguyen & McDaniel, 2001; Weekly & Jones, 1999).

1.2. Validity and subgroup differences of SJTs

SJTs have gained increasing popularity in recent years. This popularity has been driven both by the validity of the tests (McDaniel *et al*., 2001) and by findings of smaller mean differences among racial subgroups as compared with traditional cognitive ability tests (Motowidlo *et al*., 1990; Motowidlo & Tippens, 1993; Pulakos & Schmitt, 1996; Clevenger *et al*., 2001). Nguyen, McDaniel, and Whetzel (2005) have meta-analytically summarized the magnitude of mean racial and gender differences in SJTs. Black/White mean differences average .38 standard deviations favoring Whites. Gender differences average .10 favoring females. However, the mean differences cannot be readily interpreted because the magnitude of the mean racial differences are strongly moderated by the correlation between the SJT and measures of cognitive ability. More cognitively loaded SJTs (i.e., SJTs with high correlations with cognitive ability) show greater mean racial differences than less cognitively loaded SJTs. The gender differences are moderated by the personality correlates of the SJTs. Nguyen, Biderman, and McDaniel (2005) called for substantially more research and race and sex differences in SJTs so that the magnitude of subgroup differences can be accurately estimated within moderator subgroups.

As the validity of SJTs has been established (McDaniel *et al*., 2001), it is important to understand how these tests add utility to a selection battery by further understanding the relationship between SJTs and other constructs, such as personality and cognitive ability. Specifically, it is important to identify the incremental validity that can be attributed to these types of instruments over other established predictors used in employee selection (Chan & Schmitt, 2002). In five samples, Weekly and Jones (1997, 1999) found significant incremental validity for SJTs over cognitive ability and job experience. Clevenger *et al*. (2001) used three samples to examine the incremental validity of situational judgment measures over cognitive ability, conscientiousness, job experience, and job knowledge and reported incremental validity in two of the three samples. Chan and Schmitt (2002) found a SJT to have substantial validity in predicting task performance, and overall job performance. However, the SJT used in this study reported an unusually small correlation with cognitive ability ($r = -.02$). Meta-analytic research by McDaniel *et al*. (2001) examined the estimates of the correlation between situational judgment, cognitive ability, and job performance and offered the conclusion that SJTs may show incremental validity. Their conclusion was tempered by the fact that the correlations between cognitive ability and SJTs varied widely and that the validity of both cognitive and situational tests were affected by moderators. In summarizing this literature, McDaniel, Whetzel, Hartman, Nguyen, and Grubb (2006) concluded that SJTs typically show incremental validity over cognitive ability tests and called for more research addressing incremental validity over both cognitive ability and personality tests.

1.3. Multidimensional job performance

When investigating predictors of job performance, it is important to address the multidimensional nature of job performance, which suggests that job performance is comprised of two major dimensions: task and contextual performance (Borman & Motowidlo, 1993). Task performance refers to expected work behaviors that are required to perform the job successfully. Contextual performance refers to extra-role behaviors or organizational citizenship behaviors, things that are not dictated by job requirements and act to benefit the organization. Research in this area has concluded that certain constructs are differentially related to the two dimensions of job performance. Specifically, cognitive ability is most strongly related to task performance, and non-cognitive measures, such as conscientiousness, add incremental value above cognitive ability measures to the prediction of contextual performance (Borman & Motowidlo, 1993; Hattrup, O'Connell, & Wingate, 1998). With the exception of Chan and Schmitt (2002), previous research has not addressed the extent to which SJTs relate to task or contextual performance domains.

1.4. Goals of the current research

This paper addresses three goals. First, it responds to the call by Nguyen *et al.* (2005) for more research on mean racial and sex differences on SJTs. Consistent with their findings of cognitive and personality moderators of subgroup differences, our results also provide SJT correlates with cognitive and personality variables. Second, the paper is responsive to the McDaniel *et al.* (2006) call for more research on the incremental validity of SJTs in combination with both cognitive and personality predictors. Third, the paper extends the contribution of Chan and Schmitt (2002) by examining the validity of SJTs for both task and contextual performance.

2. Method

2.1. Participants

Analyses of data drawn from seven different organizations using the same test battery are reported. Test data were collected concurrently with the criterion measures. All organizations were manufacturing companies, including: two heavy truck manufacturers ($N=72$ and 69), one truck engine manufacturer ($N=92$), one custom engineered materials manufacturer ($N=43$), one fiberglass and flat glass manufacturer ($N=461$), one electronics/communications manufacturer ($N=226$), and one television manufacturer ($N=177$). All participants were entry-level assembly or manufacturing employees. All data were collected in the United States between 1993 and 1999. Across the organizations, there were 757 males, 361 females, and 22 individuals for which gender was not known. About 20% of the sample was Black ($N=221$) and 65% was White ($N=735$). The remaining sample members were Asian, Hispanic, Native American or individuals for whom race is not known.

2.2. Measures

The tests used in the study were all part of a computer-based assessment system designed for use in a broad range of manufacturing occupations. Descriptions of parts of this computer-based assessment system, the Select Assessment for Manufacturing® have appeared elsewhere in the literature (Hattrup, O'Connell, & Labrador, 2005; Bott, O'Connell, Ramakrishnan, & Doverspike, 2007; O'Connell, Doverspike, Gillikin, & Meloun, 2001). All participants completed the same assessment battery. The test battery consisted of a cognitive test, several personality tests, and a SJT.

2.2.1. Cognitive ability

A 35-item computer-administered cognitive ability measure was used. This scale, or variants of it, has appeared elsewhere in the literature (cf. Hattrup *et al.*, 2005; Cober, Cober, Lawrence, & O'Connell, 2003; Bott *et al.*, 2007). The measure is a composite of four reasoning subtests.

2.2.2. Personality

The personality items yielded five scale scores: conscientiousness, agreeableness, attention to detail, locus of control, and positive affectivity. All personality items were single statements to which the respondent used a sliding pointer to indicate agreement or disagreement. These items were continuous variables with a range from 1 to 5, strongly disagree to strongly agree. Based on a normative sample of over 3000, individual scale reliabilities ranged from .65 to .88 (O'Connell & Kato, 2001). These personality scales have been used in other studies and more detailed descriptions can be found in those studies, (cf. O'Connell & Smith, 1999, 2000; Hattrup *et al.*, 2005; Bott *et al.*, 2007).

2.2.3. Situational judgment test

The SJT consisted of 10 interpersonal scenarios that might be encountered by employees working in manufacturing, warehouse, or assembly environments. It was designed to evaluate candidates' skills on working effectively with others (e.g., giving advice, diffusing an interpersonal dispute, and demonstrating empathy). Participants were asked to rate the effectiveness of each of four possible alternatives to each situation on a five-point Likert type scale. Based on results from over 3000 individuals, the 10-item scale had an internal consistency reliability of .72 (O'Connell & Kato, 2001). Because SJTs are construct heterogeneous, this reliability is an underestimate.

2.2.4. Job performance

Supervisory ratings, collected for research purposes only, were used as criteria. Some organizations used a long form of the rating scale and other organizations used a shorter form of the same rating scale. Both scales contained 12 common items: six items assessed task performance and six items assessed contextual performance. These items were used to form two criterion scales: task performance and contextual performance. The contextual scale was composed of items measuring leadership, teamwork, and positive attitude, consistent with the interpersonal facilitation dimension of contextual performance described by Van Scotter and Motowidlo (1996), as well as items related to conscientiousness, initiative, and work motivation, which are consistent with the job dedication dimension of performance described by Van Scotter and Motowidlo (1996). The task performance scale included technical knowledge and problem solving items, which are consistent with the dimension of task performance (Borman & Motowidlo, 1993; Motowidlo *et al.*, 1997).

2.3. Data analysis

Means, standard deviations, reliabilities, and a full correlation matrix were calculated. Differential prediction analyses by race were conducted using the Cleary (1968) model and implemented through moderated regression analyses (Bartlett, Bobko, Mosier, & Hannan, 1978).

Incremental validity analyses were conducted using hierarchical regression. Three separate sets of hierarchical regression analyses for predicting both task and contextual performance criteria were conducted. These analyses investigated the incremental validity of the SJT over (1) cognitive ability, (2) personality, and (3) both cognitive ability and personality using two steps. In the first step of these hierarchical regression analyses, cognitive ability and/or personality variables were included and situational judgment was added in the second step. We also report the partial correlations for each of the predictors in the regression. These partial correlations are a measure of relative incremental prediction across all predictors in the regression. To aid in understanding of the incremental validity results, regression analyses were used to predict the SJT score from cognitive ability and the personality variables.

We also examined the substitutability of four predictor composites. The first composite consisted of the cognitive ability test and the SJT. The second composite consisted of the cognitive ability test and the personality measures. The third composite consisted of the personality measures and the SJT. The fourth composite consisted of cognitive ability, SJT, and the personality tests. Each composite was expressed in two forms, which differed in the weights applied to the predictors. The first form of each composite was based on the predictor regression weights for predicting task performance. The second form of each composite was based on the predictor weights for predicting contextual performance. We evaluated each composite with respect to its validity and its magnitude of Black–White race differences.

In response to a reviewer request, analyses were also conducted to estimate the number of Blacks and Whites and men and women who would be hired using the various composites. For each composite, we rank ordered the score and set the passing point at the 80th percentile such that 20% of the sample 'passed.' For each composite, we reported the count and percentage of the subgroups who passed.

3. Results

The means, standard deviations, and correlations of the predictors and criteria are presented in Table 1. The reliabilities are in the diagonal. The cognitive ability test was a composite of four reasoning scales. The reliability of the cognitive test was calculated as a linear composite of the four subscales (Nunnally & Bernstein, 1994, pp. 268–269). A test–retest or parallel form reliability would have been preferred for the SJT because it measures multiple constructs, however, an alternate form of the SJT was unavailable. Our α reliability estimate is based on α and is best viewed as an underestimate of the reliability due to the measure's heterogeneity. The reliabilities reported for the remaining measures are α reliabilities. We note that α reliability is an overestimate of the reliability of the two job performance measures because it ignores the errors associated with inter-rater reliability. We concur with Viswesvaran, Ones, and Schmidt (1996) that the best estimate of the reliability of a supervisory rating is .52.

The SJT shared variance with several other predictors: cognitive ability (.33), conscientiousness (.33), and agreeableness (.31), positive affectivity (.26), internal locus of control (.24), and attention to detail (.21). Thus, consistent with past research (Chan & Schmitt, 1997; Clevenger *et al.*, 2001; McDaniel & Nguyen, 2001; Nguyen & McDaniel, 2001), this SJT is construct heterogeneous.

The two performance criteria (task performance and contextual performance) are correlated .64. This correlation will likely cause the degree of differential prediction to be an underestimate of that which could be obtained if the two criterion constructs were assessed so as to be less correlated (Viswesvaran, Schmidt, & Ones, 2005). However, because the two facets are theoretically different, we discuss them separately. Task performance had somewhat higher magnitude correlations with the predictors than contextual performance. The primary correlates of task performance are general mental ability (.15), conscientiousness (.14), and situational judgment (.14). The primary correlates of the contextual performance are conscientiousness (.13), internal locus of control (.12), agreeableness (.11), and the SJT (.10).

In response to a reviewer request, we conducted differential prediction analyses of the SJT. Results of these moderated regression analyses (Bartlett *et al.*, 1978) indicated that differential prediction was not present. More detailed results are available from the first author.

Hierarchical regression analyses were conducted to determine the incremental validity of SJTs over the other predictor measures. The results for each of the three sets of regression analyses for task and contextual performance are presented in Table 2. We will first discuss the analyses that used task performance as the criterion. This first set of analyses examined the incremental prediction of the SJT over general cognitive

Table 1. Means, standard deviations and inter-correlations of predictors and criteria

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Cognitive ability	4.99	1.44	(.79)								
2. Conscientiousness	5.31	2.36	.12	(.79)							
3. Attention to detail	4.63	2.52	.08	.53	(.66)						
4. Agreeableness	3.65	2.39	.01	.41	.32	(.72)					
5. Locus of control	6.09	1.73	.11	.47	.64	.31	(.65)				
6. Positive affect	5.15	2.19	.10	.51	.36	.36	.41	(.71)			
7. Situational judgment	3.83	1.74	.33	.33	.21	.31	.24	.26	(.72)		
8. Task performance	4.96	.88	.15	.14	.07	.07	.09	.08	.14	(.92)	
9. Contextual performance	5.62	.81	.07	.13	.08	.11	.12	.08	.10	.64	(.87)

Note: N = 1140. All correlations are uncorrected. Reliability estimates are reported in parentheses on the diagonal. All correlations are significant $p < .05$ except for the .01 correlation between cognitive ability and conscientiousness. Cognitive ability was a composite of four cognitive tests. Its reliability was determined using the reliability of linear combination (Nunnally & Bernstein, 1994, pp. 268–269). The α reliability of the SJT is a likely underestimate due to the heterogeneity of the measure. SJT, situational judgment tests.

Table 2. Incremental prediction of SJT

Independent variables	Task performance					Contextual performance				
	β	R	R^2	ΔR^2	Partial *r*	β	R	R^2	ΔR^2	Partial *r*
a.										
Step 1		**.15**	**.024**				**.07**	**.005**		
Cognitive ability	**.15**					**.07**				
Step 2		**.18**	**.033**	**.011**			**.11**	**.012**	**.007**	
Cognitive ability	**.12**				.11	.04				.04
SJT	**.10**				.10	**.10**				.09
b.										
Step 1		**.14**	**.021**				**.16**	**.025**		
Conscientiousness	**.13**				.10	**.10**				.08
Attention to detail	−.03				−.02	−.04				−.03
Agreeableness	.01				.01	.06				.05
Locus of control	.05				.03	**.08**				.06
Positive affect	.00				.00	−.02				−.01
Step 2		**.18**	**.031**	**.010**			.16	.027	.003	
Conscientiousness	**.11**				.08	**.09**				.07
Attention to detail	−.30				−.02	−.03				−.03
Agreeableness	−.01				−.07	.05				.04
Locus of control	.04				.03	**.08**				.06
Positive affect	−.12				−.01	−.02				−.02
SJT	**.11**				.10	.06				.05
c.										
Step 1		**.20**	**.040**				**.17**	**.028**		
Cognitive ability	**.14**				.14	.06				.06
Conscientiousness	**.12**				.10	**.09**				.07
Attention to detail	−.03				−.02	−.04				−.03
Agreeableness	.02				.02	.06				.06
Locus of control	.04				.03	.08				.06
Positive affect	−.01				−.01	−.02				−.02
Step 2		**.21**	**.043**	**.003**			.17	.029	.001	
Cognitive ability	**.12**				.11	.05				.04
Conscientiousness	**.11**				.08	**.09**				.07
Attention to detail	−.03				−.02	−.03				−.03
Agreeableness	.01				.01	.05				.05
Locus of control	.03				.02	.07				.06
Positive affect	−.01				−.01	−.02				−.02
SJT	**.07**				.06	.04				.04

Note: Bold numbers are significant $p < .05$. SJT, situational judgment tests.

ability. Cognitive ability correlated .15 with task performance. Adding the SJT improved the multiple correlation to .18, an incremental *R* change of .03 ($p < .05$). Thus, the SJT added some incremental validity over general cognitive ability for task performance. We note that .03 is not a large increment but we suspect that many would be willing to add a SJT to a battery if it could raise the validity of the battery from .15 to .18. An inspection of the partial correlations showed that the partial for cognitive ability (.11) was about the same as the partial for situational judgment (.10).

The second set of analyses for task performance examined the incremental prediction of SJT over five personality variables (conscientiousness, attention to detail, agreeableness, locus of control, and positive affect). The set of four personality variables correlated .14 with task performance. The multiple *R* for the predictor set of personality variables and the SJT was .18, an incremental *R* change of .04 ($p < .05$). We note that .04 is not a large increment, but we suspect that many would be willing to add a SJT to a battery if it raised the validity of the battery from .14 to .18. An analysis of the partials shows that the SJT (.10) and conscientiousness (.08) have the largest partials.

In the third set of analyses for task performance, the incremental prediction of the SJT over that achieved by the combination of general cognitive ability and the five personality variables was examined. The multiple *R* for the predictor set of general cognitive ability and personality variables was .20. The addition of the SJT raised the multiple correlation to .21 for an increment of .01 ($p < .05$). We note that .01 is a small increment. We suspect that many would not be willing to go through the cost of creating a SJT if only adds .01 to the prediction battery. An analysis of the partials showed that general cognitive ability had the largest partial (.11), followed by conscientiousness (.08), and SJT (.06). The remaining four partials were substantially lower (−.02 to .02).

Results for the three sets of regression analyses for contextual performance also are also presented in Table 2. The first set of analyses examined the incremental prediction of the SJT over general cognitive ability. Cognitive ability correlated .07 with contextual performance. Adding the SJT improved the multiple correlation to .11, an incremental *R* change of .04 ($p < .05$). Although .04 is not a large increment, we suspect that many would add a SJT to a battery if it would raise the validity of the battery from .07 to .11. An inspection of the partial correlations showed that the partial for the SJT, although small, was more than twice as large as the partial for cognitive ability (.09 vs .04). We suspect that partial for the SJT was larger than that of the cognitive ability tests because the SJT has substantial non-cognitive variance in common with the criterion.

The second set of analyses for contextual performance examined the incremental prediction of SJT over five personality variables (conscientiousness, attention to detail, agreeableness, locus of control, and positive affect). The set of personality variables correlated .16 with contextual performance. The multiple *R* for the predictor set of personality variables plus the SJT was also .16. Thus, most practitioners would be unwilling to add a SJT to a battery already containing five personality scales. The variables with the largest partials were conscientiousness (.07), locus of control (.06) and the SJT (.05).

The third set of analyses for contextual performance examined the incremental prediction of the SJT over that achieved by the combination of general cognitive ability and the five personality variables. The multiple *R* for the predictor set of general cognitive ability and personality variables was .17. The addition of the SJT did not raise the multiple *R*. Again, most practitioners would be unwilling to add an SJT into a battery already containing five personality tests and a cognitive ability test for no increase in validity. The three largest partials in the analysis were conscientiousness (.07), locus of control (.06) and agreeableness (.05).

Table 3 presents standardized mean differences (Cohen, 1977) and percentiles for the various predictors and criteria. The results contrast Black and White respondents as well as male and female respondents. There were too few respondents in other race groups for a meaningful analysis. For the d_{WB}, positive values of *d* reflect higher scores in the White subgroup. The Black percentile represents the normal distribution of the Black mean relative to a value at the 50th percentile for the White mean. For the d_{MF}, positive values of *d* reflect higher scores in the male group. Female percentile represents the normal distribution of the female mean relative to a value at the 50th percentile for the male mean.

On the predictor side, Whites obtained higher scores than Blacks on all predictors except positive affect. On the criterion side, Whites obtained higher ratings than Blacks on both criteria, with the largest difference in task performance. The finding of larger mean racial differences in task ($d = .47$) than contextual performance ($d = .18$) are consistent with a major review of mean racial differences (McKay & McDaniel, 2006) but are larger than those reported by McKay and McDaniel (task $d = .21$; contextual $d = .13$). The SJT shows much lower race differences ($d = .38$) than the cognitive test ($d = .66$) and larger race differences than the personality scales (−.01 to .17). The correlation between the vector of Black–White standardized mean differences of all predictors (excluding cognitive ability) and the vector containing the correlations of the test with cognitive ability is .91, indicating that differences across the predictors in mean racial differences is best attributed to the cognitive loading of the tests. This finding is consistent with Spearman's hypothesis that the magnitude of White–Black differences on various tests is directly related to the tests' cognitive-loading (Spearman, 1927, p. 379).

On the predictor side, females obtain lower scores on all predictors except for conscientiousness ($d = -.05$), agreeableness ($d = -.24$), and situational judgment ($d = -.27$). On the criteria side, females obtained somewhat lower ratings than males on both criteria, with the largest difference in task performance (.11 vs .05).

Table 4 presents the results concerning the substitutability of predictor composites. For task performance, the validity of the composites ranged from .18 to .21. All of the composites yielded higher validity than the cognitive ability test alone ($r = .15$, see Table 1) and lower race differences. The composite of cognitive ability and the SJT showed the largest race difference ($d = .65$) which was nearly the same as the race difference for cognitive ability alone, ($d = .66$, see Table 3). Although the sole composite without cognitive ability showed the smallest race difference ($d = .34$) it also showed the lowest validity of the composites ($r = .18$). If we were to choose a composite for the task performance, we would choose the composite containing cognitive ability, the SJT, and the personality scales because it had the largest validity ($r = .21$) and slightly increased the magnitude of the race differences ($d = .59$ vs .56) over the second most valid composite of cognitive ability and personality.

For the prediction of contextual performance, three of the four composites yield the same validity ($r = .17$). If we were only interested in the prediction of contextual performance, we would choose the personality and SJT composite because it has substantially smaller race difference ($d = .24$) than the other composites. However, we would find it odd to be solely concerned with contextual performance.

We also ran a regression to predict the SJT score from the cognitive ability test and the five personality scales. The multiple R was .49 with cognitive ability, conscientiousness, and agreeableness having statistically significant β weights. Dropping locus of control and attention to detail as predictors also yielded a multiple R of .49 in which all the predictors had statistically significant β weights. In this regression, the SJT score was a function of cognitive ability, conscientiousness, agreeableness and positive affect.

Table 3. Standardized mean differences and percentiles for predictors and criterion for major subgroups

Variables	d_{WB}	Black percentile	d_{MF}	Female percentile
Predictors				
Cognitive ability	.66	26	.30	38
Conscientiousness	.17	43	−.05	52
Attention to detail	.13	45	.11	46
Agreeableness	.01	50	−.24	59
Locus of control	.11	46	.06	48
Positive affect	−.01	50	−.05	52
Situational judgment	.38	35	−.27	60
Criteria				
Task performance	.47	32	.11	45
Contextual performance	.18	43	.05	48

Note: N for Blacks = 221; N for Whites = 735. d_{WB} = White and Black differences. For the d_{WB}, positive values of d reflect higher scores in the White subgroup. Black percentile represents the normal distribution of the Black mean relative to a value at the 50th percentile for the White mean. N for males = 757; N for females = 361. d_{MF} = male and female differences. For the d_{MF}, positive values of d reflect higher scores in the male group. Female percentile represents the normal distribution of the female mean relative to a value at the 50th percentile for the male mean. Discrepancies between the d values and the percentiles are due to rounding.

Table 4. Substitutability of predictors. Effects on validity and race differences

Independent variables	Task performance			Contextual performance		
	R	d_{WB}	Black percentile	R	d_{WB}	Black percentile
Personality and SJT	.18	.34	37	.17	.24	41
Cognitive ability and SJT	.18	.65	26	.11	.56	29
Cognitive ability and personality	.20	.56	29	.17	.34	37
Cognitive ability, SJT, and personality	.21	.59	28	.17	.37	36

Note: N for Blacks = 221; N for Whites = 735. d_{WB} = White and Black differences. For the d_{WB}, positive values of d reflect higher scores in the White subgroup. Black percentile represents the normal distribution of the Black mean relative to a value at the 50th percentile for the White mean. Discrepancies between the d values and the percentiles are due to rounding. Effects on validity and race differences. SJT, situational judgment tests.

Table 5. Counts and percents of Blacks, Whites, men, and women in the top 20% of rank order test scores

Predictor	Black (23.2%)	White (76.8%)	Men (67.7%)	Women (32.3%)
1. Cognitive ability alone	21 (10.9%)	171 (89.1%)	181 (80.4%)	44 (19.6%)
2. Personality alone	42 (21.8%)	151 (78.2%)	150 (66.4%)	76 (33.6%)
3. SJT alone	37 (13.9%)	229 (86.1%)	171 (57.8%)	125 (42.2%)
4. Cognitive ability + SJT	15 (7.8%)	177 (92.2%)	165 (73.3%)	60 (26.7%)
5. Personality + SJT	30 (15.6%)	162 (84.4%)	132 (58.4%)	94 (41.6%)
6. Cognitive ability + personality	22 (11.5%)	170 (88.5%)	169 (75.1%)	56 (24.9%)
7. Cognitive ability + personality + SJT	23 (12.0%)	168 (88.0%)	163 (72.4%)	62 (27.6%)

SJT, situational judgment tests.

Table 5 presents results on the percentage of Blacks and Whites and males and females who would pass each test and composite if the passing point were set at the 80th percentile. The Black and White analyses were based on the 956 members of the sample who were either Black or White. The analyses of men and women were based on the 1118 sample members whose sex was known. The first column of the table shows which test or composite is being used. The possibilities are (1) the cognitive test alone, (2) the five personality measures (used as set) alone, (3) the SJT alone, (4) cognitive ability plus the SJT, (5) the five personality scales plus the SJT, (6) cognitive ability plus the five personality scales plus SJT, and (7) the cognitive ability test plus the five personality scales, plus the SJT. When the predictor contained more than one measure, the composite was formed using the weights from the regression equation for task performance. The next two columns show the percentage of the top 20% who are Black or White. The final two columns show the percentage of the top 20% who are men or women.

For the Black and White analyses, 23.2% of the sample of 956 Black or White individuals are Black and 76.8% are White. These are the percentages of Blacks and Whites who would be in the top 20% on average if 20% were selected at random. Note that 20% of 956 Blacks and Whites results in a sample of 191 individuals. The counts of Blacks and Whites in the top 20% for any test composite will sum to more than 191 when there are ties at the score received by the 191st person in a rank-ordered list. We note that there were many people tied at the 80th percentile score for the SJT.

Personality only has the largest percentage of Blacks above the 80th percentile (21.8%). Composites containing cognitive ability (10.9%) and cognitive ability with SJT (7.8%) had the smallest percentage of minorities. Although the composite containing only SJT had a smaller percentage of blacks (13.9%), the total number of blacks (37) in the top 20% was only slightly less than the personality composite (42) resulting from a large number of tie scores on the SJT.

For analyses comparing men and women, 67.7% of the 1118 individuals of known gender are male and 32.3% are female. These are the percentages of men and women who would be in the 20% on average if 20% were selected at random. Note that 20% of 1118 individuals results in a sample of 224 individuals. The counts of men and women in the top 20% for any test will sum to more than 224 when there are ties at the score of the 224th person. We note that there were many people tied at the 80th percentile score for the SJT.

The composite containing only SJT had the largest percentage of women above the 80th percentile (42.2%) and the composite with personality and SJT had the second largest percentage (41.6%). The composite containing only cognitive ability had the smallest percentage (19.6%).

4. Discussion

This paper had three goals. First, it responded to the call by Nguyen *et al.* (2005) for more research on mean racial and sex differences on SJTs. Consistent with their findings of cognitive and personality moderators of subgroup differences, we provide SJT correlates with cognitive and personality variables. Second, the paper is responsive to the McDaniel *et al.* (2006) calls for more research on the incremental validity of SJTs in combination with both cognitive and personality predictors. Third, the paper extended the contribution of Chan and Schmitt (2002) by examining the validity of SJTs for both task and contextual performance.

With respect to the first goal, we contributed to the literature on mean racial and gender differences in SJTs. The Black–White mean difference in the SJT was .38 that is identical to the mean offered by Nguyen *et al.* (2005) in a preliminary meta-analysis of subgroup differences in SJTs. The gender difference is in this study ($d = -.27$), which favored females, was larger than that found in Nguyen *et al.* (2005) who reported a d of $-.10$, which also favored females. Nguyen *et al.* had argued for substantially more research on subgroup differences in SJTs and the results of our large sample study makes an incremental contribution to this literature.

Although not an explicit goal of the present study, we do report subgroup differences in the job performance measures (see Table 3). McKay and McDaniel (2006) noted that mean racial effect sizes reported in journals were smaller, on average, than those reported in unpublished technical reports and suggested that this might be due to publication bias. McDaniel, McKay, and Rothstein (2006) offered evidence that low magnitude mean racial differences are being systematically suppressed in the journal literature. This most likely happens because journals give authors the discretion of reporting mean racial differences in job performance and authors tend to report such differences when they are small and not when they are large. Consistent with McDaniel *et al.*'s (2006) recommendation that reporting of mean racial differences in job performance be required by journals, we report our values in Table 3.

Our paper's second goal was met by analyses examining the incremental validity of SJTs over cognitive and personality measures. Our third goal was met by conducting such analyses separately by task and contextual performance. We discuss these accomplishments below.

For task performance, cognitive ability had a validity of .15 and the SJT added a .03 validity increment to a cognitive test. The personality predictors had a validity of .18, and the SJT added a .04 validity increment to a battery composed of five personality predictors. A battery of the five personality scales and the cognitive ability tests had a validity of .20, and the SJT added a .01 validity increment to a battery. Thus, for task performance, it would be useful to supplement a cognitive ability test with a SJT. It would also be useful to supplement a personality battery with a SJT. However, if the test battery contains both personality tests and a cognitive ability tests, the contribution of the SJT will be marginal.

For contextual performance, cognitive ability had a correlation of .07 and the SJT added an increment of .04 to cognitive ability. Situational judgment did not, however, add incrementally over a composite of five personality tests. Nor did it contribute incrementally to a test battery composed of five personality tests and a cognitive ability tests. Thus, for contextual performance, it was useful to supplement the cognitive ability test with the SJT. However, in the prediction of contextual performance, it was not useful to add a SJT to a battery of five personality tests or to a battery with five personality tests and a cognitive ability test.

From these results, we conclude that SJTs are a useful component of a selection battery to predict task performance and that SJTs are also a useful complement to predict task performance when added into a battery of personality or when used to supplement a cognitive ability test. Although we obtained a statistically significant incremental validity when adding a SJT to a cognitive ability test and a personality battery, the incremental validity was small (.01). The incremental validity of any test is a function of its correlation with other tests, its correlation with the criterion, and the extent to which the criterion variance it captures is also captured by the other predictors. The SJT used in this study was moderately correlated with the cognitive ability measure (.33) and moderately correlated with each of the five personality measures (.21–.33). In fact, one obtains a multiple *R* of .49 when predicting the SJT from cognitive ability and a set of personality variables. This degree of cognitive and personality saturation in the SJT situation makes it difficult for this SJT to predict over and above a battery containing cognitive ability and personality. If an SJT could be built with lower correlations with these variables, it would have a greater chance of showing incremental validity above both the cognitive ability tests and the five personality scales.

The test battery was less able to predict contextual performance than task performance. The SJT was able to predict over and above cognitive ability in the prediction of contextual performance. In part, this is due to the very low correlation ($r = .07$) between the cognitive ability test and contextual performance. However, the SJT did not provide any meaningful or statistically significant incremental prediction over personality or a composite of personality and cognitive ability.

4.1. Limitations of the study

We note that all correlations presented are for observed data. No reliability corrections or range restriction corrections have been made. Thus the validity estimates presented are underestimates (i.e., downwardly biased) of their population values. We also note that these jobs are relatively low in cognitive complexity demands. One would expect higher validities for the cognitive ability measures and the cognitively correlated measures (i.e., SJT) in occupations with higher cognitive complexity (Gandy, 1986; Gutenberg, Arvey, Osburn, & Jeanneret, 1983; Hunter, 1983; McDaniel, 1986).

We also note that the data were collected on incumbents. This raises two issues. First, there is likely more range restriction in the predictors than would be found in applicant samples. The criterion measures would also be restricted by the effects of whatever selection was used to screen the employees. Also, the criterion is likely affected by range restriction because some employees may have left the organizations due to poor job performance. The second issue is that job incumbents who completed the measures likely had less motivation to fake than applicants. The validity for applicant samples is usually slightly lower than the validity for concurrent samples. Thus, the validity of the personality tests may be slightly lower in operational screening. Nguyen *et al.* (2005) noted that SJTs with knowledge response instructions appear to be faking resistant. Knowledge instructions ask the respondent to indicate the effectiveness of the SJT response options as opposed to behavioral tendency instructions which request respondents to indicate their behavioral tendencies (e.g., what would you most likely do?). The SJT in this study was a knowledge instruction SJT and therefore likely to be more faking resistant than the personality test. In an operational setting, the somewhat lower validity of the personality test might cause the SJT to make a greater incremental contribution over and above personality. This is clearly speculative but is offered as the type of reasoning in which one may engage when using concurrent validity data to make inferences about the likely results in a predictive setting.

SJTs are measurement methods. Any given SJT may show different population-level correlations with other tests than another SJT. Because the intercorrelations of the tests in part control the incremental validity of a test, there could well be some SJTs which show no useful levels of incremental validity and other SJTs that show substantially larger levels of incremental validity. Thus, we discourage an overreliance on these findings until there are substantially more studies examining incremental validity using a variety of situational judgment measures. Just as McDaniel *et al.* (2001) showed moderate variability in the validity of situational judgment measures, we anticipate that across studies, there will be meaningful levels of variability in the incremental validity associated with SJTs. Likewise, it is also important that the publication decisions on such studies not depend on the magnitude of the incremental effects. Should studies showing incremental validity be more likely to be published than studies showing no or low incremental effects, cumulations of this literature will be inaccurate due to publication bias and magnitude of incremental validity effects for situational judgment studies will be overestimated.

5. Conclusion

This literature sorely needs some taxonomy of SJTs or items that can help explain the variance in the magnitude of the tests with other predictors and with criteria. In addition, future studies should examine prediction differences for both task and contextual performance. Although the findings of this study are limited by its choice of predictors, and potentially by the restriction of its sample to manufacturing jobs, the study is useful in extending our knowledge of SJTs. The present study adds to the cumulative literature on the usefulness of SJTs for incremental prediction in both task and contextual performance and adds large sample results to the growing body of literature of subgroup differences in SJTs.

Acknowledgements

The authors thank Mari Kato for her assistance in preparing the data file. The paper has benefited substantially from comments provided by Neal Schmitt.

References

Bartlett, C.J., Bobko, P., Mosier, S.B. and Hannan, R. (1978) Testing for fairness with a moderated multiple regression strategy: an alternative to differential analysis. *Personnel Psychology*, **31**, 233–241.

Borman, W.C. and Motowidlo, S.J. (1993) Expanding the criterion domain to include elements of contextual performance. In: Schmitt, N. and Borman, W.C. (eds), *Personnel selection in organizations*. San Francisco: Jossey-Bass, pp. 71–98.

Bott, J.P., O'Connell, M.S., Ramakrishnan, M. and Doverspike, D.D. (2007) Practical limitations in making decisions regarding the distribution of applicant personality test scores based on incumbent data. *Journal of Business and Psychology*, (in press).

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

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

Chan, D. 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.

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

Cleary, T.A. (1968) Test bias: prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, **5**, 115–124.

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

Cober, R.T., Cober, A.T., Lawrence, A.D. and O'Connell, M.S. (2003) *Predictors of multi-tasking ability for selection: attitudes versus ability*. Paper presented at the 18th annual meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Cohen, J. (1977) *Statistical power analysis for the behavioral sciences*, (Rev. edn). New York: Academic Press.

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

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

Hattrup, K., O'Connell, M.S. and Labrador, J.R. (2005) Incremental validity of locus of control after controlling for cognitive ability and conscientiousness. *Journal of Business Psychology*, **19**, 461–481.

Hattrup, K., O'Connell, M.S. and Wingate, P.H. (1998) Prediction of multidimensional criteria: distinguishing task and contextual performance. *Human Performance*, **11**, 305–320.

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*. US Department of Labor, USES Test Research Report No. 45. Washington, DC: US Department of Labor.

McDaniel, M.A. (1986) *The evaluation of a causal model of job performance: the interrelationships of general mental ability, job experience, and job performance*. Doctoral dissertation, The George Washington University.

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

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A. and Braverman, E.P. (2001) Predicting job performance using situational judgment tests: a clarification of the literature. *Journal of Applied Psychology*, **86**, 730–740.

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

McDaniel, M.A., Whetzel, D.L., Hartman, N.S., Nguyen, N. and Grubb, W.L. (2006) Situational judgment tests: validity and an integrative model. In: Ployhart, R. and Weekley, J. (eds), *Situational judgment tests: theory, measurement, and application*. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 183–204.

McKay, P. and McDaniel, M.A. (2006) A re-examination of black–white mean differences in work performance: more data, more moderators. *Journal of Applied Psychology*, **91**, 531–554.

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. and Tippens, 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.

Nguyen, N., McDaniel, M.A. and Whetzel, D.L. (2005) *Subgroup differences in situational judgment test performance: a meta-analysis*. Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology. Los Angeles, April.

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

Nguyen, N.T. and McDaniel, M.A. (2001) *Constructs assessed in situational judgment tests: a meta-analysis*. Paper presented at the 16th Annual Convention of the Society for Industrial and Organizational Psychology, San Diego, CA, April.

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

Nunnally, J.C. and Bernstein, I.H. (1994) *Psychometric theory*. New York: McGraw Hill.

O'Connell, M.S., Doverspike, D., Gillikin, S. and Meloun, J.M. (2001) Computer anxiety: effects on computerized testing performance and implications for e.cruiting. *Journal of e. Commerce and Psychology*, **1**, 25–39.

O'Connell, M.S. and Kato, M. (2001) *Psychometric properties of the select assessment™ for manufacturing system*. Technical Report, Pittsburgh, PA: Select International Inc.

O'Connell, M.S. and Smith, M.E. (1999) *Normative and validity results for entry-level assessment scales*. Technical Report, Pittsburgh, PA: Select International Inc.

O'Connell, M.S. and Smith, M.E. (2000) *Meta-analysis of the select assessment™ for manufacturing system*. Technical Report, Pittsburgh, PA: Select International Inc.

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

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, TX.

Spearman, C. (1927) *The abilities of man*. New York: Macmillan.

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.

Van Scotter, J.R. and Motowidlo, S.J. (1996) Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology*, **81**, 525–531.

Viswesvaran, C., Ones, D.S. and Schmidt, F.L. (1996) Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, **81**, 557–574.

Viswesvaran, C., Schmidt, F.L. and Ones, D.S. (2005) Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, **90**, 108–131.

Weekly, J.A. and Jones, C. (1997) Video-based situational testing. *Personnel Psychology*, **50**, 25–491.

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