Can Basic Individual Differences Shed Light on the Construct Meaning of Assessment Center Evaluations?

J. M. Collins*, Michigan State University
F. L. Schmidt, University of Iowa
M. Sanchez-Ku, Texas A&M University
L. Thomas, Texas A&M University
M. A. McDaniel Virginia Commonwealth University
H. Le University of Iowa

The construct meaning of assessment center evaluations is an important unresolved issue in I/O psychology. This study hypothesized that Cognitive Ability and personality traits are primary correlates of evaluators' overall assessment ratings (OARs). Meta-analysis results based on 65 correlations indicate the following mean construct-level correlations with OARs: .67 for Cognitive Ability, .50 for Extraversion, .35 for Emotional Stability, .25 for Openness, and .17 for Agreeableness; yielding a multiple R of .84. These findings support our hypothesis and cast light on the construct meaning of assessment center evaluations.

Introduction

Since their introduction into industry (Bass, 1954), assessment centers have been used world-wide for personnel purposes. Assessment centers are combinations of simulated work exercises and other assessment procedures designed to assess the job-related skills and abilities of job applicants and incumbents. In addition to group-based exercises like the Leaderless Group Discussion (LGD), assessment centers typically include such individual assessments as in-depth interviews and mental ability tests. The research evidence indicating that assessment centers are valid for predicting job performance, personnel potential, and other measures (Gaugler, Rosenthal, Thornton, and Bentson, 1987; also, Klimoski and Brickner, 1987; McEvoy and Beatty, 1989; Schmitt, Gooding, Noe, and Kirsch, 1984) resolves any question of the criterion-related validity of assessment centers. However, the construct meaning of assessment ratings is still uncertain (Guion, 1998). In fact, evidence has been reported *against* the construct validity of intended measures of constructs in assessment centers (as described below).

The construct validity of ACs has been researched on several different levels (e.g., see Thornton, 1992, p. 100). The most specific ratings obtained in ACs are the within-exercise dimension ratings, for example, evaluations made on leadership and problem solution creativity within the Leaderless Group Discussion exercise. As discussed below, research findings indicate little construct validity evidence at this level. The next level of generality is that of final dimension ratings – ratings made by averaging within-exercise dimension ratings across exercises. For example, several exercises might include a leadership rating and these ratings would be averaged across exercises. At this level, the construct validity findings are slightly less negative (Shore, Thornton, and Shore, 1990; Lievens and Klimoski, 2001). Finally, at the highest level of generality, one has the OAR (the Overall Assessment Rating), a final summary score reflecting overall assessment center performance. In addition to dimension ratings at both lower levels, the OAR also reflects information from other sources, such as peer ratings, personality inventories, ability test scores, etc. (The present study is focused mostly at the OAR level.) As noted below, construct validity evidence at this level is disappointing.

* Address for correspondence: Judith Collins, Leadership and Management Program, Michigan State University, East Lansing, MI 48907, USA. E-mail: judithc@msu.edu

There is considerable evidence for low construct validity at the level of within-exercise dimension ratings. For example, Sackett and Dreher (1982) found that measures of different constructs within the same exercise correlated higher with each other than did measures of the same construct across exercises. In fact, instead of the expected construct-based factors, the results revealed exercise-based factors. Others studies also have called into question the construct validity of assessment centers at this level. There is considerable evidence for the lack of convergent and discriminant validity of assessment center exercises and dimensions within the exercises (cf. Brannick, Michaels, and Baker, 1989; Crawley, Pinder, and Herriott, 1990; Gaugler and Thornton, 1989; Sackett and Dreher, 1982; Sackett and Harris, 1988; Shore *et al*., 1990; Turnage and Muchinsky, 1982). Other negative construct evidence comes from factor analytic studies of overall assessment center ratings (OARs) that found either one underlying factor (Sackett and Harris, 1988; Turnage and Muchinsky, 1982) or two factors (Archambeau, 1979; Howard and Bray, 1988; Outclat, 1988; Russell, 1985; Shore *et al*., 1990). In light of these findings, Guion (1998) has concluded that construct validity is the biggest unresolved issue in the assessment center area.

One time-honored approach to shedding light on the construct validity of a measure is examination of its correlates (APA Test Standards, 1999; Cronbach and Meehl, 1995). Such correlates can be especially informative if the measures correlated are 'basic' or fundamental constructs and if the correlations are large. In this article, we used this approach to probe the construct meaning of the final evaluations produced by assessment centers, the overall assessment ratings (OARs). We used meta-analysis to combine correlations from existing studies to estimate the extent to which fundamental individual difference traits – general mental ability and the Big Five personality traits – predict final OARs. If these traits prove effective in predicting OARs – that is, if the magnitude of the correlations is large – it would appear likely that the behaviors and performances observed during the assessment center exercises, upon which the assessors presumably base their evaluations, are influenced or shaped by these fundamental individual difference traits. Note that such findings would not necessarily indicate that assessors observe these traits themselves but, rather, that the behaviors and performances that they observe reflect (are caused by) these basic traits. In this sense, then, such a finding would shed light on the construct meaning, and hence the construct validity, of assessment centers. However, as we note later, large trait-OAR correlations could also result from assessors weighting personality and ability test scores heavily in producing their OARs. Either way, these correlations provide information about the construct meaning of OARs.

Theoretical Background

Two commonly used assessment center exercises are in-basket simulated work exercises and leaderless group discussions. In-basket tasks are written exercises designed to test a candidate's problem-solving abilities. Several studies have found positive correlations between in-basket scores and verbal and numerical ability (e.g., Gaugler *et al*., 1987; also see Bray and Grant, 1966; Carlton, 1970; Hakstian, Woolsey, and Schroeder, 1986; Huck, 1973; Moses, 1973; Russell and Kuhnert, 1992; Schmitt, 1977; Thornton and Byham, 1982; Wolfson, 1985 as cited in Klimoski and Brickner, 1987). The most likely explanation for these correlations is that verbal and numerical tests are measures of general mental ability (GMA), which is known to be correlated with performance on cognitively complex tasks of all kinds (Hakstian *et al*., 1986, Hunter and Schmidt, 1996; Schmidt and Hunter, 1998).

In-basket tasks are work samples of major and specific duties required on the job. The in-basket score is based either on performance-related behaviors, or on the resulting product, or on both, and the construct ostensibly assessed by the work sample is the ability to perform the tasks (Guion, 1991). Furthermore, in-baskets, being work samples, are examples of task performance, which is predicted by general mental ability (Schmidt and Hunter, 1992). We therefore hypothesize that general mental ability is correlated with in-basket scores and therefore with overall assessment center ratings, because those ratings are based in part on in-basket tasks.

We further hypothesize that leaderless group discussion scores are more strongly correlated with personality traits than are in-basket scores. This hypothesis is not original: Bass (1954) hypothesized long ago that personality was a determinant of ratings on leaderless group discussions (LGDs). But that hypothesis has not been tested, perhaps because for many years personality tests were not part of assessment centers. However, in recent years assessment centers have begun to measure personality (Gaugler *et al*., 1987; also, Levy-Leboyer, 1963; Russell, 1987) and there is now a sufficiently large literature database to allow hypothesis testing using meta-analysis.

In LGDs candidates are observed in small group settings as they interact with and attempt to influence others. LGD ratings are based on the assessor's observations of behaviors that may be affected by personality characteristics such as aggressiveness, cooperation, adaptability, and sociability (Bass, 1954). However, candidates must also offer solutions to the LGD problem-solving task, and this depends on general mental ability. We therefore expect that LGD scores are also related to general mental ability but for several reasons we believe the relationship with personality will be stronger.

First, whereas task-related performance refers to duties specific to the job task, such as the problem-solving tasks found in in-basket exercises, leaderless group discussions involve interpersonal interactions and therefore may be related to the concept of contextual performance. Borman and Motowidlo (1993) introduced contextual performance, distinguished this concept from task performance by defining it as those job behaviors that surround and support the performance of job tasks, and hypothesized that personality is a major determinant of contextual performance.

Second, the quality and nature of interpersonal interactions depend on the personality of the interactors. That is, personality traits are expected to manifest themselves in behaviors such as 'social skill' or adroitness in interacting with others (Derlega, Winstead, and Jones, 1991).

A third reason for our hypotheses linking in-baskets predominantly with general mental ability and LGDs predominantly with personality is the empirical literature indicating that task performance is predominantly predicted by general mental ability, while contextual performance is better predicted by personality (e.g., Hunter and Hunter, 1984; Hunter and Schmidt, 1996; McHenry, Hough, Toquam, Hanson, and Ashworth, 1990; Motowidlo and Van Scotter, 1994; Peterson, Hough, Dunnette, Rosse, Houston, Toquam, and Wing, 1990; Pulakos, Borman, and Hough, 1988; Schmidt and Hunter, 1992). Finally, one study suggested that performance on in-basket tasks was more strongly related to Cognitive Ability than to personality and that LGD ratings were more highly correlated with personality (Shore *et al*., 1990).

Our more general hypothesis – and the major one examined in this article – is that OARs are determined, either directly or indirectly, by cognitive ability and personality traits. Indirect determination occurs if these traits determine performance on exercises such as in-basket tasks and LGDs, and these performances, observed by the assessors, determine the OARs in turn. However, general cognitive ability and personality measures may be related to OARs for another reason: These measures, obtained as part of the assessment process, are available to the assessors when they assign final OAR ratings. Hence substantial correlations between these measures and OARs might reflect the fact that, from a policy capturing point of view, assessors rely heavily on these measures in assigning OARs. However, the conclusion is the same in either case. That is, regardless of the reason why these basic traits and OARs are highly correlated (if they indeed are), such correlations reflect the construct meaning of the OARs. (A reviewer suggested that correlations produced by the latter process would represent a 'confound'. This is not the case, as explained here.) A finding that these traits substantially predict final OARs would support our general hypothesis and would shed light on the construct meaning assessment center OARs.

Method

The Meta-Analysis Procedure

We used the Hunter-Schmidt interactive meta-analytic procedure (Hunter and Schmidt, 1990; Schmidt and Hunter, 1992; Schmidt, Law, Hunter, Rothstein, Pearlman, and McDaniel, 1993) to generate estimates of both the operational validity and the construct-level validity of Cognitive Ability and personality dimensions for the prediction of overall assessment center ratings. The interactive procedure uses a non-linear correction for range restriction and uses the observed mean correlation in the formula for sampling error (Schmidt *et al*., 1993), both of which have been shown in computer simulation studies to enhance accuracy (Law, Schmidt, and Hunter, 1994a; 1994b). The interactive procedure determines how much of the observed variance of the validity coefficients is attributed to differences in studies in the amount of range restriction, sampling error, and predictor and criterion unreliability. In this first analysis, we did not correct for predictor unreliability because the research goal in that phase was to estimate the operational validity for the prediction of assessment center ratings. However, in a second analysis examining construct level relations, we did make this latter correction. This analysis allows us to examine these relationships at the trait (or construct) level, providing a basis for theoretical interpretations. Reliabilities used for the Big Five measures were coefficient alpha values given in Costa and McCrae (1992) for their normative sample. The reliability used for general cognitive ability measures was the mean coefficient alpha value presented in Schmidt (1999). This value (.95) is the average across seven well-known measures of general cognitive ability.

These reliability estimates for both the personality and general cognitive ability measures lead to conservative corrections. This is because coefficient alpha (an estimate of the coefficient of equivalence; CE) overestimates actual reliability because it does not control for transient error (Schmidt and Hunter, 1996; 1999). A more accurate measure of reliability would be the coefficient of equivalence and stability (CES); i.e., the correlation between two parallel forms of the measures administered on two different occasions (two different days). Both the CE and the CES control for random response error and specific factor measurement error. But the CES, unlike the CE, also controls for day-to-day fluctuations in mood and mental state–transient measurement error (Schmidt and Hunter, 1996; 1999). Because it controls for all three sources of measurement error, the CES is a more accurate (and somewhat smaller) index of reliability of a scale. However, we were unable to locate estimates of the CES for these measures. Hence our corrected correlations must be viewed as somewhat conservative (i.e., downwardly biased) estimates of the true score (construct level) correlations. The same is true of our

true score multiple correlation. This must be kept in mind in interpreting our findings.

Study variables and key search words. We defined personality using the Big Five traits, the dominant framework used today in personality research. We conducted a comprehensive search of the published literature using the PsychInfo database and the following combination of search terms and words: 'assessment center', 'Cognitive Ability', 'intelligence', and 'neuroticism', 'Emotional Stability', 'Extraversion', 'Openness to Experience', 'Intellect', 'Agreeableness', 'Conscientiousness', and markers or facets of these Big Five personality traits (cf., Goldberg, 1990; Costa and McCrae, 1992; Hogan and Hogan, 1992) or their synonyms (Landau and Bogus, 1990). The list of personality traits and synonymous markers and facets may be obtained from the first author.

As part of the literature search, we used a snowballing technique: we checked references of retrieved articles against a list of references composed from articles already retrieved; articles not already in the database were then also retrieved and those references were checked against the continuously updated list, until all available published articles were identified, collected, and reviewed. Through this process we discovered two meta-analyses of OARs, both conducted in Germany, and a range of predictor correlates (Maukisch, 1986; Scholz and Schuler, 1993). The German reviews as well as the Gaugler *et al*. (1987) meta-analysis were also subjected to the snowballing procedure. The German reviews were based primarily on studies in American journals, most of which were either already in the present database or which did not meet our criteria for inclusion. In addition, the German reviews omitted many of the studies we located.

Criteria for including studies. Studies were retained for use in the database if they met the following criteria: (a) the studies were done on assessment centers; (b) sample sizes were reported; (c) correlations (or statistics allowing computation of correlations) were reported between personality or general mental ability measures and overall assessment center ratings (OARs), defined as ratings summed across two or more assessment center exercises or dimensions; (d) cognitive ability and personality test scores were generated using paper and pencil tests; and (e) the samples were from field (versus laboratory) settings. Altogether, 524 articles were located, retrieved, and then reviewed by two of the study authors for possible use in the meta-analysis. Personality measures were retained only if they could be classified as measuring one of the Big Five personality traits, based on the classification methods of Barrick and Mount (1991). The coding schemes for the correlations and for the interrater reliabilities were simple and there were therefore only a few coding discrepancies. These were easily resolved upon discussion. Most of the 524 articles fell into the following unusable categories: (a) theoretical articles that contained no statistics; (b) review articles that cited statistics reported in other studies already in the database; (c) articles that examined the construct validity of assessment center exercises and dimensions and that reported no OARs; (d) articles in which OARs were reported but the correlations were not; and/or (e) articles that did not use paper and pencil measures of Cognitive Ability or personality. Overall, 80 correlations met the study criteria. These comprised the database.

The database. Overall, 26 of the 80 correlations were from independent samples and 54 correlations from non-independent samples. That is, some studies reported correlations between an overall rating and two (or more) measures of Cognitive Ability. In these cases of statistical dependence, we computed composite correlations, using the Hunter and Schmidt (1990, pp. 458) formula. For multiple Cognitive Ability tests administered to a sample, the composite correlation is the correlation between the sum of the ability test scores and the overall assessment center rating. The computation of this composite requires either (a) the intercorrelation matrix for the different measures of Cognitive Ability within that study, or (b) the average intercorrelation among Cognitive Ability measures. Six studies in the meta-analysis did not report the intercorrelation matrix for the study variables. In those cases, the composite correlation was computed using .80 as the average of the correlations among the cognitive ability tests, the value reported by Jensen (1980) as the average of all the values of correlations among cognitive ability tests. There were no reported correlations of assessment center ratings with any measure of the conscientiousness trait, precluding its use in the meta-analysis. After computation of composite correlations, the overall database consisted of 65 correlations and the total sample size was $N = 9{,}738$.

Criterion artifact distribution. We used an artifact distribution of interrater reliabilities generated from the assessment center literature to correct for attenuation due to unreliability in assessment center ratings. We compiled this distribution from reliabilities having the following three characteristics. The reliabilities were: (1) identified as interrater reliabilities (versus internal consistency or rate-rerate); (2) computed on assessor's overall ratings, or on a series of exercises or dimensions within exercises that comprised the overall rating; and (3) computed prior to discussion among the evaluators (i.e., ratings were independent). Many studies reported interrater reliabilities computed *after* consensus discussions, and many other studies reported coefficient alphas or test-retest reliabilies; we did not include these. Eleven studies reported information meeting these criteria, and these

reliabilities ranged from .68 to .85 (cf., Bray and Grant, 1966; Frederiksen, Saunders, and Wand, 1957; Lowry, 1994; Schmitt, 1977). Appendix A lists the distribution of interrater reliabilities and the primary studies that contributed to the distribution.

Range restriction. The concept of restriction is that persons below some score on the predictor are unacceptable applicants and therefore are excluded from a validity study. In the validity generalization procedure, restriction of range is quantified in terms of u, the ratio of the restricted sample standard deviation to the unrestricted population standard deviation (Hunter and Schmidt, 1990; Schmidt, Hunter and Urry, 1976). For corrections for restriction in range for cognitive ability we used the empirically derived distribution reported by Alexander, Carson, Alliger, and Cronshaw (1989). This distribution is shown in Appendix B. In the case of personality measures, the studies did not report data from which we could compute a range restriction distribution and appropriate empirically derived published distributions were not available. For example, the well-known Barrick and Mount (1991) meta-analysis of the validity of personality measures did not report a distribution of u values. They did, however, report a mean u value of .94, indicating minimal levels of range restriction. In light of this, we did not correct for range restriction in personality measures. Although this can be expected to produce a conservative bias in our correlation estimates, this bias is minimal as indicated by a value of .94. (In fact, we ran the calculations both with and without the range restriction correlations. After rounding to two decimal places, the results were identical. Hence we reported only the uncorrected results.)

Results

Table 1 presents the meta-analytic results. In Table 1, $\bar{\rho}_o$ is the mean operational validity, corrected for range restriction and for unreliability in the criterion (overall ratings), but not for unreliability in the predictor. SD_{ρ_o} is the standard deviation of ρ_o, indexing the estimated amount of variation in operational validity population correlations; the percent variance accounted for is the sum of sampling error variance and variance due to between-study differences in reliability of OARs and in the degree of range restriction, divided by total observed variance (times 100). The 90% credibility value is the value of ρ_o at the 10th percentile in the ρ_o distribution.

Table 1 also shows that mean ($\bar{\rho}_T$) and standard deviation (SD_{ρ_T}) of the trait-level (or construct-level) correlations. Unlike the operational validities, these values are corrected for measurement error in personality scales and GMA measures. This correction increases all correlations slightly.

The meta-analysis results showed the following mean operational validities for the prediction of overall ratings: $\bar{\rho}_o = .65$ for Cognitive Ability; $\bar{\rho}_o = .16$ for Agreeableness; $\bar{\rho}_o = .47$ for Extraversion; $\bar{\rho}_o = .34$ for Emotional Stability; and $\bar{\rho}_o = .23$ for Openness to experience. For Cognitive Ability, the relatively large SD_{ρ_o} (.16) and the moderate variance accounted for points to a moderator or moderators of the relationship between overall assessment center ratings and Cognitive Ability. For Agreeableness and Openness, the small SD_{ρ_o} (.03 and .00) and the large percent variance accounted for (91.61 and 100.00) indicates that the magnitude of these validities do not vary much across the studies in the meta-analysis. For Emotional Stability, however, the large SD_{ρ_o}(.26) together with the small percent variance accounted for (20.99) suggests that the magnitude of ρ_o may vary across studies due to unknown and unaccounted for artifacts or moderators.

As discussed earlier, the assessment literature suggests that performance on different exercises may depend on different characteristics (e.g., Brannick *et al*., 1989; Crawley *et al*., 1990; Gaugler and Thornton, 1989; Sackett and Dreher, 1982; Sackett and Harris, 1988; Shore *et al*., 1990; Turnage and Muchinsky, 1982). For example, Cognitive Ability may be a better predictor than personality of performance on in-basket tasks, and personality may be a stronger predictor than Cognitive Ability of performance in leaderless group discussions. To conduct the moderated meta-analysis, we retrieved studies from the database that reported correlations separately for in-basket and LGD exercises. Because a goal of most assessment centers is to generate an overall evaluation by integrating the exercise scores and other information, few correlations were reported for independent exercises. Only 15 studies in the database reported correlations between LGD and Cognitive Ability and only four studies reported correlations between in-basket scores and Cognitive Ability. For personality variables, there were only six correlations – three correlations between LGD and Extraversion, and three correlations between LGD and Agreeableness. Because of the limited number of correlations and the associated possibility of second order sampling error, we did not meta-analytically test those moderator hypotheses. We therefore report only the results for Cognitive Ability.

These results are reported in Table 2. LGD results showed that the operational validity for Cognitive Ability for predicting the LGD was .57 ($SD_{\rho_o} = .15$) and 49.19% of the variance was accounted for by artifacts. These figures are based on 15 studies and a total N of 2,697. Although this finding does not strictly disprove our hypothesis that personality is the dominant construct underlying LGD scores, the .57 validity indicates that Cognitive Ability is an important construct underlying the performance in leaderless group discussions. Indeed,

Table 1: Validities for the prediction of overall assessment center rating from cognitive ability and personality variables

Predictor	K	N	$\bar{r}$	SD_r	$SD_{residual}$	$\bar{\rho}_o$	$SD\rho_o$	90% Credibility Value of ρ_o	% Variance Accounted	$\bar{\rho}_T$	SD_{ρ_T}
Cognitive Ability	34	5419	.43	.14	.11	.65	.16	.44	38.62	.67	.16
Agreeableness	7	830	.12	.09	.02	.16	.03	.12	91.61	.17	.03
Extraversion	13	1847	.36	.13	.09	.47	.12	.31	42.95	.50	.13
Emotional Stability	6	1023	.26	.22	.20	.34	.26	.00	12.85	.35	.27
Openness	5	619	.18	.08	.00	.23	.00	.23	100.00	.25	.00

Notes: K = Number of validities in the meta-analysis; N = Total sample size across all validities; $\bar{r}$ = Sample size weighted mean effect size; SD_r = Standard deviation of the observed distribution; $SD_{residual}$ = Standard deviation of observed validities remaining after artifactual variance is removed; $\bar{\rho}_o$ = Operational validity (corrected for unreliability in criterion but not in predictors); SD_{ρ_o} = Standard deviation of operational validity after accounting for artifacts; 90% Credibility Value = Value of ρ_o at the 10^{th} percentile, computed using SD_{ρ_o}; % Variance Accounted = Percent of variance in the observed validities due to artifacts; $\bar{\rho}_T$ = True score correlation between individual difference predictors and overall assessment center ratings (corrected for unreliability in both predictors and criterion); SD_{ρ_T} = Standard deviation of true score correlation after accounting for all the artifacts.

Table 2: Moderator analysis of correlations between assessment center exercises and cognitive ability

Predictor	K	N	$\bar{r}$	SD_r	$SD_{residual}$	$\bar{\rho}_o$	SD_{ρ_o}	% Variance Accounted	90% Credibility Value of ρ_o
LGD									
Cognitive Ability	15	2697	.32	.12	.08	.57	.15	49.19	.37
In-basket									
Cognitive Ability	4	557	.36	.03	.00	.55	.00	100.00	.55

Notes: K = Number of validities in the meta-analysis; N = Total sample size across all validities; $\bar{r}$ = Sample size weighted mean effect size; SD_r = Standard deviation of the distribution; $SD_{residual}$ = Standard deviation in validities remaining after sampling error variance is removed; $\bar{\rho}_o$ = True score validity; SD_{ρ_o} = True variation after accounting for artifacts; % Variance Accounted = Percent of variance in the observed validity due to artifacts; 90% Credibility Value = Value at the 10^{th} percentile, computed using SD_{ρ_o}. This table presents operational validities only; construct level correlations are not presented.

Table 3: Observed correlations used to compute the multiple correlation

	GMA	A	E	ES	O	OAR
GMA	1.00					
A	.01	1.00				
E	.07	.04	1.00			
ES	.14	.25	.21	1.00		
O	.30	−.02	.40	−.02	1.00	
OAR	.65	.16	.47	.34	.23	1.00

Note: GMA = General Mental Ability; A = Agreeableness; E = Extraversion; ES = Emotional Stability; O = Openness to Experience; OAR = Overall Assessment Rating. The correlations in column 1 between General Mental Ability and the four personality variables were reported in Ackerman and Heggestad (1997). The values reported there are fully corrected for measurement error in both variables. We have attenuated them using the appropriate reliability of GMA and personality variables to obtain the observed correlations reported in this table. The correlations between the four personality variables are those reported in Costa and McCrae (1992) for their national norm group sample. The correlations between the OARs and GMA and the personality variables are those obtained in the present study (Table 1).

it seems unlikely that any personality trait, or even any combination of personality traits, could produce a correlation larger than .57. For the in-basket exercise, the operational validity was .55 ($SD_\rho o = .00$ and 100% of the variance was accounted for by artifacts. These results support the hypothesis proposed by others over the years and formally tested here that performance on in-basket tasks depends greatly on Cognitive Ability. We note, however, that this finding is based on only four correlations (although the total sample size is 557).

We estimated the relative importance of the traits to the OAR using standard regression procedures. In the first regression analysis, we regressed the OARs on the GMA and personality *measures*. In the second analysis, we regressed the OARs on the GMA and personality *constructs*. The first analysis addressed applied questions and the second, theoretical questions. The correlations among the independent variable measures were estimated from the literature. Specifically, the correlations among Cognitive Ability and the personality traits of Agreeableness, Extraversion, Emotional Stability, and Openness were taken from Ackerman and Heggestad (1997), and the intercorrelations among the four personality traits were those reported in Costa and McCrae (1992) for their national norm group sample. The correlations among the independent variables and the OARs are from the present meta-analyzed validities (Table 1). The correlation matrix for the observed score regression is shown in Table 3 and that for the trait-level regression is shown in Table 4. Results of the regression analyses are shown in Table 5.

As seen in Table 5, the observed score multiple correlation for predicting OARs from Cognitive Ability,

Table 4: Construct-level correlations used to compute the true score multiple correlation

	Reliability	GMA	A	E	ES	O	OAR
GMA	.95	1.00					
A	.86	.01	1.00				
E	.89	.08	.05	1.00			
ES	.92	.15	.28	.23	1.00		
O	.87	.33	−.02	.45	−.02	1.00	
OAR	.74	.67	.17	.50	.35	.25	1.00

Note: GMA = General Mental Ability; A = Agreeableness; E = Extraversion; ES = Emotional Stability; O = Openness to Experience; OAR = Overall Assessment Rating. The correlations in column 1 between General Mental Ability and the four personality variables were reported in Ackerman and Heggestad (1997). The values reported there are fully corrected for measurement error in both variables. We have attenuated them using the appropriate reliability of GMA and personality variables to obtain the observed correlations reported in this table. The correlations between the four personality variables are those reported in Costa and McCrae (1992) for their national norm group sample. The correlations between the OARs and GMA and the personality variables are those obtained in the present study (Table 1).

Table 5: Regression Models Predicting OAR from GMA and Personality Variables

	Beta weight GMA	Beta weight Agreeableness	Beta weight Extraversion	Beta weight Emotional Stability	Beta weight Openness	Multiple R
Operational Validity Model	.64	.10	.45	.13	−.14	.81
Construct Validity Model	.68	.11	.51	.10	−.20	.84

Agreeableness, Extraversion, Emotional Stability, and Openness is .81. The following beta weights were associated with each variable: .64 (Cognitive Ability), .10 (Agreeableness), .45 (Extraversion), .13 (Emotional Stability), −.14 (Openness to Experience).[1] The multiple correlation of .81 represents quantitatively the extent to which Cognitive Ability and personality *measures* can be used operationally to predict final OAR ratings. However, this value is likely to be a lower bound value because of our inability to include conscientiousness in the set of personality measures.

Table 5 also presents the trait-level beta weights and multiple correlation. The multiple correlation of .84 indicates how well OARs could be predicted from the actual traits of GMA, Agreeableness, Extraversion, Emotional Stability, and Openness to Experience. As noted earlier, this multiple correlation must be considered an underestimate because use of alpha reliability estimates (CE estimates) instead of CES reliability estimates results in under-corrections for measurement error. The true score multiple correlation could not be realized in an applied setting because measures of these traits free of measurement error are not operationally attainable. However, the results of this analysis are important theoretically, since this analysis estimates what is happening at the level of actual traits or constructs. Hence this analysis is more relevant than the observed score regression to the question of the construct meaning of assessment center OARs. Although the trait-level multiple correlation increases only marginally over that of the observed score regression (.84 − .81 = .03), the trait level regression shows that observed score regression underestimates the importance of GMA (beta of .64 instead of .68) and Extraversion (beta of .45 vs. .51 at the trait level). The construct level analysis makes it clearer that GMA and Extraversion are the most important traits affecting assessment center performance (as measured by OARs).

Discussion

These findings support our hypothesis that personality and cognitive ability substantially affect the overall assessment center ratings of candidates. Using Cohen's (1977) guidelines as a benchmark, the multiple correlation of .84 between Cognitive Ability, personality and OARs is large in magnitude. (Correlations of around .20 indicate a small effect; correlations approximating .50 suggest a moderate effect; and correlations of .70 or greater are considered large in magnitude.) This large correlation, generated using the meta-analyzed OAR validities, suggests that the constructs underlying OARs are Cognitive Ability and personality traits, probably operating at least in part through task and contextual forms of behavior in in-basket exercises, LGD exercises, and other exercises included in assessment centers (e.g., role playing exercises). Both the trait-level and observed score multiple correlations would probably have been larger if it had been possible to include conscientiousness in the personality traits studied. There is also a possibility that there are personality traits beyond the Big Five that would contribute to prediction. The present findings suggest the two most important traits determining assessment center performance are general intelligence and Extraversion. However, if more personality traits are included in the regression equation, findings could indicate a reduced role for Extraversion.

The finding of a large operational multiple correlation (.81) for predicting OARs from trait *measures* suggests that it may be possible to substitute standardized paper and pencil tests for more costly and lengthy assessment centers, at least for purposes of selection. This result supports previously proposed hypotheses that until this study were not formally tested (Bass, 1954; Klimoski and Strickland (1981), as cited in Klimoski and Brickner, 1987; Tziner and Dolan, 1982).

Based on a survey of 215 organizations, Slavenski (1986) estimated that assessment center costs are approximately $1,730 per assessee; he also reported another estimate for entry-level marketing and management personnel of $4,000 per assessee. These expenses are magnified by the fact that a large percentage of individuals within an organization participate more than once in the assessment center (Spychalski, Quiñones, Gaugler, and Pohley, 1997). Furthermore, these estimates under-represent the actual cost because they do not include the expenses incurred for the development and validation of the assessment center exercises, the training of assessors and administrators, the time-off-task from regular jobs to run the assessment centers, the continuous up-grading of assessment center

exercises, and other unknown and unanticipated costs. These additional and omitted costs are difficult to estimate, but they can be substantial for many organizations.

These cost factors and the findings in the present study raise the question of the appropriate role for assessment centers. In recent years, assessment centers have been used to select into training programs (Moulton, 1993), develop employee skills (Rea, Rea, and Moomaw, 1990), sort employees into skill levels (Jackson, 1985), identify employees who are team-oriented (Kirksey and Zawacki, 1994), and for executive development (Moulton, 1993). For these and other developmental activities where employee strengths and weaknesses can be evaluated within the different exercises, assessment centers can be useful. For example, exercises designed for the selection of teams based on group interactions or for selection into training programs in which participants learn to emphasize or express their job-related characteristics (e.g., learning to behave in an extraverted manner for a particular sales job). For these (and perhaps other) personnel purposes, assessment centers can be useful. However, one of the primary purposes for assessment centers, especially for larger corporations, continues to be for personnel selection. Despite higher costs, AC use in selection has been defended on the grounds of reduced adverse impact for minorities (in comparison with ability tests) (Baron and Janman, 1996) and more positive applicant reactions than for ability tests and personality inventories (Macan, Avedon, Paese, and Smith, 1994). Nevertheless, for purposes of selection, the findings of this study suggest that paper and pencil tests might be an economically and psychometrically viable substitute in some cases.

Note

1. The negative beta weight for Openness to Experience indicates that Openness to Experience is functioning as a suppressor variable. The validity of Openness for the prediction of the OAR is .24, but its (observed) correlation with Extraversion is .40. The negative beta weight indicates that Openness is suppressing invalid variance in Extraversion, thereby increasing the beta weight on Extraversion. For a detailed discussion of suppressor variables, see Collins and Schmidt (1997).

Appendix A

Artifact Distributions Used in the Meta-Analysis: Interrater Reliabilities and Range Restrictions

Interrater Reliabilities Used to Correct for Criterion (OAR) Unreliability

Interrater Reliability	Study
.68	Frederiksen, N., Saunders, D.R. and Wand, B. (1957). The in-basket test. *Psychological Monographs: General and Applied, 71(9)*, 1–28.
.69	Frederiksen *et al*. (1957).
.83	Frederiksen *et al*. (1957).
.73	Schectman, Z. (1992). A group assessment procedure as a predictor of on-the-job performance of teachers. *Journal of Applied Psychology, 77*, 383–387.
.82	Tett, R.P. and Jackson, D.N. (1990). Organization and personality correlates of participative behaviors using an in-basket exercise. *Journal of Occupational Psychology, 63*, 175–188.
.85	Lowry, P.E. (1994). Selection methods: Comparison of assessment centers with personnel records evaluations. *Public Personnel Management, 23(3)*, 383–395.
.84	Tziner, A. and Dolan, S. (1982). Validity of an assessment center for identifying future female officers in the military. *Journal of Applied Psychology, 67(6)*, 728–736.
.56	Pynes, J.E. and Bernardin, H.J. (1989). Predictive validity of an entry-level police officer assessment center. *Journal of Applied Psychology, 74(5)*, 831–833.
.68	Bray, D.W. and Grant, D.L. (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs: General and Applied, 80(17)*, Whole No. 625, 1–27.
.76	Borman, W.C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology*, 67(1), 3–9.
.68	Schmitt, N. (1977). Interrater agreement in dimensionality and combination of assessment center judgments. *Journal of Applied Psychology, 62(2)*, 171–176.

Appendix B

Artifact Distribution for Range Restriction: Cognitive Ability

(Alexander, R.A., Carson, K.P., Alliger, G.M. and Cronshaw, S.F. [1989]. Empirical Distributions of Range Restricted SDx in Validity Studies. *Journal of Applied Psychology, 74*, 253–258.)

u value	frequency
.559	5
.603	15
.649	20
.701	20
.766	20
.849	15
1.000	5

Note: u = s/S, where s = the restricted standard deviation of the independent variable and S = the unrestricted standard deviation of the independent variable.

References

Ackerman, P.L. and Heggestad, E.D. (1997) Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, **121**, 219–245.

Alexander, R.A., Carson, K.P., Alliger, G.M. and Cronshaw, S.F. (1989) Empirical distributions of range restricted SDx in validity studies. *Journal of Applied Psychology*, **74**, 253–258.

American Psychological Association (1999) *APA Test Standards*. Washington, DC: APA.

Archambeau, D.J. (1979) Relationships among skill ratings assigned in an assessment center. *Journal of Assessment Center Technology*, 7–20.

Baron, H. and Janman, K. (1996) Fairness in the assessment center. *International Review of Industrial and Organizational Psychology* (Vol. 11, pp. 61–114). Chichester: John Wiley & Sons Ltd.

Barrick, M.R. and Mount, M.K. (1991) The Big Five personality dimensions and job performance. *Personnel Psychology*, **44**, 1–26.

Barrick, M.R. and Mount, M.K. (1996) Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology*, **81**, 261–272.

Bass, B.M. (1951) Situational tests, II: Variables of the leaderless group discussion. *Educational and Psychological Measurement*, **11**, 196–207.

Bass, B.M. (1954) The leaderless group discussion. *Psychological Bulletin*, **51**, 465–492.

Bass, B.M. and Coates, C.H. (1952) Forecasting officer potential using the leaderless group discussion. *Journal of Abnormal Social Psychology*, **47**, 321–325.

Bass, B.M., McGehee, C.R., Hawkins, W.C., Young, P.C. and Gebel, A. Personality variables related to leaderless group discussion behavior. *Journal of Abnormal Psychology*, **48**, 120–128.

Bass, B.M. and Wurster, C.R. (1953) Effects of company rank on LGD performance of oil refinery supervisors. *Journal of Applied Psychology*, **37**, 100–104.

Bass, B.M., Wurster, C.R., Doll, P.S. and Clair, D.J. (1953) Situational and personality factors in leadership among sorority women. *Psychological Monographs*, 67 (Whole No. 366).

Borman, W.C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 67(1)*, 3–9.

Borman, W.C. and Motowidlo (1993) Expanding the criterion domain to include elements of contextual performance. In Neal Schmitt and Walter C. Borman and Associates (eds), *Personnel Selection in Organizations*, pp. 71–98. San Francisco: Jossey-Bass.

Brannick, M.T., Michaels, C.E., Baker, D.P. (1989) Construct validity of in-basket scores. *Journal of Applied Psychology*, **73**, 736–742.

Bray, D.W. (1964) The management progress study. *American Psychologist*, **19**, 419–420.

Bray, D.W. and Campbell, R.J. (1968) Selection of salesmen by means of an assessment center. *Journal of Applied Psychology*, **52**, 36–41.

Bray, D.W. and Grant, D.L. (1966) The assessment center in the measurement of potential for business management. *Psychological Monographs*, 80 (17, Whole No. 625).

Carlton, F.O. (1970) Relationships between follow-up evaluations and information developed in a management assessment center. Paper presented at the annual meeting of the American Psychological Association, Miami Beach, FL.

Cohen, J. (1977) *Statistical Power Analysis for the Social Sciences*. New York: Academic Press.

Collins, J.M. and Schmidt, F.L. (1997) Can suppressor variables enhance criterion-related validity in the personality domain? *Educational and Psychological Measurement*, **57**, 924–936.

Costa, P.T., Jr. and McCrae, R.R. (1992) *Revised NEO Five-Factor Inventory Professional Manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Crawley, B., Pinder, R., Herriott, P. (1990) Assessment center dimensions, personality, and aptitudes. *Journal of Occupational Psychology*, **63**, 211–216.

Cronbach, L.J. and Meehl, P.E. (1995) Construct validity in psychological tests. *Psychological Bulletin*, **52**, 281–302.

Derlega, V.J., Winstead, B.A. and Jones, W.H. (1991) Personality: an introduction. In Valerian J. Derlega, Barbara A. Winstead and Warren H. Jones (eds), *Personality: Contemporary Theory and Research*, p. 3. Chicago: Nelson-Hall Inc.

Dugan, B. (1988) Effects of assessor training on information use. *Journal of Applied Psychology*, **73**, 743–748.

Dulewicz, V. and Fletcher, C. (1982) The relationship between previous experience, intelligence and background characteristics of participants and their performance in an assessment centre. *Journal of Occupational Psychology*, **55**, 197–207.

Fitzgerald, L.F. and Quaintance, M.K. (1982) Survey of assessment center use in state and local government. *Journal of Assessment Center Technology*, **5**, 9–21.

Fletcher, C.A. and Dulewicz, V. (1984) An empirical study of a UK-based assessment centre. *Journal of Management Studies*, **21**, 83–97.

Frederiksen, N., Saunders, D.R. and Wand, B. (1957) The in-basket test. *Psychological Monographs: General and Applied*, **71**(9), 1–28.

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

Gaugler, B.B. and Thornton, G.C. (1989) Number of

assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology*, **74**, 611–618.
Goldberg, L.R. (1990) An alternative 'Description of Personality': The Big-Five factor solution. *Journal of Social and Personality Psychology*, **59**, 1216–1229.
Gough, H.G. (1996) Theory, development, and interpretation of the CPI Socialization Scale. *Psychological Reports*, Monograph Supplement 1-V75, 651–700.
Guion, R.M. (1991) Personnel assessment, selection, and placement. In Marvin D. Dunnette and Leaetta M. Hough (eds), *Handbook of Industrial and Organizational Psychology*. Palo Alto, CA: Consulting Psychologists Press, Inc.
Guion, R.M. (1998) *Assessment, Measurement, and Prediction for Personnel Decisions*. Mahwah, NJ: Lawrence Erlbaum Associates.
Hakstian, A.R., Woolsey, L.K. and Schroeder, M.L. (1986) Development and application of a quickly-scored in-basket exercise in an organizational setting. *Educational and Psychological Measurement*, **46**, 385–396.
Hardesty, D.L. and Jones, W.S. (1968) Characteristics of judged high potential management personnel – the operations of an industrial assessment center. *Personnel Psychology*, **21**, 85–98.
Hogan, R. and Hogan, J. (1992) *Hogan Personality Inventory Manual*. Tulsa, OK: Hogan Assessment Systems, Inc.
Howard, A. (1983) Measuring management abilities and motivation. *New Directions for Testing and Measurement*, **17**, 31–44.
Howard, A. and Bray, D. (1988) *Managerial Lives in Transition: Advancing Age and Changing Ttimes*. New York: Guilford Press, Inc.
Huck, J.R. (1973) Assessment centers: A review of the external and internal validities. *Personnel Psychology*, **26**, 191–212.
Huck, J.R. and Bray, D.W. (1976) Management assessment center evaluations and subsequent job performance of white and black females. *Personnel Psychology*, **29**, 13–30.
Hunter, J.E. and Hunter, R.F. (1984) Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, **96**, 72–98.
Hunter, J.E. and Schmidt, F.L. (1990) *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage.
Hunter, J.E. and Schmidt, F.L. (1996) Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, and Law*, **2**, 447–472.
Jackson, C.N. (1985) Training's role in the process of planned change. *Training and Development Journal*, **39**, 70–74.
Jensen, A.R. (1980) *Bias in Mental Testing*. New York: Macmillan.
Jones, A., Herriot, P., Long, B. and Drakeley, R. (1991) Attempting to improve the validity of a well-established assessment centre. *Journal of Occupational Psychology*, **64**, 1–21.
Kelly, E.L. and Fiske, D.W. (1951) *The Prediction of Performance in Clinical Psychology*. New York: Greenwood Press.
Kirksey, J. and Zawacki, R.A. (1994) Assessment center helps find team-oriented candidates. *Personnel Journal*, **73**, 92.
Klimoski, R.J. and Brickner, M. (1987) Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, **40**, 243–260.
Klimoski, R.J. and Strickland, W.J. (1977) Assessment centers: Valid or merely prescient. *Personnel Psychology*, **30**, 353–363.
Klimoski, R.J. and Strickland, W.J. (1981) A comparative view of assessment centers: A case analysis. Unpublished manuscript.
Kraut, A.I. (1973) Management assessment in international organizations. *Industrial Relations*, **12**, 172–182.
Landau, S.I. and Bogus, R.J. (eds) (1990) *The Bantam Roget's Thesaurus*. New York: Bantam.
Le, H. and Schmidt, F.L. (2001a) The multi-faceted nature of measurement error and its implications for measurement error corrections: The case of job satisfaction. Paper under review.
Le, H. and Schmidt, F.L. (2001b) Discriminant validity of measures of job satisfaction and organizational commitment: A mirage due to measurement error? Paper under review.
Levy-Leboyer, C. (1963) Social behavior and individual characteristics: A study of members of a controlled group. *Monographies Françaises de Psychologie*, **10**, 105.
Lievens, F. and Klimoski, R.J. (2001) Understanding the assessment centre process: Where are we now? In C.L. Cooper and I.T. Robertson (eds), *International Review of Industrial and Organizational Psychology* (Vol. 16, pp. 245–286). Chichester: John Wiley & Sons Ltd.
Lowry, P.E. (1994) Selection methods: Comparison of assessment centers with personnel records evaluations. *Public Personnel Management*, **23**, 383–395.
Macan, T.H., Avedon, M.J., Paese, M. and Smith, D.E. (1994) The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology*, **47**, 715–738.
Maukisch, H. (1986) Erfolgskontrollen von Assessment Center-systemen: Der Stand der Forschung. *Organisations Psychologie*, **30**, (NF4) 2, 86–91.
McConnell, J.J. and Parker, T.C. (1972) An assessment center program for multi-organizational use. *Training and Development Journal*, **March**, 6–14.
McEvoy, G.M. and Beatty, R.W. (1989) Assessment centers and subordinate appraisals of managers: A seven-year examination of predictive validity. *Personnel Psychology*, **42**, 37–52.
McHenry, J.J., Hough, L.M., Toquam, J.L., Hanson, M.A. and Ashworth, S. (1990) Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology*, **43**, 335–353.
Mead, D.F. and Christal, R.E. (1970) Development of a constant standard weight equation for evaluating job difficulty. *USAF AFHRL Technical Report*. No. 70-44, 11 p, November.
Meyer, H.H. (1970) The validity of the in-basket test as a measure of managerial performance. *Personnel Psychology*, **23**, 297–307.
Mitchel, J.O. (1975) Assessment center validity: A longitudinal study. *Journal of Applied Psychology*, **60**, 573–579.
Moses, J.L. (1973) The development of an assessment center for the early identification of supervisory potential. *Personnel Psychology*, **26**, 569–580.
Motowidlo, S.J. and Van Scotter, J.R. (1994) Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology*, **79**, 475–480.
Moulton, H.W. (1993) *Executive Development: Preparing for the 21st Century*. New York: Oxford University Press.
Muchinsky, P.M. (1990) *Psychology Applied to Work*, 3rd edn. Pacific Grove, CA: Brooks/Cole Publishing Company.
Outclat, D. (1988) A research program on General Motors' foreman selection assessment center: Assessor/assessee characteristics and moderator analysis. Paper presented at the 16th International Congress on the Assessment Center Method, Tampa, FL.
Peterson, N.G., Hough, L.M., Dunnette, M.D., Rosse, R.L., Houston, J.S., Toquam, J.L. and Wing, H. (1990) Project A:

Specification of the predictor domain and development of new selection/classification tests. *Personnel Psychology*, **43**, 247–276.
Pulakos, E.D., Borman, W.C. and Hough, L.M. (1988) Test validation for scientific understanding: Two demonstrations of an approach to studying predictor-criterion linkages. *Personnel Psychology*, **41**, 703–716.
Pynes, J.E. and Bernardin, H.J. (1989). Predictive validity of an entry-level police officer assessment center. *Journal of Applied Psychology, 74(5)*, 831–833.
Rea, P., Rea, J. and Moomaw, C. (1990) Use of assessment centers in skill development. *Personnel Journal*, **69**, 126.
Rousseau, D.M. and Aquino, K. (1993) Fairness and implied contract obligations in job terminations: The role of remedies, social accounts, and procedural justice. *Human Performance*, **6**, 135–149.
Russell, C.J. (1985) Individual decision process in an assessment center. *Journal of Applied Psychology*, **70**, 737–746.
Russell, C.J. (1987) Person characteristic vs. role congruency explanations for assessment center ratings. *Academy of Management Journal*, **30**, 817–826.
Russell, C.J. and Kuhnert, K.W. (1992) New frontiers in management selection systems: Where measurement technologies and theory collide. *Leadership Quarterly*, **3**, 109–135.
Sackett, P.R. and Dreher, G.F. (1982) Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology*, **67**, 401–410.
Sackett, P.R. and Harris, M.M. (1988) A further examination of the constructs underlying assessment center ratings. *Journal of Business Psychology*, **3**, 214–219.
Schectman, Z. (1992). A group assessment procedure as a predictor of on-the-job performance of teachers. *Journal of Applied Psychology, 77*, 383–387
Schmidt, F.L. (1999) Measurement error and cumulative knowledge. Paper presented at Purdue University, 26 March.
Schmidt, F.L. and Hunter, J.E. (1981) Employment testing: Old theories and new research findings. *American Psychologist*, **36**, 128–137.
Schmidt, F.L. and Hunter, J.E. (1992) Development of causal models of processes determining job performance. *Current Directions in Psychological Science*, **1**, 89–92.
Schmidt, F.L. and Hunter, J.E. (1996) Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, **1**, 199–223.
Schmidt, F.L. and Hunter, J.E. (1998) The validity and utility of selection methods in personnel psychology. Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, **124**, 262–274.
Schmidt, F.L. and Hunter, J.E. (1999) Theory testing and measurement error. *Intelligence*, **27**, 183–198.
Schmidt, F.L., Law, K., Hunter, J.E., Rothstein, H.R., Pearlman, K. and McDaniel, M. (1993) Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology*, **78**, 438–443.
Schmidt, F.L., Le, H. and Ilies, R. (2001) Beyond Alpha: An empirical examination of the effects that different sources of measurement error have on reliability estimates for individual differences constructs. Manuscript under review.
Schmitt, N. (1977) Interrater agreement in dimensionality and combination of assessment center judgments. *Journal of Applied Psychology*, **63**, 171–176.
Schmitt, N., Gooding, R., Noe, R. and Kirsch, M. (1984) Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, **37**, 407–422.
Scholz, G. and Schuler, H. (1993) Das nomologische Netzwerk des Assessment Centers: Eine Metaanalyse (The nomological network of the assessment center: A meta-analysis). *Organisations Psychologie*, **37**, (N.E. 11) 2, 73–85.
Shore, T.H. and Thornton, G.C. and Shore, L. (1990) Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology*, **43**, 1–12.
Slavenski, L. (1986) Matching people to the job. *Training and Development Journal*, **August**, 54–57.
Spychalski, A.C., Quiñones, M.A., Gaugler, B.B. and Pohley, K. (1997) A survey of assessment center practices in organizations in the United States. *Personnel Psychology*, **30**, 71–90.
Sternad, J.A. (1976) Use of the 16PF to predict pre-service principals' assessment center scores. Unpublished dissertation. University of Akron, Ohio.
Stopler, R. (1953) An investigation of some hypotheses concerning empathy. Unpublished master's thesis. Louisiana State University.
Tett, R.P. and Jackson, D.N. (1990) Organization and personality correlates of participative behaviors using an in-basket exercise. *Journal of Occupational Psychology*, **63**, 175–188.
Thornton, G.C. III (1992) *Assessment Centers in Human Resource Management*. Reading, MA: Addison-Wesley.
Thornton, G.C. and Byham, W.C. (1982) *Assessment Centers and Managerial Performance*. New York: Academic Press.
Turnage, J.J. and Muchinsky, P.M. (1982) Transituational variability in human performance with assessment centers. *Organizational Behavior and Human Performance*, **30**, 174–200.
Tziner, A. (1984) Prediction of peer rating in a military assessment center. *Canadian Journal of Administrative Science*, **1**, 146–160.
Tziner, A. and Dolan, S. (1982) Validity of an assessment center for identifying future female officers in the military. *Journal of Applied Psychology*, **67**, 728–736.
Vernon, P.E. (1948) The validation of civil service observation method in the selection of trainee executives. *Occupational Psychology*, **32**, 587–594.
Vernon, P.E. (1950) The validation of civil service selection board procedures. *Occupational Psychology*, **24**, 75–95.
Werner, J.M. and Bolino, M.C. (1997) Explaining US courts of appeals decisions involving performance appraisal: Accuracy, fairness, and validation. *Personnel Psychology*, **50**, 1–24.
Wilson, J.E. and Tatge, W.A. (1973) Assessment centers – Further assessment needed? *Personnel Journal*, 172–179.
Wurster, C.R. and Bass, B.M. (1953) Situational tests, IV: Validity of leaderless group discussion among strangers. *Educational and Psychological Measurement*, **13**, 122–132.