Journal of Applied Psychology
1987, Vol. 72, No. 3, 493-511

0021-9010/87/$00.75

Journal of Applied Psychology Monograph

Meta-Analysis of Assessment Center Validity

Barbara B. Gaugler, Douglas B. Rosenthal, George C. Thornton III, and Cynthia Bentson
Colorado State University

Meta-analysis (Hunter, Schmidt, & Jackson, 1982) of 50 assessment center studies containing 107 validity coefficients revealed a corrected mean and variance of .37 and .017, respectively. Validities were sorted into five categories of criteria and four categories of assessment purpose. Higher validities were found in studies in which potential ratings were the criterion, and lower validities were found in promotion studies. Sufficient variance remained after correcting for artifacts to justify searching for moderators. Validities were higher when the percentage of female assessees was high, when several evaluation devices were used, when assessors were psychologists rather than managers, when peer evaluation was used, and when the study was methodologically sound. Age of assessees, whether feedback was given, days of assessor training, days of observation, percentages of minority assessees, and criterion contamination did not moderate assessment center validities. The findings suggest that assessment centers show both validity generalization and situational specificity.

Since the first industrial application of assessment centers in 1956 by AT&T (Bray & Grant, 1966), a growing number of companies have used the assessment center method. Today, it is estimated that more than 2,000 organizations are currently using some type of assessment center program. Organizations use assessment centers for a wide variety of purposes, including selection, placement, early identification of management potential, promotion, development, career management and training. Although assessment centers are most frequently used for assessing managers, they have also been developed to assess college students, engineers, salespersons, military personnel, rehabilitation counselors, school administrators, and blue-collar workers.

The increasing popularity of the assessment center method has stimulated a great amount of research concerning its effectiveness. Reviewers have accumulated research findings from a variety of types of assessment centers and have concluded that assessment centers have predictive validity for a variety of criteria (Byham, 1970; Cohen, Moses, & Byham, 1977; Howard, 1974; Thornton & Byham, 1982). Although the predictive validity coefficients of assessment centers are generally high, some assessment centers have low predictive validity. In fact, the observed validity coefficients of the assessment centers reviewed by us ranged from −.25 to +.78. The current meta-analysis was designed to estimate the true validity of assessment centers and to understand the reasons for the variability in observed predictive validity coefficients.

Meta-analysis is a collection of methods used to aggregate results across studies quantitatively. It helps us draw more accurate conclusions about inconsistent findings in a particular area of research. Statistical procedures replace the traditional literature review, which has been criticized for its "great information-gate-keeping potential" (Cooper & Rosenthal, 1980, p. 442). Literature reviews are highly influenced by the biases of the reviewer, may neglect large amounts of information provided in the original research reports, and imprecisely weight conclusions with regard to the amount of research covered. Statistical techniques of aggregation have been suggested as an alternative to the literature review.

There are a number of different meta-analytic procedures that vary in sophistication (Rosenthal, 1978). The method developed by Schmidt and Hunter (1977) was used in this study for a variety of reasons. First, it provides specific formulas for statistically cumulating effect sizes across studies. Second, it was developed specifically for use with correlational data (e.g., validity coefficients). Third, it rests on the assumption that much of the variation in observed results is due to statistical artifacts and methodological problems rather than to true differences in underlying population correlations. Artifacts include sampling error due to studies having sample sizes less than infinity, unreliability of predictor and criterion measurement, differential restriction of range across samples, and various typographical and other data analysis and reporting errors. Formulas are used to estimate the amount of true variation in validity coefficients and the amount of observed variation that is due to artifacts.

Validity generalization studies of other selection techniques suggest that a substantial amount of the variability in predictive validities is due to statistical artifacts. Studies of programmer and clerical aptitude tests (Pearlman, Schmidt, & Hunter, 1980;

Parts of this article were presented at the 93rd Annual Convention of the American Psychological Association held in August 1985 in Los Angeles.

We would like to acknowledge the helpful comments of three anonymous reviewers.

Douglas B. Rosenthal is now at HUMRRO in Alexandria, Virginia. Cynthia Bentson is now at Bentson Associates, Fort Collins, Colorado.

Correspondence concerning this article should be addressed to Barbara B. Gaugler, who is now at Rice University, Psychology Department, Houston, Texas 77251.

Schmidt, Gast-Rosenberg, & Hunter, 1980), mechanical and chemical comprehension tests (Schmidt, Hunter, & Caplan, 1981), weighted biographical inventories (Brown, 1981), tests of general intelligence (Pearlman et al., 1980; Schmidt et al., 1981), and also verbal, quantitative, reasoning, spatial/mechanical and motor ability, perceptual speed, memory, and performance tests (Pearlman et al., 1980) have shown that at least 60% of the variation in single predictor–criterion relations can be accounted for by sampling error, predictor and criterion unreliability, and range restriction.

The predictive validity of assessment centers may be susceptible to the same artifacts. First, sampling error contributes to variability in validity coefficients. Because of the expense and time required by the assessment center process, many of the studies of the criterion validity of assessment centers have relatively small sample sizes. Thornton and Byham (1982) reported that sample sizes varied from 12 to 5,943, with a median of 55. Most studies used 40 to 50 candidates, and only a few studies had over 100 subjects. Because sampling error accounted for most of the artifactual variance in the previous validity generalization studies, and because sample sizes of assessment center studies are relatively small and vary to such an extent, it was predicted that sampling error would account for much of the variance in assessment centers' predictive validity coefficients.

Second, assessment center studies show moderate to severe levels and variation in range restriction. When assessment center results are used for operational purposes, not everyone who is assessed is selected or promoted.

Reliability of supervisory ratings of performance and potential, a common criterion in assessment center studies, may be low. As Thornton and Byham (1982) noted, "Problems with supervisors' ratings are legion. Leniency, halo, and restriction-in-range biases may occur" (p. 298). In addition, the reliability of the criterion measured to validate assessment centers varies considerably. For example, the mean reliability of the criteria reviewed for this meta-analysis ranged from .61 to 1.00.

Although there is reason to believe that variability in validity coefficients of assessment centers may be partially due to methodological artifacts, the diversity in makeup of assessment centers suggests the possibility that certain variables may moderate the predictive validity of assessment centers. There is such a variety of assessment center procedures that a *typical* assessment center does not exist (L. Alexander, 1979; Bender, 1973; Byham, 1978a, 1978b; Thornton & Byham, 1982). "There is no standardization of content in assessment centers or of the way they are administered, and there is no uniform method of treating the performance evaluation data generated by assessment centers" (Bender, 1973, p. 56). Such variety makes informed comparisons across studies extremely difficult. Meta-analysis provides a quantitative method of examining correlates of the predictive validity coefficients.

Many of the individual investigations of moderators of assessment centers, using single samples, have yielded conflicting and inconclusive evidence. Moderators that have been tested in individual studies include the following: candidate's age (Burroughs, Rollins, & Hopkins, 1973; Neidig, Martin, & Yates, 1978), candidate's minority group status (S. Alexander, 1975; Clingenpeel, 1979; Huck, 1974; Huck & Bray, 1976; Jaffee, Cohen, & Cherry, 1972; Marquardt, 1976; Moses, 1973a, 1973b; Moses & Boehm, 1975; Russell, 1975), sex of candidate (S. Alexander, 1975; Clingenpeel, 1979; Hall, 1976; Marquardt, 1976; Moses, 1973a, 1973b; Moses & Boehm, 1975), composition of the assessee group (Byham, 1981; Schmitt & Hill, 1977), type of criterion (Klimoski & Strickland, 1977, 1981), and time at which criterion measures are taken (Finley, 1970; Hinrichs, 1978; A. Howard, personal communication, February 16, 1979; Mitchel, 1975; Moses, 1972; Slivinski & Bourgeois, 1977). In addition, other parameters have been suggested for investigation (Thornton & Byham, 1982): types of evaluation devices, operating procedures, ratio of assessees and assessors, evaluation of observed dimensions, process of integrating information, uses made of performance data, and purposes of assessment. Meta-analysis provides a method of examining variability in validity coefficients *across* studies of different populations.

In the present investigation, both validity generalization (i.e., whether the lower bound of some confidence interval around the average validity is greater than zero) and situational specificity (i.e., whether nonartifactual variance in validities exist) were studied. It is possible and meaningful to find any combination of results. Hunter, Schmidt, and Jackson's (1982) meta-analytic procedures were applied to the results of 50 studies that investigated the relation between the overall assessment rating and various criteria. The purpose was threefold: (a) to estimate the true validity of assessment centers, (b) to determine the extent to which varied results across studies are due to statistical artifacts and methodological problems, and (c) to discover which characteristics of assessment centers moderate the predictive validity of assessment centers.

Method

Selection of Studies

A review of the literature was undertaken using *Psychological Abstracts*, reference lists of previous reviews (Cohen et al., 1977; Howard, 1974; Huck, 1973; Thornton & Byham, 1982), and personal contact with primary researchers in the field. From this pool, published and unpublished studies were selected to be included in the meta-analysis that met the following criteria: (a) The manuscript described an assessment center, as delineated by the Standards for Assessment Centers (Task Force on Assessment Center Standards, 1980), and (b) a correlation between the overall assessment rating and some criterion was provided or calculable from the data given. Studies included experimental studies in which there was no operational use of assessment center data, studies with no feedback to participants, studies that compared the subsequent performance of assessed and nonassessed groups (i.e., control groups), correlational studies with feedback to assessees and management, and concurrent validity studies. No study was excluded on the basis of poor method or quality. However, the quality of various design features, adequacy of information provided, and external validity of each study were rated by the authors.

Ratings of Characteristics

A number of variables were believed on theoretical or empirical grounds, or both, to contribute to the relation between the overall assessment rating and various criteria. We also examined demographic and other variables of interest. The following information was recorded for each study: (a) identification information—coder identification number, study identification number, effect size identification number, publication year, publication form, and country of study; (b) candidate characteristics—average age at time of assessment, educational level,

current position, percentage of men, and percentage with minority status; (c) assessment center description—types of assessment techniques used (e.g., in-basket, leaderless group discussion), number of types of assessment techniques, ratio of assessees to assessors, names of dimensions assessed, process of integrating information, uses made of assessment data, purposes of assessment, types of criteria, and time lag between the assessment center and when the criterion was measured; (d) study design and reporting—study design (e.g., experimental), number of assessees, number of assessees on whom criterion measures were taken, reliability of the overall assessment rating, reliability of the criterion, presence of a systematic method of identifying dimensions, presence of potential restriction of range, potential for criterion contamination through knowledge of assessment results, threats to the validity of research (e.g., inadequate representation; Cook & Campbell, 1976), general index of validity of research, and adequacy of information provided in the report; and (e) conclusions—uncorrected correlation value, statistic given, and author's conclusion about calculated correlation.

The four authors of this article served as coders of the characteristics of the studies. Following training and practice, interjudge agreement in coding among the four authors was calculated, once prior to and once during the coding of the studies. Because the characteristics were coded on a variety of scales of measurement, several indices of interrater reliability were needed. Acceptable levels of interrater agreement (i.e., >85%) were attained for all types of variables for both assessments. For example, for categorical variables, there was total agreement on 78% of the items, agreement among three judges on 11% of the items, and agreement by two judges on 11% of the items. Kuder-Richardson 20 reliabilities performed on the dichotomous items averaged .91 and .88 for the two interrater reliability assessments. Detailed data are available from the first author.

Analytic Procedures

Combining validities within studies. Many of the studies reported multiple validity coefficients. In some cases, researchers obtained several independent samples of assessees and calculated separate validities on each sample. Validities from multiple samples were considered statistically independent and were therefore entered unchanged into the cumulation formulas. (Validities for multiple samples may not be independent because of similarities in exercises, biases of assessors, biases of supervisors providing criterion ratings, and generalized features of the organization climate. Supplementary analyses were conducted that combined validities judged to be dependent within a research report.) More frequently, however, researchers used several criterion measures for the same sample (e.g., supervisor performance ratings, salary advancement, number of promotions). Validities calculated on the same sample were considered statistically dependent and therefore were combined, following the recommendations of Hunter et al. (1982, p. 118). In most cases, a simple mean was calculated across dependent validities within a single study. In a few cases, intercorrelations among criteria were reported that allowed us to compute a composite validity (Hunter et al., 1982, p. 120). The advantage of using a composite validity instead of a mean validity is that it reflects the validity the researchers would have obtained, had they originally summed each assessee's criterion scores, and then correlated them with overall assessment ratings. Dependent validities were only combined within five general categories of criteria. Table 1 presents the five categories and the individual criteria subsumed under each.

Some studies reported multiple validities on the same sample from a single criterion measured at various times (e.g., number of promotions in the 1 year, 5 years, and 10 years since the assessment center). In a preliminary analysis, a small, nonsignificant correlation between the time when criteria were obtained and the magnitude of validities was found. (Concurrent studies were excluded from this analysis.) We therefore decided to combine all validities taken at different times for the same criterion category within studies. Because subject attrition over time generally occurred in these studies, a mean sample size was calculated for each study.

Table 1
Five Categories of Criteria

Category	Criteria within category
Rating of job performance	An overall performance rating
	Field observation of manager's performance
	Field interview with supervisor of the manager
	A rating on some aspect of job performance other than an assessment center dimension
Potential ratings	A rating of manager's potential
Dimension ratings	Rating of manager's job performance on the dimensions used in the assessment center
Performance in training	Performance of manager in a training program
Career advancement	Change in salary over time
	Absolute level of salary obtained
	Number of promotions
	Absolute job level obtained
	Turnover

Cumulating effect sizes across studies. First, the mean validity and variance of validities weighted by sample size were calculated. Thus, large studies are given more importance than small ones. Second, the mean and variance were corrected for statistical artifacts. Using distributional formulas presented by Hunter et al. (1982, p. 90), the weighted mean validity was corrected for restriction of range and unreliability in the criteria.[1] The validities were not corrected for predictor unreliability because we were unable to obtain a reasonable estimate of the distribution of reliabilities for the overall assessment ratings, and to correct for unreliability in the overall assessment rating would yield a mean validity that assumes overall assessment ratings are perfectly reliable. Such a mean would be an overestimate of the validity of assessment centers as currently practiced. It is important to note that variability in the unreliability in the overall assessment ratings is an artifactual source of variance in assessment center validities and, ideally, should be removed from the variance. However, we could not do this without reliability estimates of the overall assessment ratings. The correction formulas used in this study are the same correction formulas used by Schmidt, Hunter, and their colleagues in some of their early validity generalization work on personnel selection research (see Hunter et al., 1982, p. 91). In these validity generalization studies the selection tests were treated as fixed (i.e., variance due to predictor unreliability was ignored).

Table 2 presents the criterion reliability estimates used to correct the validity means and variances. Means and variances of the square roots of the reliabilities are presented because they are the actual numbers used in the correction formulas. It was assumed that the reliability distributions for performance ratings, ratings of potential, and dimension ratings would be identical because they are all ratings of on-the-job performance. This reliability estimate was calculated by combining the reliabilities for performance, potential, and dimension ratings reported in our assessment center studies and reliabilities from other research on performance evaluation. The result was a list of 286 reliabilities for which means and variances were computed.[2] Our estimate of reliabili-

[1] A typographical error appears in Hunter, Schmidt, and Jackson (1982, p. 90) for the formula to compute "c". It should read $\bar{r}^2$, not $\bar{r}$. The analyses appearing in the present article were computed using the correct formula.

[2] These are available from the second author.

Table 2
Means and Variances of Assumed Reliability Distributions

Criterion	Mr_{yy}	$M\sqrt{r_{yy}}$	$\sigma_{r_{yy}}^2$	$\sigma^2\sqrt{r_{yy}}$
Performance, potential, and dimension ratings	.61	.774	.034	.016
Performance in training	.80	.894	.007	.002
Career advancement	1.00	1.000	.000	.000
Total sample	.77	.871	.039	.015

Note. Mr_{yy} = estimate of mean criterion reliability; $M\sqrt{r_{yy}}$ = estimate of the mean of the square roots of criterion reliabilities; $\sigma_{r_{yy}}^2$ = variance of criterion reliabilities; $\sigma^2\sqrt{r_{yy}}$ = variance of the square roots of criterion reliabilities.

ties for training criteria came from an assumed distribution reported by Pearlman et al. (1980, p. 375). For the criterion measures in the career advancement category, we assumed a mean validity of 1 and a variance of 0 because we could find no relevant data and we wished to use a conservative (i.e., high) figure. To obtain the reliability values for the total sample of studies, a distribution was created using the reliability estimates from each category in proportion to the number of studies we had in each category. Means and variances were then calculated from this distribution.

Moderator analyses. Moderator analyses were not undertaken until it was determined that enough variance in correlations remained after correcting for statistical artifacts to warrant such a search (Hunter et al., 1982). In all, 20 potential moderators were tested. Continuous and dichotomous moderators were tested by correlating them with study validities. Although other studies have tested dichotomous moderators by dividing validities into groups and comparing corrected means and variances, we chose to compute point-biserial correlations because this gave us statistics comparable to the other moderator analyses using correlations. We did however, test potential moderators with three or more categories by comparing corrected means and variances of each group of validities.

Several factors influenced our choice of variables to test as potential moderators. First, there had to be sufficient variance on the variable to allow a meaningful test. Second, the variable had to meet one of the following criteria: Past research suggested its moderating effects, the results would have relevance to concerns for equal employment opportunity, or the results might have practical relevance to the design or administration of assessment centers.

Table 3
Means and Standard Deviations for Variables Tested as Moderators of Assessment Center Validities

Variable	*N*	*M*	*SD*
Publication year	108	74.00	7.77
Mean age of assessees	57	30.15	6.89
Percentage men	68	63.98	43.42
Percentage minority status	37	15.69	28.83
Total no. of devices	104	7.33	1.99
Days of assessor training	67	7.52	5.27
No. of hours per assessee spent integrating information	53	1.62	.52
Quality of the study as measured by an overall judgment made by the authors of this article	108	2.14	1.00
Quality of the study as measured by summing ratings of threats to validity	105	1.68	1.52

Table 4
Descriptive Information for Variables Tested as Moderators of Assessment Center Validities

Variable	*N*	Variable	*N*
Publication form	109	Days of observation of assessees	96
Published	63	1	17
Journal	60	2	57
Book	2	3 or more	22
Thesis	1	Psychologists vs. managers as assessors	76
Unpublished	21	Psychologists	10
Used an intelligence test	106	Managers	66
Yes	78	Use of peer evaluations	93
No	28	Yes	44
Ratio of assessees to assessors	80	No	49
1:1	6	Feedback given to assessees	87
2:1	57	Yes	50
3:1	11	No	37
4 or more:1	6	Feedback given to immediate supervisor	77
		Yes	16
		No	61

Tables 3, 4, 5, 7, and 11 list the potential moderators. Additional comment is warranted on the last two variables, type of criterion and purpose of the assessment. It was felt that differences in criterion type and purpose of the assessment center implied conceptually distinct types of validity information. It was also suspected that the other 18 potential moderators might operate differently within the categories of these two variables. So, at one point in our analyses, validities were sorted on the basis of criterion type and assessment purpose and then were tested for moderators within each of these sortings.

Large-sample studies. The distribution of sample size in our studies was positively skewed due to the presence of three relatively large studies. Moses and his colleagues (Moses, 1972; Moses & Boehm, 1975; Ritchie & Moses, 1983) used samples of 5,943, 4,846, and 1,097 assessees, respectively. These samples are substantially larger than the next largest sample, which contained 471 assessees. Because there is a chance that weighted means and variances could be misleading when samples of this magnitude exist (Hunter et al., 1982, p. 41), our meta-analysis was carried out twice, once including the disparate studies and once excluding them. Within the total sample of studies, only a small, nonsignificant decrease was found in the corrected mean and variance when the large studies were removed. However, the three large studies were excluded from subsequent calculations in our meta-analysis because it was suspected that they would predominate in the subgroup analyses that contained fewer studies.

Results

Descriptive Information

Table 6 lists a number of the characteristics of the studies in our meta-analysis. There is wide diversity in the design of assessment centers and their predictive validity studies. Among the studies (28% of the total) that reported minority status, on the average 17% of the assessees were minorities. The total number of different types of assessment devices ranged from 1 to 11 with a mean of 7 per study. The number of days of obser-

Table 5
Ratings of Quality of Studies Tested as Moderators of Assessment Center Validities

Ratings on individual threats	No plausible threat	Minor problems	Plausible threat	Could explain most of results
Inadequate representation[a]	41	30	37	
Motivation differences[a]	96	4	7	
Job experience[a]	78	21	8	
Criterion contamination	50	25	25	5
Other	91	4	9	1

[a] Summed to form a composite rating of study quality.

vation ranged from one to three days. For most studies (64%), managers served as assessors; some employed both psychologists and managers (20%), whereas a few used only psychologists as assessors (10%).

Most studies were published in journals (52%), others were presented at conferences (22%) or were prepared as in-house technical reports (22%). The plurality of the assessment centers reviewed were conducted for promotional purposes (46%); however, others were carried out for the purpose of selection (22%), developmental planning (4%), early identification of managerial talent (16%), or basic research (6%). Most of the studies (52%) used a predictive validation design and provided feedback to assessees regarding their performance in the assessment center. Others used a predictive design but did not provide feedback (19%), were pure research experiments (16%), used a control group (4%), or used a concurrent validation design (20%). In addition, many studies either used job performance ratings ($n = 28$) or measures of career advancement ($n = 25$) as criteria, whereas others used ratings of potential ($n = 9$), measures of performance in training ($n = 7$), or dimension ratings ($n = 5$). The variety of assessment centers described here substantiates the contention that a typical assessment center does not exist.

The 50 studies reported 220 validities that when combined within criterion types within each study resulted in 112 validity coefficients. Three large-sample studies and two studies that failed to report sample size were excluded from our analyses, yielding a final total of 107 validity coefficients. Fifteen studies contributed approximately one half of these validities, but several were computed from independent samples of subjects. Supplementary analyses were conducted after additional combinations of potentially dependent validities.

Test for Validity Generalization

Cumulation of effect sizes across studies. Table 7 presents unweighted means and variances for the total set of validities and for validities sorted into categories of criteria type and purpose of the assessment center. The unweighted mean validity ($\bar{r}$) across the total sample was .32. Most of the mean validities within the criteria and purpose categories were close to this value. The two exceptions were the validities for studies using ratings of potential as the criterion ($\bar{r} = .45$), and for research studies ($\bar{r} = .42$). None of the mean validities changed significantly when they were recalculated, including the three outliers.

The last column of this table contains the raw or unweighted variances (s_r^2) across validities. The s_r^2s of most categories ranged from .03 to .04. However, validities using dimension ratings as the criterion had a relatively large variance (.071), whereas assessment centers conducted for the purposes of research, early identification, and career advancement had relatively little variance.

Correction for artifacts. Table 8 presents the weighted mean validities and variances corrected for statistical artifacts. According to Hunter et al. (1982), weighting by sample size increases the accuracy of population estimates. The relative magnitude of the weighted and unweighted means and variances are quite similar. However, in all cases, the weighted means and variances (Table 8) are slightly lower than the unweighted values (Table 7). For example, the unweighted mean and variance across the total set of validities are .32 and .030, respectively. Weighting by sample size reduces these numbers to .29 and .023. These reductions result from a negative correlation, $r = -.24$ ($p < .05$), between sample size and size of validities.

The column headed by $\bar{\rho}_{xU}$ presents the weighted means corrected for range restriction and unreliability in the criterion.[3] The corrected mean across the total sample of validities is .37. The corrected means within the criterion categories are about .35, with the exception of the average validity for studies using ratings of potential, which had a noticeably larger mean of .53. In the purpose categories, early identification, selection, and research studies have corrected mean validities between .41 and .48, whereas mean validity for promotion studies was somewhat smaller (.30).

These corrected means were computed by dividing the uncorrected weighted means by the product of the estimates of the average of the square root of the criterion reliabilities, $E(r_{yy}^{1/2})$, and values incorporating estimates of range restriction, $\bar{c}$. (See Hunter et al., 1982, p. 90.) Small c values indicate more range restriction than large c values.

A comparison of Column 1 with Column 3 reveals that the correction for unreliability and range restriction boosted some validities more than others. Within the criterion categories, performance, potential, and dimension means increased by at least .11. However, training and career progress mean validities increased by about only one half this amount. In studies in which training performance was the criterion, the smaller increase was due to the lack of range restriction (i.e., $\bar{c} = .977$). In studies

[3] Using Hunter, Schmidt, and Jackson's (1982) notation, "U" represents both range restriction ($\bar{U}$) and corrected criterion scores in the notation of the mean validity corrected for artifacts ($\bar{\rho}_{xU}$).

(*text continues on page 501*)

Table 6
Studies Included in the Meta-Analysis

Author(s)	Form of study[a]	% of minority	No. of devices	No. of days of assessor observation	Type assessor	Peer evaluation	Purpose of AC	Design of study	Criterion type	Validities used in the meta-analysis[b]
Alexander, Buck, & McCarthy (1975)	Journal	NR	1	NR	NR	No	Early ID	Concurrent	Job performance	.23
American Airlines (1976)	Unpublished	3	4	2	Managers	Yes	Promotion	Predictive (w/feedback)	Career	.28 (3)
								Concurrent	Potential	.32 (2)
Anstey (1966)	Journal	NR	8	2	Both	No	Selection	Predictive (w/feedback)	Career	.22
Anstey (1971)	Journal	NR	8	2	Both	No	Selection	Predictive (w/feedback)	Job performance	.30
									Career	.30
Anstey (1976)	Unpublished	NR	8	2	Both	No	Selection	Predictive (w/feedback)	Career	.35
Bentz (1980)	Presentation	NR	6	1	Both	No	Selection	Predictive (w/o feedback)	Job performance[c]	.38, .28, .27
									Career	.40 (2)
Borman (1982)	Journal	NR	6	1	NA	No	Promotion	Predictive (w/feedback)	Training	.34 (6)
Bray (1964)	Journal	NR	6	3	Managers	Yes	Promotion	Control group	Job performance	.26
									Potential	.31
Bray (1982)	Journal	0	7	3	Psychologists	Yes	Promotion	Experimental	Career	.24, .33
Bray & Campbell (1968)	Journal	NR	7	2	Managers	NR	Selection	Predictive (w/o feedback)	Job performance	.28 (2)
									Training	.04
Bray, Campbell, & Grant (1974)	Book	0	7	3	Psychologists	Yes	Research	Experimental	Career	.40, .42
Bray & Grant (1966)	Journal	0	7	3	Psychologists	Yes	Research	Experimental	Career	.44 (2), .50 (2)
Bullard (1969)	Unpublished	NR	7	2	Managers	No	Promotion	Predictive (w/feedback)	Job performance	.14[d] (4)
Campbell & Bray (1967)	Journal	NR	6	3	Managers	Yes	Promotion	Control group	Job performance	.13
									Potential	.22
Carleton (1970)	Presentation	NR	9	3	Both	No	Promotion	Predictive (w/feedback)	Potential	.65
									Dimension	.23 (12)
Dunne, Komar, Wise, & Norton (1981)	Presentation	NR	5	NR	Managers	No	Promotion	Predictive (w/feedback)	Job performance	.28
									Career	.22 (3)
Erpenbach (1968)[e]	Unpublished	NR	4	1	Managers	No	Promotion	Concurrent	Job performance	.74
Gardner & Williams (1973)	Journal	NR	6	2	Both	No	Selection	Predictive (w/feedback)	Job performance	.14
									Training	.21
									Career	.19 (2)
Grossner (1974)	Thesis	NR	8	3	Managers	Yes	Early ID	Predictive (w/feedback)	Job performance	.23 (2)
Haynes (1978)	Unpublished	NR	NR	NR	NR	NR	Promotion	Experimental	Job performance	.17
									Potential	.24
Hinrichs (1969)	Journal	NR	8	2	Managers	No	Early ID	Concurrent	Job performance	.46
									Career	.37

Table 6 (*continued*)

Author(s)	Form of study[a]	% of minority	No. of devices	No. of days of assessor observation	Type assessor	Peer evaluation	Purpose of AC	Design of study	Criterion type	Validities used in the meta-analysis[b]
Hinrichs (1978)	Journal	NR	8	2	Managers	No	Early ID	Predictive (w/o feedback)	Career	.46
Howard (1979)	Presentation	0	7	3	Psychologists	Yes	Research	Experimental	Career	.33, .44
Huck & Bray (1976)	Journal	0, 100, 28	8	2	Managers	Yes	NR	Predictive (w/feedback)	Job performance	.41, .35, .40, .44
								Predictive (w/o feedback)	Potential	.59, .54, .52, .64
									Dimension	.47[c] (6), .41[c] (6)
Klimoski & Strickland (1981)	Unpublished	NR	7	2	Managers	Yes	Promotion	Predictive (w/feedback)	Job performance Potential Career	.02 .37 .22 (3)
Kraut & Scott (1972)	Journal	NR	10	2	Both	NR	Early ID	Predictive (w/feedback)	Career	.28, .29, .33, .30, .26, .27
McConnell & Parker (1972)	Journal	33, 0	11	1	Managers	Yes	Promotion	Concurrent	Job performance	.55, .48, .64, .28
McElroy (1979)	Unpublished	6	NR	NR	Both	No	Promotion	Predictive (w/feedback)	Job performance	.51, .41
								Concurrent	Career	.52
Metropolitan Transit Authority (1972)	Unpublished	20	5	1	Managers	NR	Development Planning	Concurrent	Job performance	.71
Mitchel (1975)	Journal	NR	9	3	Managers	Yes	Promotion	Predictive (w/feedback)	Career	.16 (2), .19, .25 (3)
Moses (1972)[f]	Presentation	NR	6	2	Managers	NR	Promotion	Predictive (w/feedback)	Career	.44
Moses & Boehm (1975)[f]	Journal	NR	6	2	Managers	Yes	Promotion	Predictive (w/feedback)	Career	.25 (3)
Moses & Wall (1975)	Journal	77	4	1	Managers	NR	Selection	Predictive (w/o feedback)	Job performance	.60
Norton (1980)	Presentation	NR	4	NR	Managers	No	Early ID	Concurrent	Job performance	.30
Parker (1980)	Journal	6	11	1	Managers	Yes	Promotion	Concurrent	Job performance	.36
Ritchie (1980)	Presentation	20	4	1	Managers	NR	Promotion	Experimental	Job performance Training Career	.10 .14 .03
Ritchie & Moses[f] (1983)	Journal	NR	5	2	Managers	No	Development Planning	Predictive (w/feedback)	Career	.42
Schmitt, Noe, Meritt & Fitzgerald (1984)	Presentation	15	6	2	Managers	No	Promotion	Predictive (w/feedback)	Job performance Dimension	.25, .29, .09 −.03 (7), −.04 (7)
Slivinski, Grant, Bourgeois, & Pederson (1977)	Unpublished	NR	10	2	Managers	No	Promotion	Predictive (w/feedback)	Job performance Career	.38 (3), .14 (3) .42 (6), .29 (4)

Table 6 (*continued*)

Author(s)	Form of study[a]	% of minority	No. of devices	No. of days of assessor observation	Type assessor	Peer evaluation	Purpose of AC	Design of study	Criterion type	Validities used in the meta-analysis[b]
Slivinski, McCloskey, & Bourgeois (1979)	Presentation	NR	5	NR	NR	No	NR	Predictive (w/o feedback)	Career	.30
Thomson (1969)	Journal	NR	9	3	Managers Psychologists	Yes	Promotion	Predictive (w/feedback)	Potential Dimension	.64, .64 .57[c] (12), .61[c] (12)
Tziner (1982)	Unpublished	NR	5	2	Managers	Yes	Selection	Experimental	Training	.35
Tziner (1984)	Journal	NR	4	2	Managers	Yes	Selection	Predictive (w/feedback)	Job performance Training	.30 .60
Tziner & Dolan (1982)	Journal	NR	6	2	Managers	No	Selection	Experimental	Training	.38
Vernon (1950)	Journal	NR	8	2	Both	No	Selection	Predictive (w/feedback)	Job performance[b] Training	.25, .13, .22, .16 .42
Warriner (1981)	Presentation	NR	10	2	Managers	NR	Early ID	Predictive (w/feedback)	Career	.39
Wilson (1948)	Journal	NR	8	2	Both	No	Selection	Predictive (w/feedback)	Job performance	.50
Wissman & Rankin (1982)	Presentation	NR	5	NR	Managers	No	Promotion	Predictive (w/o feedback)	Job performance Dimension	<u>−.02, .0</u> .06 (14)
Wollowick & McNamara (1969)	Journal	NR	10	2	Managers	No	Early ID	Predictive (w/feedback)	Career	.37
Worbois (1975)	Journal	8	11	1	Managers	No	Promotion	Concurrent	Job performance[b]	.38 (5)

Note. NR = not reported by the study. AC = assessment center. Double entries for % minority, type assessor, peer evaluation, purpose, and design reflect an article reporting multiple independent investigations. Double and triple entries for criteria type and validities used, reflects either multiple investigations or our desire to categorize criterion type into one of five types.

[a] When a study appeared in a journal and in other forms (e.g., presentation), we recorded it only as a journal.

[b] The numbers in parentheses indicate number of validities combined in the average or composite. Values that are underlined were combined in the supplementary analyses of potentially dependent validities.

[c] Summed or averaged dimension ratings used as the criterion were recorded as a measure of job performance.

[d] Rather than testing an average, we combined the validities into a composite.

[e] J. J. Erpenbach (personal communication, March 11, 1971).

[f] These studies were excluded from the bulk of the analyses because of their relatively large sample sizes.

Table 7
Unweighted Means and Variances of Validities

Sample	No. of studies	No. of validities	Sample range	Total sample	Unweighted $\bar{r}$[a]	Unweighted s_r^2
Total	47	107	12–471	12,235	.32	.030
Criteria						
Performance	29	44	12–471	4,180	.31	.032
Potential	9	13	20–425	1,338	.45	.037
Dimension	5	9	35–122	748	.25	.071
Training	8	8	50–269	1,062	.31	.031
Career	22	33	30–437	4,907	.32	.011
Purpose[b]						
Promotion	21	52	13–53	5,201	.29	.040
Early ID	8	15	24–437	2,068	.31	.009
Selection	12	24	55–301	3,198	.30	.019
Research	3	6	125–144	837	.42	.003

Note. Unweighted $\bar{r}$ = simple mean validity (i.e., not weighted by sample size). Unweighted s_r^2 = simple variance of validities (i.e., not weighted by sample size).
[a] The mean correlation cefficients did not change when we recalculated them including the three large-sample studies.
[b] The total of the purpose categories fails to sum to 107 because we excluded two validities in a developmental planning category and were unable to classify several others.

with career progress measures, the smaller increase was due to the high reliability estimates for this criterion.

Similar observations can be made within the purpose categories. Mean validities of early identification and selection studies increased by at least twice the amount for promotion and research studies. The relatively smaller increases for promotion and research were due to lack of range restriction within these studies.

In column 4, $\sigma_{\rho_{xU}}^2$ represents the variances of the weighted validities corrected for all statistical artifacts. These values were computed using a formula presented by Hunter et al. (1982, p. 90). Schmitt, Gooding, Noe, and Kirsch (1984) noted that this correction may be inaccurate when applied to small samples. Thus, caution should be exercised when interpreting the corrected variances for the research and training studies, and studies using dimension ratings as criteria. In most categories, the correction reduced the original weighted variances. For career, early identification, selection, and research validities, most or all of the variance appears to have been artifactual. Partial support for this conclusion was found using a chi-square test developed by Hunter et al. (1982, p. 47), which was applied to uncorrected weighted variances. However, the results of the chi-square analyses should be interpreted with caution. Whereas nonsignificant results, as found for early identification, career progress, selection, and research validities, suggest that no true variance exists among these populations of validities, significant results are ambiguous and can be caused by artifacts, true variance, or both. In addition, even when significant results are based on true variances, the amounts may be trivial in size.

The last three columns provide information about the distributions of corrected validities. The lower 90% credibility value is the point above which lie 90% of the true validities. This statistic can be used to assess the likelihood that any given assessment center will be at least minimally valid. This value exceeds zero for all of the studies except those in which dimension ratings were used as criteria. The final two columns depict the lower and upper bounds for the 95% confidence interval created around the corrected mean. Using this more stringent criterion, all categories of studies except those using dimension ratings and those conducted for promotional purposes, appear to be at least minimally valid.

Supplementary analyses were conducted to examine whether potentially nonindependent validities within studies affected the results of this meta-analysis. Each study in Table 6 that contained multiple validities was reexamined. A judgment was made by the third author about whether the separate coefficients were potentially dependent. Validities were considered potentially dependent and, thus, were combined (a) if it appeared that the same or quite similar assessors were involved, (b) if the study was done in the same small organization or division, (c) if the assessments were done at about the same time, or (d) if the criterion measures came from the same types of raters. Validities for different criteria or other variables under study were not combined. Key results for the original total sample and the supplementary, combined total sample are shown in Table 9. Differences are quite small, and we decided to proceed with analyses on the 107 validities.

Table 10 presents the variance of the weighted validities decomposed into their artifactual and nonartifactual components. Values of σ_e^2 depict the amount of variance due to sampling error (see Hunter et al., 1982, p. 44). Values of $\sigma^2\sqrt{r_{yy}} + \bar{U}$ represent the amounts of variance due to the combination of unreliability of the criterion and range restriction. These quantities are calculated using terms and their operations to the right of the minus sign in the numerator of the formula for $\sigma_{\rho_{xU}}^2$ (see Hunter et al., 1982, p. 90). Values of σ_1^2 represent the amount of variance remaining when σ_e^2 and $\sigma^2\sqrt{r_{yy}} + \bar{U}$ are removed. The final column contains the percentage of variance in the original validities that is not explained by statistical artifacts.

In the total sample and in five of the subgroups of validities, more than 40% of the variance in correlations could not be explained by artifacts. However, all of the variance for career progress, early identification, and research validities appears to be artifactual.

Table 8
Weighted Means and Variances Corrected for Artifacts

Sample	Weighted $\bar{r}$	Weighted s_r^2	$\bar{\rho}_{xU}$	$\sigma_{\rho_{xU}}^2$	$\bar{c}$	σ_c^2	$E(r_{yy}^{1/2})$	$\sigma^2\sqrt{r_{yy}}$	Lower 90% credibility value	95% confidence interval	
										Lower bound	Upper bound
Total	.29	.0228	.37	.0171	.896	.032	.871	.015	.21	.11	.63
Criteria											
Performance	.25	.0233	.36	.0203	.902	.031	.774	.016	.18	.08	.64
Potential	.40	.0330	.53	.0373	.974	.004	.774	.016	.28	.15	.91
Dimension	.22	.0606	.33	.0998	.883	.028	.774	.016	−.07	−.29	.95
Training	.30	.0219	.35	.0197	.977	.004	.894	.002	.17	.07	.63
Career	.30	.0087	.36	.0000	.837	.051	1.000	.000	.36	.36	.36
Purpose											
Promotion	.24	.0304	.30	.0293	.939	.011	.871	.015	.08	−.04	.64
Early ID	.30	.0032	.46	.0000	.746	.056	.871	.015	.46	.46	.46
Selection	.29	.0166	.41	.0032	.805	.059	.871	.015	.34	.30	.52
Research	.42	.0027	.48	.0000	1.000	.000	.871	.015	.48	.48	.48

Note. Weighted $\bar{r}$ = mean validity weighted by sample size; weighted s_r^2 = variance of validities weighted by sample size; $\bar{\rho}_{xU}$ = mean validity corrected for statistical artifacts; $\sigma_{\rho_{xU}}^2$ = variance corrected for statistical artifacts; $\bar{c}$ = a measure of range restriction (1 = none, 0 = severe); σ_c^2 = variance of c; $E(r_{yy}^{1/2})$ = average of square roots of reliabilities across criterion measures; $\sigma^2\sqrt{r_{yy}}$ = variance of $\sqrt{r_{yy}}$.
[a] When outliers were included, the total mean was .33, the career mean .34, and the promotion mean .32; all other means remained unchanged.
[b] The chi-square test of variance was significant ($p < .05$) for all but the career, early ID, selection, and research categories.

Moderator Analyses

Tables 11 and 12 contain the results of the moderator analyses. Table 11 presents the analyses for the two variables that had more than two categories: study design and publication form. The corrected means for all categories of both variables are similar. The corrected mean validities for different study designs ranged from .36 for experimental studies to .43 for predictive studies without feedback. The corrected mean validities for studies published in different forms ranged from .33 for presentations to .39 for unpublished technical reports. Thus, it appears that neither design of the study nor publication form moderate assessment center validity.

In one of our analyses we subdivided our total sample of validities into subgroups of validities based on both criterion type and assessment purpose and *then* attempted to assess differential moderating effects for study design and publication form within each of these subdivisions. We do not report these analyses here because in some subcategories there were too few validities (i.e., less than five) to ensure stability of the results. In other subcategories, we judged that insufficient true variance remained for a particular criterion or purpose category to permit the operation of moderators. The latter judgment was made on the basis of the absolute amount of true variance found in the category, the results of the chi-square test on the uncorrected variances, and the percentage of the variance in the original correlations that could be explained by statistical artifacts. In a few subcategories, in which sufficient numbers of validities and variance existed, we found no support for differential moderating effects of study design or publication form within studies of different criteria and purpose. (These results may be obtained from the second author.)

Table 9
Meta-Analytic Results on Total Sample: Before and After Combining Potentially Nonindependent Validities Within Studies

Sample	No. of validities	Weighted $\bar{r}$	Weighted s_r^2	$\bar{\rho}_{xU}$	$\sigma_{\rho_{xU}}^2$
Before combining	107	.2913	.02281	.3732	.01711
After combining	89	.2854	.02425	.3600	.01987

Table 12 contains the results for analyses of potential moderators that are continuous and dichotomous variables. These variables were tested within the total sample of validities, within the job performance, potential, and dimension criterion categories, and within studies done for promotion purposes. Studies using career advancement criteria and studies conducted for selection, early identification, and research purposes were not analyzed because we judged that insufficient true variance existed (see Table 10).

The first row of entries for each potential moderator are Pearson product–moment and point-biseral correlations between the moderator variable and the effect size. The second row contains these correlations corrected for sampling error (Hunter et al., 1982, p. 52). The entries in parentheses are the number of validities used in the calculations. Due to the likelihood of capitalizing on chance with small samples, we excluded those correlations from the table that were based on fewer than nine validities. In sum, 69 correlations were computed and 25 were found significant ($p < .05$). The probability of this occurring by chance is extremely small ($CR = 14.77$, $p < .001$; Brozek & Tiede, 1952).

A few variables demonstrated significant correlations across samples of validities. The results suggest that assessment center

Table 10
Percentage of Weighted Variance Unexplained by Artifacts

Sample	Weighted s_r^2	σ_e^2	$\sigma^2\sqrt{r_{yy}} + \overline{U}^2$	σ_1^2	% variance unexplained by artifacts
Total	.0228	.0073	.0051	.0104	46
Criteria					
Performance	.0233	.0092	.0041	.0100	43
Potential	.0330	.0069	.0049	.0212	64
Dimension	.0606	.0109	.0031	.0466	77
Training	.0219	.0062	.0006	.0151	69
Career	.0087	.0055	.0067	.0000	0
Purpose					
Promotion	.0304	.0089	.0019	.0196	65
Early ID	.0032	.0060	.0107	.0000	0
Selection	.0166	.0058	.0093	.0015	9
Research	.0027	.0049	.0035	.0000	0

Note. Weighted s_r^2 = variance of validities weighted by sample size. σ_e^2 = variance due to sampling error; $\sigma^2\sqrt{r_{yy}} + \overline{U}^2$ = variance due to range restriction and unreliability on the criterion; σ_1^2 = variance left over after removing artifacts.

validities are higher when the percentage of male assessees is low, when a larger number of assessment devices are used, when assessors are psychologists rather than managers, when peer evaluations are used, and when the studies are judged to be of higher quality.

Other variables, however, operated as moderators only within a group of studies that were conducted for a single purpose or that used a particular criterion. For example, within the group of studies done for promotion purposes, validities are higher when the percentage of minority assessees is low. When predicting job performance, lower validities were found when assessors spend more days observing assessees. In addition, when ratings of potential are the criterion, validities are higher when feedback is given to assessees than when it is not.

Discussion

Generalizability of Assessment Centers

The findings of this meta-analysis support the widely held contention that assessment centers have predictive validity (Byham, 1970; Cohen et al., 1977; Howard, 1974; Huck, 1977; Hunter & Hunter, 1984; Thornton & Byham, 1982). Assessment center validities, corrected for sampling error, restriction of range, and criterion unreliability yielded a mean validity coefficient of .37. The average corrected validity coefficients for the various purposes of assessment centers ranged from .30 in promotional studies to .48 in basic research studies. Mean corrected validity coefficients for the prediction of different criteria ranged from .33 for dimensional ratings to .53 for ratings of management potential. Given the lower bound of the 90% credibility value for the average corrected validity coefficient in total sample, .21, we conclude that the validity of assessment centers does generalize.

The results of this meta-analysis must be interpreted with caution because of the complex and variable nature of assessment centers. Reliance on these results assumes that a new assessment center application will be designed and administered as well as or better than the average assessment center reviewed in this study. The *Standards and Ethical Considerations for Assessment Center Operations* (Task Force on Assessment Center

Table 11
Corrected Means and Variances for Study Design and Publication Form Calculated Across the Total Sample of Validities

Sample	No. of studies	No. of validities	Weighted $\bar{r}$	Weighted s_r^2	$\bar{\rho}_{xU}$	$\sigma_{\rho_{xU}}^2$	$\bar{c}$	σ_c^2
Study design								
Experiment	7	15	.32	.0189	.36	.0161	1.000	.000
Predictive (w/o feedback)	7	14	.30	.0311	.43	.0107	.809	.052
Predictive (w/feedback)	23	59	.29	.0234	.39	.0186	.855	.039
Concurrent	10	15	.36	.0184	.42	.0035	1.000	.000
Publication form								
Journal	25	58	.30	.0188	.38	.0110	.916	.030
Unpublished	10	21	.32	.0267	.39	.0194	.927	.021
Presentation	10	25	.23	.0272	.33	.0303	.812	.041

Note. Weighted $\bar{r}$ = mean validity weighted by sample size; weighted s_r^2 = variance of validities weighted by sample size; $\bar{\rho}_{xU}$ = mean validity corrected for artifacts; $\sigma_{\rho_{xU}}^2$ = variance corrected for statistical artifacts; $\bar{c}$ = a measure of range restriction (1 = none, 0 = severe); σ_c^2 = variance of c across validities.

Standards, 1980) provides guidance on the essential features of an assessment center.

Note that the present corrected validity coefficients differ from those calculated by Hunter and Hunter (1984). Hunter and Hunter found median corrected correlations of .63 for potential and .43 for performance, compared to the present mean correlations of .53 and .36, respectively. Considering that we corrected for sampling error, range restriction, and differences in unreliability in the criterion, whereas Hunter and Hunter (1984) corrected only for the first artifact, one would expect our values to be higher. Our lower values may be due to two factors: (a) We included a wider selection of studies, both published and unpublished, and additional studies conducted in the last 11 years, and (b) more recent studies tend to have lower validity as indicated by the slight negative correlation between publication year and assessment center validities. Taken together, the two meta-analyses suggest that assessment centers show validity generalization.

It should be recognized that the validity coefficients used for this meta-analysis may reflect a subtle form of *criterion contamination* not ferreted out in our moderator analyses of study design, study quality, and type of criterion (all of which are discussed later). We are referring to a set of perceptions about the qualities of a good manager that may be shared by the assessors (usually managers themselves) and anyone who provides criterion data later (e.g., performance ratings or promotion decisions). What we call a *validity coefficient* may be partially determined by a prototype (Feldman, 1981) of "a good manager" held in common among the various people providing both predictor and criterion data. This hypothesis deserves further investigation.

Situational Specificity of Assessment Centers

Our results also provide support for the situational specificity of assessment centers. Whereas recent validity generalization studies have shown that sampling error, unreliability of predictors and criteria, and range restriction account for about 75% of the observed variance across test validation studies (Hunter, 1980; Lilienthal & Pearlman, 1983; Pearlman, 1984; Pearlman et al., 1980; Schmidt et al., 1980; Schmidt & Hunter, 1977; Schmidt et al., 1981; Schmidt, Hunter, Pearlman, & Shane, 1979), these statistical artifacts accounted for only 54% of the observed variance of the total sample in the present study. Therefore, almost one half of the variance remains unexplained. Percentages of variance unexplained for studies conducted for certain purposes and involving some criteria were even higher. These results may be explained by the fact that the assessment center is a general method characterized by different procedures.

In addition to the percentage of remaining observed variance, it is important to note that the absolute level of variance is substantial. For the total sample, the standard deviation of true validities is .13, and for studies using dimension ratings it is .32. Even if artifacts for which we did not correct could account for one half of this remaining variance, there would still be enough variance remaining to conclude that assessment centers do show different true levels of validity. This finding is consistent with the conclusions of Schmidt and Hunter and their colleagues that validity generalization is frequently possible even when the situational specificity hypothesis cannot be rejected (Pearlman et al., 1980, 1981; Schmidt et al., 1980). Support for both validity generalization and situational specificity has been found for weighted application blanks (Brown, 1981), intelligence and arithmetic tests (Schmidt et al., 1981), and the Law School Aptitude Test (Linn, Harnisch, & Dunbar, 1981). Given the utility work of Brown (1981), which suggests that even small *real* differences in validity coefficients across situations can have practical monetary implications, we suggest that designers of assessment centers take into consideration the variables found to moderate validities in this study.

Moderators of Assessment Center Validity

We looked for moderators within several coding categories: assessee characteristics, evaluation device characteristics, other assessment center characteristics, and validation study characteristics. These findings must be interpreted with caution because of the small sample sizes in some analyses.

Assessee characteristics. Assessee age, sex, and minority status were analyzed as potential moderators. There was no relation between average age of assessees and predictive validity of the assessment center. However, results suggest that assessment centers are more valid when the composition of the group consists of a larger proportion of women and a smaller proportion of minorities. Two explanations are possible: (a) Assessment centers may be more valid for women and for minorities, or (b) group composition alters the dynamics of the assessment process such that the overall assessment rating is more accurate when the assessee group includes a large portion of women, and less accurate when the assessee group includes a large portion of minorities.

The first explanation is not supported by previous studies of assessment centers, which have found no differential validity for Blacks and Whites (Huck, 1974; Huck & Bray, 1976), or for men and women (S. Alexander, 1975; Clingenpeel, 1979; Hall, 1976; Marquardt, 1976; Moses, 1973b; Moses & Boehm, 1975). Because women may be more self-disclosing than men (see, e.g., Fletcher, 1981; Fletcher & Spencer, 1984), they may provide assessors with more or better information to help make assessment ratings.

It is also possible to rule out other explanations for the present finding that sex and predictive validity of the assessment center are related, by examining correlations of sex and other moderators. For example, percentage of women is inversely related to the use of peer evaluations and the use of psychologists as assessors. Because studies that use peer evaluations and psychologists as assessors have higher validities, we can have greater confidence that sex is itself a moderator.

Evidence related to the second explanation, which is that group composition affects the dynamics of assessment, comes from a study by Schmitt and Hill (1977), who found that peer and assessor average ratings were minimally influenced by the proportion of men and women, or Blacks and Whites, in the group. The ratings of Black women on some dimensions were somewhat lower when the group consisted of a large portion of White men. Further study is needed to determine whether group composition or differential validity explains the findings of this meta-analysis.

It may initially appear unsettling that for promotional stud-

ies, the greater the proportion of minority assessees, the lower the validity of the assessment center. However, in the spirit of affirmative action and in response to pressures from compliance agencies, organizations may be promoting greater numbers of minority candidates even though they have received relatively low assessment ratings.

Evaluation device characteristics. The analyses suggest that assessment centers are more predictively valid when a greater number of different types of exercises are used. One might discount this finding if the variable is positively correlated with other moderators. In fact, we found that the number of types of exercises is negatively correlated with the use of peer evaluations, which also moderates assessment center validities, and thus we can be somewhat more confident that it actually moderates validities. This supports the advice (Thornton & Byham, 1982) that a broad spectrum of types of exercises should be used to attain content representativeness in assessment centers.

Although less than one half of the predictive validity studies reviewed used some form of peer evaluation to help evaluate assessees, assessment centers that did were found to be more valid than those that did not. This finding is not surprising given the substantial amount of evidence that shows that peer assessment can be both a reliable and valid predictor of performance (Kane & Lawler, 1978). Although organizations are reluctant to use peer ratings for fear of increasing competitiveness among assessees, they should be used more often in the future to supplement the ratings of trained assessors if such reactions could be minimized.

Other assessment center characteristics. We also analyzed a number of other assessment center characteristics to see whether they moderated assessment center validity. These were, type of assessor used (i.e., manager or psychologist), amount of training assessors were given, the number of days assessors spent observing assessees and the number of hours they spent integrating information, the ratio of the number of assessees to assessors, and whether feedback was given to assessees or to their immediate supervisors.

Type of assessor was the only variable among these assessment center characteristics that moderated validities in the total sample. In contrast to other researchers (Greenwood & McNamara, 1969; Thomson, 1970) who have found no difference in the assessment center ratings of professional (i.e., psychologists) and nonprofessional (i.e., in-house managers) assessors, we found evidence that assessment centers that use psychologists as assessors are significantly more valid than those that use managers as assessors. Many people in the field believe that managers are better able to interpret the meaning of different behaviors for a particular job than are psychologists, because they are more familiar with the requirements of the job. However, the results of this meta-analysis suggest that psychologists provide more valid assessment center ratings than do managers. In fact, this moderator is particularly robust given that it is negatively related to many other moderators.

The following variables, all thought to be related to assessment center validity, were not significantly correlated with validities in this meta-analysis: ratio of assessees to assessors, number of days of observation and days of assessor training, hours spent integrating information, feedback to assessees and their supervisors, and criterion contamination. Providing feedback to assessees and their supervisors seems to have little effect on assessment center validities. (Only when potential ratings are the criterion does feedback to assessees seem to inflate validities.) These findings suggest that criterion contamination is not the sole explanation for the high correlation of assessment center and follow-up ratings.

One finding that initially surprised us was that amount of assessor training did not affect the validity of assessment centers. Thorough assessor training is thought to be essential to producing reliable and valid ratings (Task Force on Assessment Center Standards, 1980). However, given the mixed success of assessor training for related skills (see Landy & Farr, 1980), the lack of relation among assessor training and assessment center validity found in this meta-analysis is not very surprising.

The present results should be interpreted with caution, however, because research reports do not always give adequate descriptions of the amount or type of assessor training. Although there were no reports of research that did *not* train assessors, a number of researchers failed to mention whether assessors were trained and if they were, for how long. Therefore, we were unable to discern whether there is a significant difference in the validity of ratings of assessors who have been trained compared to those who have not. However, we can conclude that within the range of number of days of training studied (.5–15), more training does not lead to high validities.

Validation study characteristics. The results of this meta-analysis support those who maintain that assessment centers are more valid for predicting an assessee's job potential ($\bar{\rho}$ = .53) than for predicting performance ($\bar{\rho}$ = .36).

These analyses are quite comparable to the analyses conducted by Cohen et al. (1977). Cohen and his colleagues concluded that predictive accuracy was highest for job potential (*Mdn r* = .63), followed by progress (*Mdn r* = .40) and job performance (*Mdn r* = .33). In addition, subsequent individual studies by Klimoski and Strickland (1981) and Turnage and Muchinsky (1984) found that assessment centers predict progress but not performance criteria.

As Klimoski and Strickland (1977) pointed out, the superior ability of assessment centers to predict potential over performance may be due to the assessment staff's intuitive grasp of organizational values and norms with regard to promotion, and to their adeptness at predicting who will get promoted in the organization. Predicting an assessee's subsequent job performance, given the variety of factors outside the assessee's immediate control (e.g., dependency on other workers, customers, raw materials) and the notorious bias of supervisory ratings is a much more challenging task.

Another validation study characteristic we investigated was study design (i.e., whether the validity study was a predictive study with or without feedback, a concurrent validation, or a pure experiment). Our results show that study design does not moderate assessment center validities, a finding that supports research on cognitive tests (Bemis, 1968; Pearlman et al., 1980). This finding, along with the lack of significant correlation between validities and potential for criterion contamination through knowledge of the assessment results, contradicts a popular belief that operational use of assessment center data inflates validity coefficients as a result of contamination via knowledge of the predictor data. If contamination was a serious problem, validities for studies that operationally used assessment center data would be much higher. Our meta-analysis found no sig-

Table 12
Moderators for the Total Sample and Selected Criterion and Purpose Categories

Moderator	Total sample	Criteria type: Job performance	Criteria type: Potential ratings	Criteria type: Dimension ratings	Purpose of assessment center: Promotion
Publication year					
r	−.13	−.08	.11	−.87*	−.52*
rho[a]	−.16	−.10	.13	−.96	−.62
No. of validities	107	44	13	9	51
Mean age of assessees					
r	.06	.15	−.51	—	−.19
rho[a]	.07	.26	−.58	—	−.23
No. of validities	57	19	10	—	23
Percentage men					
r	−.43*	−.55*	−.91*	—	−.26
rho[a]	−.52	−.71	−1.00	—	−.31
No. of validities	68	28	9	—	31
Percentage minority					
r	.03	.01	—	—	−.74*
rho[a]	.03	.02	—	—	−.88
No. of validities	37	15	—	—	19
Used a general mental ability test					
r	−.11	.17	−.24	−.91*	−.06
rho[a]	−.14	.22	−.26	−1.00	−.07
No. of validities	106	43	12	9	49
Total number of devices					
r	.25*	.19	.63*	.86*	.48*
rho[a]	.31	.25	.71	.95	.57
No. of validities	104	42	12	9	47
Ratio of assessors to assessees					
r	−.12	−.14	−.26	—	−.17
rho[a]	−.15	−.18	−.29	—	−.20
No. of validities	80	33	11	—	43
Days of observation					
r	−.02	−.50*	.14	—	.03
rho[a]	−.02	−.64	.16	—	.03
No. of validities	96	37	12	—	42
Days of assessor training					
r	.08	.00	−.35	.17	−.22
rho[a]	.10	.01	−.39	.18	−.26
No. of validities	67	28	10	9	42
No. of hours spent integrating information					
r	−.01	−.33	−.22	—	−.19
rho[a]	−.02	−.42	−.25	—	−.23
No. of validities	53	19	11	—	36
Psychologist vs. managers as assessors[b]					
r	−.21*	NV	−.34	—	−.29*
rho[a]	−.26	NV	−.39	—	−.34
No. of validities	76	31	11	—	44
Peer evaluation[c]					
r	.36*	.20	.62*	.91*	.28*
rho[a]	.44	.26	.70	1.00	.33
No. of validities	93	40	12	9	47
Feedback given to assessees[d]					
r	.10	.07	.62*	—	.18
rho[a]	.12	.09	.70	—	.21
No. of validities	87	31	12	—	39
Feedback given to immediate supervisor[d]					
r	−.14	−.17	−.15	—	−.03
rho[a]	−.17	−.22	−.58	—	−.03
No. of validities	77	25	12	—	29

Table 12 (*continued*)

Moderator	Total sample	Criteria type: Job performance	Criteria type: Potential ratings	Criteria type: Dimension ratings	Purpose of assessment center: Promotion
Criterion contamination					
r	−.07	−.24	−.18	−.17	−.17
rho[a]	−.08	−.32	−.20	−.18	−.20
No. of validities	105	43	12	9	49
Quality of study (summed rating)					
r	.15†	−.18	.66*	.90*	.14
rho[a]	.18	−.23	.74	.99	.17
No. of validities	105	43	12	9	49
Quality of study (overall rating)					
r	.26*	.23	.21	.91*	.33*
rho[a]	.31	.29	.24	1.00	.39
No. of validities	107	44	13	9	51

Note. Significance tests were not performed on the rhos because no such test exists. NV = no variance in the moderator variable. All assessors were managers in the studies in which performance was the criterion. Descriptive data in each category can be obtained from the second author.
[a] Corrected for sampling error. [b] Psychologists coded 1; managers coded 2. [c] Absence of peer evaluation coded 1; presence of peer evaluation coded 2. [d] Feedback not given coded 1; feedback given coded 2.
* $p < .05$. † $p < .06$.

nificant differences between studies that operationally used assessment center data and those that did not. In combination, the findings refute the contention that direct contamination explains the observed validities of assessment centers.

A major caveat pervading our own analyses has been the internal and external validity of the studies. We found that the degree to which validation studies are internally and externally valid is related to their predictive validity. Our rating of the quality of the study, based on the representativeness of the sample, and motivational, job experience, and training differences between assessees and present employees was highly correlated with the validity of the assessment center. This finding supports Thornton and Byham's (1982) observation that methodologically sound studies have higher validity.

One somewhat unexpected finding is the lack of a significant relation between assessment center validities and the time at which criterion measures are taken. Much of the prior research found overall assessment ratings to be more predictive over a longer period of time (Hinrichs, 1978; Mitchel, 1975; Moses, 1972). Other researchers have found no relation between validities and time of criterion measure (Finley, 1970) or have found support for a negative relation (Howard, 1979; Slivinski & Bourgeois, 1977). Clearly this issue deserves further research.

In conclusion, we recommend that assessment centers be designed to use the features that are associated with the more highly valid programs reviewed in this meta-analysis. A well-designed assessment center will probably have predictive validity, but to optimize validity, certain procedures should be followed. Our results suggest that in the future, assessment centers should include more assessment devices, use psychologists as assessors, and supplement assessor ratings with those provided by peers. Within the range of variables reviewed in this meta-analysis, it does not appear that there is a systematic relation between the size of assessment center validities and the length of assessor training, time lapsed between the assessment center and when criterion measures are taken, the number of hours assessors spend integrating information, the number of days assessees are observed, or whether assessment center data is operationally used, if precautions are taken. We also recommend that validation studies of assessment centers be conducted with adequate research methodology (e.g., ensuring adequate sample representativeness).

The Art and Science of Meta-Analysis

After completing this meta-analysis, we believe that conducting a validity generalization study is somewhat of an art. Judgments are required at many junctures. Some of the issues we found most difficult to resolve are discussed ahead.

Moderator analyses. We studied moderator variables in a number of ways. First, studies were presorted into groups when the variables were categorical and there were a priori reasons for doing so. Then meta-analyses were performed in each group. This was the approach taken with type of criterion and assessment purpose. Both variables have been handled this way in prior meta-analytic work. This approach seems appropriate when the categories such as criterion types are theoretically and logically distinct from one another. However, type of criterion can also be viewed as just one of many assessment center design variables that vary from study to study and therefore should be treated as any other potential moderator. Hence, potential moderators should be tested only within the total sample of studies. We decided to test for moderators both within the total sample and within subgroups of studies using the same type of criterion.

In theory, our research allows us to compare the results of searching for moderators within the total sample and within presorted categories. Unfortunately, we were unable to completely carry out this comparison. Although we began our analyses with 107 validity coefficients, after presorting studies by

criterion and purpose, we quickly reached the point at which there were not enough studies in some categories to make meaningful comparisons on some variables. This may be more of a problem for assessment centers than other selection devices because of their complex design. We encourage others to study this issue further, using a selection device for which a larger number of studies exist.

A second method used for analyzing moderators was to compute point-biserial correlations between dicotomous moderators and validity coefficients. If a strong relation between a variable and validity is not found, it is highly unlikely that mean correlations will vary from subset to subset. The advantage of this approach is that it requires fewer calculations than performing meta-analyses within each subset of studies. Thus, point-biserial correlations can potentially be a valuable time-saving strategy. For continuous variables, we correlated the moderator with the validities using the Pearson product–moment method.

A problem encountered when interpreting the results of the moderator analyses was the interdependency of many of the variables. One way of handling correlated moderators is to partial out the effects due to one moderator and then look at the correlation between another variable and the mean validity. Unfortunately, sample sizes were too small to enable us to meaningfully do this. Instead, we were forced to speculate about how the interdependency among variables affects the strength of the moderators individually.

Large-sample studies. The validity generalization literature is not clear about how to decide whether to exclude studies with unusually large samples. We chose to exclude three studies because we did not want them to have undue influence on the mean and variance estimates. Yet, it can be argued that large studies have little sampling error; therefore, their influence is legitimate. In the present case, comparisons showed that the results are the same regardless of whether those outliers are included or not!

Insufficient reporting. Finally, we echo Schmidt et al.'s (1980) and Orwin and Cordray's (1985) contention that reports of validity studies must be more complete for validity generalization and meta-analysis research to be maximally effective. Orwin and Cordray found that deficient reporting injects considerable noise into meta-analysis data that can lead to spurious conclusions. Although we heeded their recommendations for countering the effects of deficient reporting (i.e., computed separate reliability coefficients for individual coding items, based on appropriate estimators; incorporated data quality information into our analysis; obtained additional information on primary studies by contacting original investigators), we were still unable to code all of the study features from most reports. In fact, a few studies were totally eliminated from the analyses because they failed to report enough essential information. In particular, it was difficult to evaluate the amount of range restriction present in assessment center validity studies and to estimate the reliability of criteria. We had to obtain surrogate data from other related areas of literature to construct some distributions. For example, we also used reliabilities from the performance evaluation literature to help construct reliability distributions of assessment center criterion ratings.

Deficits in reporting reinforce our notion that validity generalization is still somewhat of an art. The general procedure is well laid out, but there are numerous stages in the analysis in which judgment comes into play. How to combine most meaningfully across effect sizes, deal with the problems of unusually large sample sizes and correlated moderators, and obtain surrogate estimates to construct distributions, are but a few.

References

Alexander, L. D. (1979). An exploratory study of the utilization of assessment center results. *Academy of Management Journal, 22,* 152–157.

Alexander, S. J. (1975). Bendix Corporation establishes early identification program. *Assessment and Development, 2,* 10.

Bemis, S. E. (1968) Occupational validity of the General Aptitude Test Battery. *Journal of Applied Psychology, 52,* 240–244.

Bender, J. M. (1973). What is "typical" of assessment centers? *Personnel, 50,* 50–57.

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

Brown, S. H. (1981). Validity generalization in the life insurance industry. *Journal of Applied Psychology, 66,* 664–670.

Brozek, J. & Tiede, K. (1952). Reliable and questionable significance in a series of statistical tests. *Psychological Bulletin, 49,* 339–341.

Burroughs, W. A., Rollins, J. B., & Hopkins, J. J. (1983). The effects of age, departmental experience, and prior rater experience on performance in assessment center exercises. *Academy of Management Journal, 16,* 335–339.

Byham, W. C. (1970). Assessment center for spotting future managers. *Harvard Business Review, 48,* 150–160.

Byham, W. C. (1978a). How to improve the validity of an assessment center. *Training and Development Journal, 32,* 4–6.

Byham, W. C. (1978b, July). *Intercultural adaptability of the assessment center method.* Paper presented at Nineteenth International Congress of Applied Psychology, Munich, FRG.

Byham, W. C. (1981). *Dimensions of managerial success.* Pittsburgh, PA: Development Dimensions International.

Clingenpeel, R. (1979, June). *Validity and dynamics of a foreman selection process.* Paper presented at the meeting of the Seventh International Congress on the Assessment Center Method, New Orleans.

Cohen, B. M., Moses, J. L., & Byham, W. C. (1977). *The validity of assessment centers: A literature review* (Rev. ed.; Monograph No. 2). Pittsburgh, PA: Development Dimensions Press.

Cook, T. D., & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 223–326). Chicago: Rand McNally.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87,* 442–449.

Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66,* 127–148.

Finley, R. M., Jr. (1970, September). *An evaluation of behavior predictions from projective tests given in a management assessment center.* Paper presented at the 78th Annual Convention of the American Psychological Association, Miami Beach.

Fletcher, C. (1981). The influence of candidates' beliefs and self-presentation strategies in selection interviews. *Personnel Review, 10,* 14–17.

Fletcher, C., & Spencer, A. (1984). Sex of candidate and sex of interviewer as determinants of self-presentation orientation in interviews: An experimental study. *International Review of Applied Psychology, 33,* 305–313.

Greenwood, J. M., & McNamara, W. J. (1969). Leadership styles of structure and consideration and managerial effectiveness. *Personnel Psychology, 22,* 141–152.

Hall, H. L. (1976). *An evaluation of the upward mobility assessment center for the Bureau of Engraving and Printing* (TM No. 76-6). Washington, DC: U.S. Civil Service Commission.

Hinrichs, J. R. (1978). An eight-year follow-up of a management assessment center. *Journal of Applied Psychology, 63,* 596–601.

Howard, A. (1974). An assessment of assessment centers. *Academy of Management Journal, 17,* 115–134.

Howard, A. (1979, June). *Assessment center predictions sixteen years later.* Paper presented at the meeting of the Seventh International Congress on the Assessment Center Method, New Orleans.

Huck, J. R. (1973). Assessment centers: A review of the external and internal validities. *Personnel Psychology, 26,* 191–212.

Huck, J. R. (1974). *Determinants of assessment center ratings for White and Black females and the relationship of these dimensions to subsequent performance effectiveness.* Unpublished doctoral dissertation, Wayne State University.

Huck, J. R. (1977). The research base. In J. L. Moses & W. C. Byham (Eds.), *Applying the assessment center method* (pp. 261–291). New York: Pergamon Press.

Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of Black and White females. *Personnel Psychology, 29,* 13–30.

Hunter, J. E. (1980). *Test validation for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB).* Unpublished manuscript, Michigan State University.

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

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Advanced meta-analysis: Quantitative methods for cumulating research findings across studies.* Beverly Hills, CA: Sage.

Jaffee, C. L., Cohen, S. L., & Cherry, R. (1972). Supervisory selection program for disadvantaged or minority employees. *Training and Development Journal, 26,* 22–28.

Kane, J. S., & Lawler, E. E., III. (1978). Methods of peer assessment. *Psychological Bulletin, 85,* 555–586.

Klimoski, R. J., & Strickland, W. J. (1977). Assessment centers: Valid or merely prescient. *Personnel Psychology, 30,* 353–363.

Klimoski, R. J., & Strickland, W. J. (1981). *A comparative view of assessment centers.* Unpublished manuscript.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72–107.

Lilienthal, R. A., & Pearlman, K. (1983). *The validity of Federal selection tests for aide technicians in the health, science, and engineering fields* (OPRD Report No. 83-1). Washington, DC: U.S. Office of Personnel Management, Office of Personnel Research and Development. (NTIS No. PB83-202051)

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Validity generalization and situational specificity: An analysis of the prediction of first year grades in law school. *Applied Psychological Measurement, 5,* 281–289.

Marquardt, L. D. (1976). *Follow-up evaluation of the second look approach to the selection of management trainees.* Chicago: Sears, Roebuck and Company, National Personnel Department, Psychological Research and Services.

Mitchel, J. O. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology, 60,* 573–579.

Moses, J. L. (1972). Assessment center performance and management progress. *Studies in Personnel Psychology, 4,* 7–12.

Moses, J. L. (1973a). Assessment center for the early identification of supervisory and technical potential. In W. C. Byham & D. Bobin (Eds.), *Alternatives to paper and pencil testing* (pp. 38–49). Pittsburgh, PA: University of Pittsburgh. [Proceedings of a conference at Graduate School of Business]

Moses, J. L. (1973b). The development of an assessment center for the early identification of supervisory potential. *Personnel Psychology, 26,* 569–580.

Moses, J. L., & Boehm, V. R. (1975). Relationship of assessment center performance to management progress of women. *Journal of Applied Psychology, 60,* 527–529.

Neidig, R. D., Martin, J. C., & Yates, R. E. (1978). *The FBI's Management Aptitude Program Assessment Center* (Research Rep. No. 1, TM 78-3). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, Applied Psychology Section.

Orwin, R. G., & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychological Bulletin, 97,* 134–147.

Pearlman, K. (1984, August). *Validity generalization: Methodological and substantive implications for meta-analytic research.* Paper presented at the 92nd Annual convention of the American Psychological Association, Toronto, Ontario, Canada.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict training success and job proficiency in clerical occupations. *Journal of Applied Psychology, 65,* 373–406.

Ritchie, R. J., & Moses, J. L. (1983). Assessment center correlates of women's advancement into middle management: A 7-year longitudinal analysis. *Journal of Applied Psychology, 68,* 227–231.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185–193.

Russell, G. (1975). Differences in minority/nonminority assessment center ratings. *Assessment and Development, 3,* 3, 7, 8.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology, 65,* 643–661.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for jobs in the petroleum industry. *Journal of Applied Psychology, 66,* 261–273.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt–Hunter Bayesian validity generalization procedure. *Personnel Psychology, 32,* 257–281.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407–422.

Schmitt, N., & Hill, T. E. (1977). Sex and race composition of assessment center groups as a determinant of peer and assessor ratings. *Journal of Applied Psychology, 62,* 261–264.

Slivinski, L. W., & Bourgeois, R. P. (1977). Feedback of assessment center results. In J. L. Moses & W. C. Byham (Eds.), *Applying the assessment center method* (pp. 143–159). New York: Pergamon Press.

Task Force on Assessment Center Standards. (1980). Standards and ethical considerations for assessment center operations. *The Personnel Administrator, 25,* 35–38.

Thomson, H. A. (1970). Comparison of predictor and criterion judgments of managerial performance using the multitrait–multimethod approach. *Journal of Applied Psychology, 54,* 496–502.

Thornton, G. C. III, & Byham, W. C. (1982). *Assessment centers and managerial performance.* New York: Academic Press.

Turnage, J. J., & Muchinsky, P. M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment paradigm. *Journal of Applied Psychology, 69,* 595–602.

Appendix

Studies Included in Meta-Analysis

Alexander, H. S., Buck, J. A., & McCarthy, R. J. (1975). Usefulness of the assessment center process for selection to upward mobility programs. *Human Resource Management, 14,* 10–13.

American Airlines. (1976). *A preliminary report on the validity of the Key Manager Human Resources Center.* New York: American Airlines, Personnel Resources Department.

Anstey, E. (1966). The Civil Service Administrative Class and the Diplomatic Service: A follow-up. *Occupational Psychology, 40,* 139–151.

Anstey, E. (1971). The Civil Service Administrative Class: A follow-up of post-war entrants. *Occupational Psychology, 45,* 27–43.

Anstey, E. (1976). Civil Service administrators: A long-term follow-up. *Behavioral Sciences Research Division* (Report No. 31). London: HMS, Civil Service Department.

Bentz, V. J. (1980, June). *Overview of Sears research with multiple assessment techniques.* Paper presented at the meeting of the Eighth International Congress on the Assessment Center Method, Toronto, Canada.

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

Bray, D. W. (1964). The assessment center method of appraising management potential. In J. W. Blood (Ed.), *The personnel job in a changing world* (pp. 225–234). New York: American Management Association.

Bray, D. W. (1982). The assessment center and the study of lives. *American Psychologist, 37,* 180–189.

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

Bray, D. W., Campbell, R. J., & Grant, D. L. (1974). *Formative years in business: A long-term AT&T study of managerial lives.* New York: Wiley.

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

Bullard, J. F. (1969). *An evaluation of the assessment center approach to selecting supervisors.* Peoria, IL: Caterpillar Tractor Company.

Campbell, R. J., & Bray, D. W. (1967). Assessment centers: An aid in management selection. *Personnel Administration, 30,* 6–13.

Carleton, F. O. (1970, September). *Relationships between follow-up evaluations and information developed in a management assessment center.* Paper presented at the 78th Annual Convention of the American Psychological Association, Miami Beach.

Dunne, G. J. Jr., Komar, D. M., Wise, W. W., & Norton, S. D. (1981, April). *An empirical look at an assessment center for R&D managers.* Paper presented at the Ninth International Congress on the Assessment Center Method, San Diego, CA.

Erpenbach, J. J. (March 11, 1971, personal communication)

Gardner, K. E., & Williams, A. P. O. (1973). A twenty-five year follow-up of an extended interview selection procedure in the Royal Navy. *Occupational Psychology, 47,* 1–13.

Grossner, C. (1974). *The assessment of the assessment center.* Unpublished doctoral dissertation, Sir George Williams University, Montreal, Quebec, Canada.

Haynes, M. E. (1978). *Operations supervisor assessment program: Five-year post program evaluation.* Unpublished report, Shell Oil Corporation, Houston, TX.

Hinrichs, J. R. (1969). Comparison of "real life" assessment of management potential with situation exercises, paper-and-pencil ability tests, and personality inventories. *Journal of Applied Psychology, 53,* 425–432.

Hinrichs, J. R. (1978). An eight-year follow-up of a management assessment center. *Journal of Applied Psychology, 63,* 596–601.

Howard, A. (1979, June). *Assessment center predictions sixteen years later.* Paper presented at the Seventh International Congress on the Assessment Center Method, New Orleans.

Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of Black and White females. *Personnel Psychology, 29,* 13–30.

Klimoski, R. J., & Strickland, W. J. (1981). *A comparative view of assessment centers.* Unpublished manuscript.

Kraut, A. I., & Scott, G. J. (1972). Validity of an operational management assessment program. *Journal of Applied Psychology, 56,* 124–129.

McConnell, J. J., & Parker, T. (1972). An assessment center program for multiorganizational use. *Training and Development Journal, 26*(3), 6–14.

McElroy, J. J. (1979). *MacSteel Assessment Centers: Evaluation and validation report* (1977–1979). Unpublished manuscript.

Metropolitan Transit Authority. (1972). *The uses of the assessment center in a government agency's management development program.* Unpublished report.

Mitchel, J. O. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology, 60,* 573–579.

Moses, J. L. (1972). Assessment center performance and management progress. *Studies in Personnel Psychology, 4,* 7–12.

Moses, J. L., & Boehm, V. R. (1975). Relationship of assessment center performance to management progress of women. *Journal of Applied Psychology, 60,* 527–529.

Moses, J. L., & Wall, S. (1975). Pre-hire assessment: A validity study a new approach for hiring college graduates. *Assessment and Development, 2*(2), 11.

Norton, S. (1980, June). *Applying the assessment center method to an upward mobility program.* Paper presented at the meeting of the Eighth International Congress on Assessment Center Method, Toronto, Ontario, Canada.

Parker, T. C. (1980). Assessment centers: A statistical study. *The Personnel Administrator, 25,* 65–67.

Ritchie, R. J. (1980, June). *The validity of an assessment center for selecting telephone directory salespeople.* Paper presented at the meeting of the Eighth International Congress on the Assessment Center Method, Toronto, Ontario, Canada.

Ritchie, R. J., & Moses, J. L. (1983). Assessment center correlates of women's advancement into middle management: A 7-year longitudinal analysis. *Journal of Applied Psychology, 68,* 227–231.

Schmitt, N., Noe, R. A., Meritt, R., & Fitzgerald, M. P. (1984). Validity of assessment center ratings for the prediction of performance ratings and school climate of school administrators. *Journal of Applied Psychology, 69,* 207–213.

Slivinski, L. W., Grant, K. W., Bourgeois, R. P., & Pederson, L. D. (1977). *Development and application of a first level management assessment centre.* Ottawa, Ontario, Canada: Personnel Psychology Centre, Managerial Assessment and Research Division.

Slivinski, L. W., McCloskey, J. L., & Bourgeois, R. P. (1979, June). *Comparison of different methods of assessment.* Paper presented at the meeting of the Seventh International Congress on the Assessment Center Method, New Orleans.

Thomson, H. A. (1969). *Internal and external validation of an industrial assessment program.* Unpublished doctoral dissertation, Case Western Reserve University.

Tziner, A. (1982). *The assessment center goes to boot camp again: An*

application to selection of officer training applicants. Unpublished manuscript.

Tziner, A. (1984). Prediction of peer rating in a military assessment center: A longitudinal follow-up. *Canadian Journal of Administrative Sciences, 1,* 146–160.

Tziner, A., & 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. (1950). The validation of Civil Service Selection Board procedures. *Occupational Psychology, 24,* 75–95.

Warriner, L. (1981, April). *Statistical vs. judgmental prediction of advancement using assessment center data.* Paper presented at the meeting of the Ninth International Congress on the Assessment Center Method, San Diego, CA.

Wilson, N. A. (1948). The work of the Civil Service Selection Board. *Occupational Psychology, 22,* 204–212.

Wissman, D. J., & Rankin, K. K. (1982). *A second look at the validity of a public sector assessment center for research and development managers.* Unpublished manuscript.

Wollowick, H. B., & McNamara, W. J. (1969). Relationship of the components of an assessment center to management success. *Journal of Applied Psychology, 53,* 348–352.

Worbois, G. M. (1975). Validation of externally developed assessment procedures for identification of supervisory potential. *Personnel Psychology, 28,* 77–91.

Received May 14, 1986
Revision received October 27, 1986
Accepted August 22, 1986 ■