PERSONNEL PSYCHOLOGY
1984, 37

METAANALYSES OF VALIDITY STUDIES PUBLISHED BETWEEN 1964 AND 1982 AND THE INVESTIGATION OF STUDY CHARACTERISTICS

NEAL SCHMITT
RICHARD Z. GOODING
RAYMOND A. NOE
and
MICHAEL KIRSCH
Michigan State University

Review and metaanalyses of published validation studies for the years 1964-1982 of *Journal of Applied Psychology* and *Personnel Psychology* were undertaken to examine the effect of (1) research design; (2) criterion used; (3) type of selection instrument used; (4) occupational group studies; and (5) predictor-criterion combination on the level of observed validity coefficients. Results indicate that concurrent validation designs produce validity coefficients roughly equivalent to those obtained in predictive validation designs and that both of these designs produce higher validity coefficients than does a predictive design which includes use of the selection instrument. Of the criteria examined, performance rating criteria generally produced lower validity coefficients than did the use of other more "objective" criteria. In comparing the validities of various types of predictors, it was found cognitive ability tests were not superior to other predictors such as assessment centers, work samples, and supervisory/peer evaluations as has been found in previous metaanalytic work. Personality measures were clearly less valid. Compared to previous validity generalization work, much unexplained variance in validity coefficients remained after corrections for differences in sample size. Finally, the studies reviewed were deficient for our purposes with respect to the data reported. Selection ratios, standard deviations, reliabilities, predictor and criterion intercorrelations were rarely and inconsistently reported. There are also many predictor-criterion relationships for which very few validation efforts have been undertaken.

With the development of metaanalytic procedures (Glass, McGaw, and Smith, 1981; Hunter, Schmidt and Jackson, 1982) and their application to personnel selection (for examples, see Pearlman, Schmidt, and Hunter, 1980; Schmidt, Gast-Rosenberg, and Hunter, 1980; Schmidt and Hunter, 1977; Schmidt, Hunter, and Caplan, 1981; and Schmidt, Hunter,

Pearlman, and Shane, 1979), several problems in the use of tests in employee selection have disappeared or seem considerably less important than previously thought. Most significantly, the body of research generated by Hunter, Schmidt, and their colleagues suggests that test validity generalizes across situations within broad occupational families (see Hunter, Note 1). Most of their validity generalization work involved the use of unpublished studies of measures of cognitive ability. However, Hunter and Hunter (Note 2) have completed metaanalyses on "alternate" predictors as well as cognitive ability tests. In that analysis of entry level jobs, they found no predictor with validity higher than that of cognitive ability tests. In this paper, we present the results of a metaanalysis completed on validation studies published in the *Journal of Applied Psychology* and *Personnel Psychology* between 1964 and 1982. No previous systematic metaanalytic work has been done on these published studies (some were included in work by Reilly and Chao, 1982; Boehm, 1982; and Hunter and Hunter, Note 2). We use our analyses to address several questions we believe are of interest to personnel researchers.

Besides materials published by Hunter, Schmidt and their colleagues, two other review efforts are relevant to the results summarized in this paper. Reilly and Chao (1982) examined the validity of eight categories of alternate predictors. Their conclusion was that only biodata and peer evaluation were supported as having validities approximately equal to those of standardized tests. In another review, Boehm (1982) examined nearly the same studies reviewed in this paper (namely, studies published in the same two journals between 1960 and 1979). Her focus was on a determination of the changes, if any, which have occurred in the volume of published research, the types of research design, occupations investigated, predictors and criteria used, and obtained validities. The overall average validity across all studies reviewed by Boehm was approximately .22 which represents no change from the results of earlier reviews for proficiency criteria (Ghiselli, 1973). However, she did not report average validities for any of the subgroups of studies she examined.

In this paper, using metaanalytic procedures outlined by Hunter, Schmidt and Jackson (1982), we analyze the validities of various subgroups of studies in an attempt to answer five questions. First, we look at validities from studies in which the research design was concurrent, purely predictive, or predictive with selection. In concurrent studies, measures of predictors and criteria are collected from job incumbents. In a purely predictive design, predictor data are collected from job applicants and hiring decisions are made with no knowledge of the predictors. A very common situation in validation research is that in which predictor information is collected from job applicants and also used as

the basis for selection producing range restriction (Thorndike, 1949). Concurrent and predictive studies in which the tests were used to make hiring decisions should yield test validities which are lower. Recently, Barrett, Phillips, and Alexander (1981) referred to four criticisms of the concurrent design that supposedly make its use less than desirable. These include "missing persons", restriction of range, motivational and demographic differences between present employees and job applicants, and confounding by job experience. However, their review of existing data indicated that these differences have a minimal impact on the magnitude of the validity coefficient. For example, an empirical comparison of concurrent and predictive validity coefficients of the General Aptitude Test Battery suggests that the two research designs yield virtually identical coefficients (Bemis, 1968). Further, Schmitt and Schneider (1983) suggest the possibility that there may also be conditions in which range enhancement occurs (obviously this would only be a problem when applying range restriction corrections to concurrent validity coefficients). On the other hand, several authors (Lee, Miller and Graham, 1982; Linn, 1983; Linn, Harnisch and Dunbar, 1981) have affirmed the appropriateness of corrections for range restriction and even their conservative nature in some instances. Examination of validities for different research designs in this paper is directed toward determining the extent to which the level of the observed validity coefficients are associated with the type of research design used in the validation effort.

A second question addressed in this paper is whether validity coefficients vary by the criterion employed in the study. In this connection, it has been standard practice for industrial psychologists to express a preference for "objective" criteria such as productivity, tenure, or salary increases and promotions while settling for "subjective" performance ratings. While a great deal of attention is currently being focused on the determinants of performance ratings (see Ilgen and Feldman, 1983; Landy and Farr, 1980; Wexley and Klimoski, 1984), no previous examination of differences in observed validity coefficients has been undertaken.

Our third question concerns the relative size of validity coefficients for various types of predictors. Similar questions have been addressed by Lent, Aurbach, and Levin (1971), Hunter and Hunter (Note 2) and Reilly and Chao (1982). The dates of the published studies (1964-1982) were set so as to ensure inclusion of all work since the Lent et al. effort and to cover the period of time since EEO concerns became important in personnel selection. Our review includes only published work whereas the Hunter and Hunter effort included much unpublished data. Finally, the Reilly-Chao review focused on predictors which may be considered alternatives to traditional paper-and-pencil measures. Of these three

reviews, only the Hunter and Hunter work included the use of metaanalytic techniques to summarize the validity data.

Fourth, for those categories in which a sufficient number of studies exist, we examined predictor-criterion combinations. This examination is particularly relevant to construct and content validity questions and represents an examination of the importance of the Wernimont and Campbell (1968) distinction among tests that are signs or samples. Wernimont and Campbell suggested that development of predictors which were intended to be actual job samples should result in increased validity coefficients. In other words, if our predictor and criterion measures are both from the same content domain, correlations should be maximized.

The final study characteristic used to subgroup validity coefficients is the occupational group which served as research participants. This, of course, represents one of the major concerns of the initial validity generalization research (Schmidt and Hunter, 1977; Hunter, Note 1). Hunter's recent work (Note 1) on virtually all jobs in the Dictionary of Occupational Titles, suggests that validities are similar within broad job categories, but that there are practically meaningful differences across these categories.

To summarize, our purpose in this paper was to apply metaanalytic methodology to examine validities as a function of five study characteristics: (1) validation research design; (2) the type of criterion used; (3) the type of predictor used; (4) predictor-criterion combinations; and (5) the occupational group studied.

Sample

All studies reporting criterion-related validity studies in *Personnel Psychology* and *Journal of Applied Psychology* between the years of 1964 and 1982 were the source of the metaanalysis reported in this paper. A total of 99 articles were reviewed; 65 came from the *Journal of Applied Psychology* and 34 from *Personnel Psychology*. References to these papers are available from the senior author.

Procedure

Each of the 99 papers was reviewed and the appropriate data coded. Specifically, of interest to this study, validity coefficients, study design, occupational group(s), predictor type(s), and criterion type(s) were coded for each study. An effort was also made to code criterion reliabilities, and the standard deviations of selected and applicant populations and/or the selection ratio but these data were available on a very small proportion of the studies. When appropriate, cross-validated correlations or

validities corrected for shrinkage were recorded.

Intercoder reliability for the various information extracted from the papers was assessed by examining the coding of a subset of 25 papers by three of the authors. Independent agreement exceeded 90 percent for all variables; subsequent discussion of the cases involving disagreement clarified the coding standards and produced agreement in all cases. The remaining studies were coded by the second author. A list and description of the coding categories is available from the senior author. A total of 840 cases or validity coefficients were coded. Most of the subgroups included sizable numbers of validity coefficients (in excess of 30). As has been true in other metaanalytic studies of validity coefficients, many of the 840 coefficients coded were nonindependent observations in the sense that several validity coefficients were computed from data collected on a single group of subjects with several intercorrelated performance criteria (see Hunter, Schmidt, and Jackson, 1982). For each independent sample within a study, validities of the various measures within a single predictor category for a single criterion category were averaged to produce a "summary" validity coefficient. This produced a total of 366 coefficients. It should be noted that while these summary validities were conceptually independent validity coefficients they were not necessarily statistically independent in the sense that criteria intercorrelations were not zero. Analyses of both the total set of validities and the 366 summary validities were conducted; only the latter are reported in this paper. There was little difference between the results of these two analyses; total analyses are available upon request from the senior author.

In averaging the validity coefficients, each coefficient was weighted by its sample size. In addition, the variance of the coefficients (σ_r^2), the variance due to sampling error (σ_e^2), variance remaining after subtracting variance due to sampling error (σ_ϱ^2) and the percent of remaining or unexplained variance were computed using formulas available in Hunter, Schmidt, and Jackson (1982).

No attempt was made to correct the variance of the coefficients for other artifacts such as differences in range restriction or criterion unreliability (Schmidt and Hunter, 1977). Data that would have made these corrections possible were unavailable in the large majority of studies. In studies in which the distributions of these artifacts are assumed or constructed based on available literature or best guesses indicate that most of the variability in validity coefficients can be explained by sampling error. For example, of the percentage of variance in validity coefficients accounted for in one validity generalization study (Schmidt, Hunter, and Caplan, 1981), approximately 90 percent was accounted for by sampling error whereas an additional 10 percent was accounted for by criterion

TABLE 1
Validity Coefficients as Function of Validation Study Design

Design	Number of validities	Sample range	Sample total	$\bar{r}$	σ_r^2	σ_e^2	σ_ϱ^2	Per cent unexplained
Concurrent	153	22-520	17838	.341	.03703	.00670	.03011	82
Predictive	99	19-68616	90552	.296	.00668	.00091	.00577	86
Predictive with selection	114	19-14738	124960	.259	.02140	.00079	.02061	96
Total	366	19-68616	233350	.280	.01750	.00133	.01617	92

reliability, test reliability, and range restriction distributions. Similar results were reported in Pearlman et al. (1980) and Schmidt, Gast-Rosenberg, and Hunter (1980). Consequently removal of artifacts other than sampling error would appear to have little or no effect on conclusions concerning the variability of validity coefficients.

Results

Study Design

In Table 1, we present data relevant to the question concerning the design of a validation study. The average overall observed validity is .28, consistent with previous reviews (Ghiselli, 1973; Boehm, 1982). There appear to be minimal differences across study designs in the average validity coefficient and contrary to conventional wisdom, the concurrent designs actually produce validity coefficients which are slightly superior to predictive designs, especially those predictive designs in which the predictor instruments were used to make hiring decisions. This direct restriction of range, then, may have more serious deflating effects on observed validity coefficients than does the indirect restriction that is assumed to have occurred through attrition and promotion in the typical concurrent study.

Sample sizes vary considerably which suggests that there be concern about averaging the sample size weighted validity coefficients, but the correlation between sample size and validity was .03. It is true, however, that sample sizes in the concurrent studies were smaller and this is reflected in the fact that a greater proportion of the variance in validity coefficients is explained by sampling error. Not surprisingly given the variety of tests, criteria, and occupational groups in these studies, sampling error did not account for much of the variability in validity coefficients.

According to Hunter, Schmidt, and Jackson (1982, pp. 47-48), a moderator variable is indicated when the average correlation varies across subgroups and the corrected variance averages lower in the subsets than

TABLE 2
Validity Coefficients as Function of Various Occupational Groups

Occupational group	Number of validities	Sample range	Sample total	$\bar{r}$	σ_r^2	σ_e^2	σ_ϱ^2	Per cent unexplained
Professional	81	19-2411	18610	.319	.02393	.00351	.02042	85
Managerial	93	24-8885	43188	.335	.01943	.00170	.01773	91
Clerical	36	25-1091	9690	.385	.02284	.00270	.02014	88
Sales	50	22-14738	31732	.170	.00845	.00149	.00696	82
Skilled labor	46	34-3964	37658	.177	.01519	.00115	.01404	92
Unskilled labor	60	47-68616	92472	.314	.00633	.00053	.00580	92
Total	366	19-68616	233350	.280	.01750	.00133	.01617	92

for the whole data. For the data summarized in Table 1, there are small differences in average validity coefficients, but the variances of the subgroups certainly are not smaller than the variance of the total. As a further effort to assess the effect of study design on the size of observed validity coefficients, three dichotomous variables were created by coding each study design 1 and the remaining studies, 0. These dichotomous variables were correlated with the validity coefficients. Correlations were .16, .00, -.17 for concurrent, predictive, and predictive with selection respectively. While the .16 and -.17 correlations are statistically significant, the correlations are certainly not large. Validity does not appear to be underestimated when researchers use concurrent strategies; if there is any difference at all, it seems that concurrent strategies result in higher estimates of validity coefficients than do predictive strategies especially those predictive studies which involve some use of the selection instruments to eliminate potentially low performing employees.

Occupational Group

The results summarizing validation studies over various occupational groups are presented in Table 2. As in Table 1, the average validity coefficients computed for the subgroups involve a wide variety of test-criterion relationships, hence it is not surprising that sampling error does not account for a large portion of the variability in observed coefficients. There are, furthermore, sizable differences in the magnitude of the coefficients for different subgroups with the Sales and Skilled Labor groups having coefficients below .20 and the other groups having coefficients above .30. The average within group corrected variance, σ_ϱ^2, was lower than the variance of coefficients for the total set of coefficients but dichotomous occupational group variables did not correlate highly ($<.16$)

TABLE 3
Predictors and Criteria Used With Various Occupational Groups

	Professional No. of validities	Managerial No. of validities	Clerical No. of validities	Sales No. of validities	Skilled labor No. of validities	Unskilled labor No. of validities
Predictor						
Special aptitude	9	4	8	1	7	2
Personality	21	17	1	6	11	6
Gen. ment. abil.	8	18	12	3	5	7
Biodata	23	4	9	31	13	19
Job Sample	8	3	3	0	4	0
Assess. center	3	15	0	3	0	0
Supervisory/ peer evaluations	4	24	0	3	0	0
Physical ability	0	1	0	0	6	15
Criterion						
Performance rating	43	31	12	15	17	22
Turnover	5	0	9	11	12	11
Achievement/grades	8	11	3	4	14	3
Productivity	7	0	0	20	0	3
Status change	4	33	0	0	0	9
Wages	13	17	0	0	0	3
Work samples	0	1	12	0	2	9

with the validity coefficients.

Some of the differences across occupational groups in validity coefficients could also be due to the particular type of predictor or criterion used in validation research. Table 3 is a summary of the number of instances a particular criterion or predictor was used for each occupational group.

Studies of Sales and Skilled Labor groups most frequently involved the use of personality and biodata as predictors and, relative to other groups, more frequently used turnover as a criterion. The generally lower validities associated with the prediction of turnover and the use of personality measures (see Tables 4 and 5) may account for the lower validities for the Sales and Skilled and Unskilled Labor groups.

Further breakdowns of the validity coefficients for predictor- criterion relationships by occupational subgroups were also done. Data from these analyses are not reported here (though available upon request) because for many of the predictor-criterion-occupational subgroup categories, the number of validity studies available was simply too few. Especially noteworthy was the fact that little information concerning physical ability measures is available even for skilled and unskilled occupational groups in which they may be useful. Also, there are few studies for any occupational subgroup which involve the use of production criteria.

TABLE 4
Validity Coefficients as a Function of Type of Predictor

Predictor	Number of validities	Sample range	Sample total	$\bar{r}$	σ_r^2	σ_e^2	σ_ρ^2	Per cent unexplained
Special aptitude	31	19-1091	4315	.268	.02083	.00619	.01464	70
Personality	62	24-3964	23413	.149	.01109	.00253	.00856	77
Gen. mental ability	53	24-8885	40230	.248	.01908	.00117	.01791	94
Biodata	99	22-14738	58107	.243	.01831	.00151	.01680	92
Work sample	18	19-1091	3512	.378	.01139	.00377	.00762	67
Assessment center	21	35-8885	15345	.407	.00250	.00095	.00155	62
Supervisor/peer evaluations	31	30-1979	6620	.427	.03046	.00313	.02733	89
Physical ability	22	55-588	3103	.315	.04865	.00575	.04290	88
Total	366	19-68616	233350	.280	.01750	.00133	.01617	92

Predictor Type

Average validity coefficients for various types of predictors are presented in Table 4. There are substantial differences in the average validity coefficients with personality measures being associated with the lowest validities and work samples, assessment centers, and supervisor and peer evaluations most highly correlated with criteria. The average within predictor type variance σ_e^2, was slightly lower than the variance of coefficients for the total set of coefficients. Two of the dichotomously scored predictor type variables did correlate with the validity coefficients ($r = -.27$ and $.15$, $p < .01$, for personality and physical ability measures, respectively). It is also noteworthy that average validities for special aptitudes and general mental ability are lower than those for predictors we classified as work samples, supervisor or peer evaluations, and assessment centers.

Criterion Type

In Table 5, average validity coefficients obtained with various criteria are listed. A concern among industrial psychologists has been the extensive use of performance rating criteria and the relative lack of use of more "objective" criteria. Due to their sensitivity to various biases, performance ratings may either inflate validity estimates or result in the inappropriate weighting of certain predictors. Performance ratings yield slightly better validity coefficients than do turnover criteria and productivity criteria but much lower validity coefficients than those associated with work samples, wages, and status changes. Feature correlations in the form of dichotomously scored criterion type variables were all less

TABLE 5
Validity Coefficients as a Function of Type of Criterion

Criterion	Number of validities	Sample range	Sample total	$\bar{r}$	σ_r^2	σ_e^2	σ_ϱ^2	Per cent unexplained
Performance ratings	140	22-520	17559	.260	.03051	.00693	.02358	77
Turnover	48	37-68616	127021	.246	.01104	.00033	.01071	97
Achievement/ grades	43	19-453	7156	.270	.03971	.00516	.03455	87
Productivity	30	50-3590	14869	.208	.00584	.00185	.00399	68
Status change	46	30-8885	52686	.359	.01303	.00066	.01237	95
Wages	33	47-443	5470	.378	.02278	.00443	.01835	81
Work samples	24	77-1091	8244	.401	.02638	.00205	.02433	92
Total	366	19-68616	233350	.280	.01750	.00133	.01617	90

than .10 except for the work sample correlation ($r = .18\, p < .01$). While these results yield no information concerning appropriate/inappropriate weighting of predictors when one uses various criteria, the results certainly do not support the belief that use of so-called subjective criteria will result in inflated validity coefficients. If anything, use of performance rating criteria results in lower validity coefficients than does the use of other criteria.

Predictor-Criterion Relationships

A question relevant to concerns about construct validity is whether validities for certain predictor-criterion relationships are higher than others. Certainly, personnel researchers will use those predictors which research, training, and experience indicate will be most useful in the prediction of given criteria, but are there any substantial differences? In Table 6, we summarize the data concerning predictor-criterion relationships.

Several points concerning the data summarized in Table 6 are worth noting. First, performance ratings are best predicted by assessment centers and supervisor-peer evaluations, both of which are themselves, rating predictors. Validity coefficients for biodata and work samples are also relatively high while those associated with paper and pencil tests (special aptitude, personality, and general mental ability) are lower. It is also true that studies involving work samples, assessment centers, and supervisory or peer evaluations as predictors were relatively few in number and the total sample size associated with these average validities, low.

Studies using turnover as a criterion have almost exclusively used biodata as a predictor presumably because of the notion that past behavior with respect to job or life changes is the best predictor of future behavior. Of those studies available, this seems to be true; though all validity coef-

TABLE 6
Average Validity Coefficients for Various Predictor-Criterion Combinations

Predictor	Number of validities	Total sample	$\bar{r}$	σ_r^2	σ_e^2	σ_ρ^2	Per cent unexplained
			Performance ratings				
Special aptitude	14	838	.162	.02841	.01584	.01257	44
Personality	32	4065	.206	.03531	.00722	.02809	80
Gen. mental ability	25	3597	.220	.01563	.00629	.00934	60
Biodata	29	3998	.317	.03566	.00587	.02979	84
Work sample	7	384	.319	.01081	.01471	—	0
Assessment center	6	394	.428	.00259	.01016	—	0
Supervisor/peer evaluations	12	1389	.315	.03140	.00701	.02439	78
			Turnover				
Personality	5	15927	.121	.00104	.00030	.00074	71
Gen. mental ability	8	12449	.141	.01877	.00062	.01815	97
Biodata	28	28862	.209	.01444	.00089	.01355	94
Physical ability	3	852	.154	.00762	.00336	.00426	56
			Achievement/grades				
Special aptitude	8	1093	.275	.03622	.00625	.02997	83
Personality	6	980	.152	.01406	.00584	.00822	58
Gen. mental ability	5	888	.437	.02209	.00369	.01840	83
Biodata	9	1744	.226	.07841	.00465	.07376	94
Work sample	3	95	.314	.01876	.02566	—	00
Assessment center	3	289	.312	.00692	.00846	—	00
Physical ability	4	976	.281	.00327	.00348	—	00
			Productivity				
Biodata	19	13655	.203	.00362	.00128	.00234	65
			Status change				
Personality	7	561	.126	.03139	.01208	.01931	61
Gen. mental ability	9	21190	.282	.00880	.00036	.00844	96
Biodata	6	8008	.332	.00144	.00059	.00085	59
Assessment center	8	14361	.412	.00151	.00038	.00113	75
Supervisor/peer evaluations	9	4224	.512	.01537	.00116	.01421	92
Physical ability	3	245	.613	.00028	.00477	—	—
			Wages				
Personality	10	1720	.268	.00903	.00501	.00402	45
Biodata	7	1544	.525	.01571	.00238	.01333	85
Work sample	4	1191	.438	.00547	.00219	.00328	60
Assessment center	4	301	.237	.00531	.00184	—	00
Supervisor/peer evaluations	4	301	.206	.00737	.01219	—	00
			Work sample				
Special aptitude	3	1793	.280	.00423	.00142	.00281	66
Gen. mental ability	3	1793	.426	.00660	.00112	.00548	83
Work sample	3	1793	.353	.01126	.00128	.00998	89
Physical ability	11	959	.419	.08924	.00784	.08140	91

[a]All predictor-criterion combinations for which less than three coefficients were available were ignored.

ficients for turnover criteria tend to be low. These low validities may occur because of the nature of the turnover criterion, but may also be due to the fact that our predictor instruments do not reflect the wide range of potential determinants of turnover such as organizational commitment, perceptions of the labor market, and job satisfaction (Mobley, Griffeth, Hand, and Meglino, 1979). Further analyses of the biodata might be helpful since only six percent of the variability in validity coefficients was explained by differences in the sample sizes across studies. Measures of achievement/grades are best predicted by general mental ability tests and least well by personality tests. Perhaps the widest variety of predictors have been used to predict achievement and grades hence the number of studies and total sample size for any given predictor category is relatively low. Somewhat surprisingly, the only selection instrument used frequently to predict productivity has been biodata and, as with turnover, the validity coefficients are modest.

Status change appears to be best predicted by supervisor/peer evaluations and assessment center ratings. Personality measures yield very low coefficients though total sample size is small. General mental ability correlates relatively low with status change though the average validity coefficient is about the same as that for the total set of coefficients.

Wages are best predicted by biodata and work samples and least well predicted by assessment centers and supervisor/peer evaluations. All results for wages are based on a small number of validity coefficients and low total sample size, however.

Work samples have been predicted most frequently by physical ability measures (though total sample size is only 959) with relatively good results (r_{xy} = .419). In reviewing Table 6, it is evident from the amount of unexplained variance in validity coefficients that there may be other moderators of the observed validity coefficient besides the type of predictor or criterion.

Discussion and Conclusions

Throughout Tables 1-6, it is evident that validities exhibit considerable variability after corrections for sample size variability. While Hunter, Schmidt, and their colleagues have usually reported that 50 to 100 percent of the variance in validities can be explained by sample size differences, the results of our metaanalyses do not indicate that much more than 25 percent of the variability is due to sample size differences. It may very well be that our data sources are more variable in that (1) different predictors are used even within a predictor category; (2) the researchers in our studies varied more in the way data were collected and analyzed; or (3) the organizational settings of the published studies which

were our source of data were more variable. The latter is purely speculative, but the fact remains that our corrections for sample size variability explained a relatively small portion of the total variability in validity coefficients.

Another obvious fact is that for many predictor-criterion relationships, we still lack adequate data with which to draw conclusions even when using metaanalysis of all available data. For some predictor-criterion relationships it would make little sense to collect data. For example, the turnover-general mental ability relationship does not seem interesting unless one has a theoretical rationale based perhaps on the complexity of the job and the relative speed with which persons with varying levels of mental ability become bored and leave. Productivity should be correlated with a wider range of predictor variables. There is of course a much larger body of unpublished literature which Hunter and Schmidt and their colleagues have summarized especially on cognitive paper-and-pencil measures (see Hunter, Note 1; Hunter and Hunter, Note 2). Unpublished literature on less widely used and nontraditional selection instruments does not seem to be as large and some potentially useful methods such as miniaturized training sessions, situational interviews, and unassembled examinations have not been frequently studied (Reilly and Chao, 1982).

As noted above, there is little evidence that concurrent studies of test validity yield different results than predictive studies. This suggests that motivational effects and/or job experience effects that are generally cited as reasons for not employing concurrent validation strategies (Guion, 1965) may not be that important. Alternatively, job experience may enhance the range of job performance and test scores in concurrent studies thus artificially inflating validity coefficients. This hypothesis has been suggested elsewhere (Schmitt and Schneider, 1983) and conflicts with the more popular belief that job experience and/or selection and attrition serve to restrict the variance of predictor and criterion in concurrent studies. Obviously data concerning the standard deviation of selection instruments and criteria collected at various times before and after employment are required to evaluate what degree of range restriction or enhancement occurs.

The use of performance ratings as criteria in validation studies does not serve to inflate observed validity coefficients. The lower observed validity coefficients for performance rating criteria counterindicate the concern that these coefficients are inflated because of various biases whereas more "objective" criteria would not be affected by these biases. One reason performance rating criteria yield lower validities may be because their reliability is lower than the reliability associated with criteria such as tenure, wages, or status changes.

Results concerning different types of predictors are consistent with previous literature reviews (Ghiselli, 1973; Guion and Gottier, 1965) which indicate that personality tests have low validity. The data are not consistent with Hunter and Hunter's conclusion (Note 2) that cognitive ability tests are superior to other predictors. The data summarized in this paper indicate that work samples, assessment centers, and supervisor/peer evaluations yield validities which are superior to those of general mental ability and special aptitude tests which are closest to those labeled ability measures by Hunter and Hunter (Note 2). There are likely several reasons for the difference between our work and that of Hunter and Hunter, but at least one occurs to the authors. Our data consist solely of published work in two journals over the past two decades while the work summarized by Hunter and Hunter (Note 2) included a large portion of unpublished data. Much of the published research was directed to the development and study of alternate predictors; more traditional tests were included only as standards of comparison or because they were available in many of these studies. In much of the work which served as the source of validity coefficients for Hunter and Hunter, the paper-and-pencil ability measures were carefully developed and standardized measures. As noted above, the studies which we reviewed may also have been more variable along several dimensions than those which were the source of the Hunter analyses.

Several other conclusions are similar to those stated by previous authors doing metaanalytic work on validation studies (Hunter, Schmidt, and Jackson, 1982; Callender and Osburn, 1980). Data for accurate assessment of the effect of artifacts in personnel selection studies is largely unavailable. Authors simply do not report the predictor or criterion standard deviations of applicant and incumbent groups nor do they report the selection ratio or the reliability of the measures they use. Second, use of credibility values for the various average validities obtained in this paper indicates one should have reasonable confidence in obtaining nonzero validity coefficients using various selection instruments and criteria for all job families (see Table 7). Our results are not as consistent with previous validity generalization work if one considers the amount of variability in validity coefficients accounted for by sampling error. Even when validities were averaged for a single occupational group for a single criterion-predictor relationship (analyses available from senior author), large portions of the validity variance remained in many cases. Finally, as is evident by examining our tables there are a wide variety of predictor-criterion relationships for which we have very little data and for which initial results are encouraging. Rather than squelching validity research, metaanalytic work should serve to redirect and stimulate both the actual validation studies and their detailed reporting.

One final note of caution in the use of the metaanalytic procedures outlined by Hunter, Schmidt, and Jackson (1982) is appropriate. In those cases where the number of validities was small (<6), we nearly always found that corrections for differences in sample size accounted for most of the validity coefficient variability. Whenever the number of studies was greater, the corrections did not account for nearly as large a proportion of the validity variance. This suggests that these corrections are likely inappropriate or misleading when the number of different validities over which one is averaging is small.

Reference Notes

1. Hunter, J. E. *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery.* Unpublished manuscript, 1982.
2. Hunter, J. E., and Hunter, R. F. *The validity and utility of alternate predictors of job performance.* Unpublished manuscript, 1981.

References

Barrett, G. V., Phillips, J. S., and Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology, 66,* 1-6.

Bemis, S. E. (1968). Occupational validity of the general aptitude test battery. *Journal of Applied Psychology, 52,* 240-249.

Boehm, V. R. (1982). Are we validating more but publishing less? (The impact of governmental regulation on published validation research—an explanatory investigation). PERSONNEL PSYCHOLOGY, *35,* 175-187.

Callender, J. C., and Osburn, H. G. (1980). Development and test of a new model for generalization of validity. *Journal of Applied Psychology, 65,* 543-558.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. PERSONNEL PSYCHOLOGY, *26,* 461-477.

Glass, G. V., McGaw, B., and Smith, M. L. (1981). *Metaanalysis in social research.* Beverly Hills, CA: Sage.

Guion, R. M. (1965). *Personnel testing.* New York: McGraw-Hill.

Guion, R. M., and Gottier, R. F. (1965). Validity of personality measures in personnel selection. PERSONNEL PSYCHOLOGY, *18,* 135-164.

Hunter, J. E., Schmidt, F. L., and Jackson, G. B. (1982). *Metaanalysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Ilgen, D. R., and Feldman, J. M. (1983). Performance appraisal: A process focus. *Research in Organizational Behavior, 5,* 141-197.

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

Lee, R., Miller, K. J., and Graham, W. K. (1982). Corrections for restriction of range and attenuation in criterion-related validation studies. *Journal of Applied Psychology, 67,* 637-639.

Lent, R. H., Aurbach, H. D., and Levin, L. S. (1971). Predictors, criteria, and significant results. PERSONNEL PSYCHOLOGY, *24,* 519-533.

Linn, R. L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement, 20,* 1-16.

Linn, R. L., Harnisch, D. L., and Dunbar, S. B. (1981). Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. *Journal of Applied Psychology, 66*, 655-663.

Mobley, W. H., Griffeth, R. W., Hand, H. H., and Meglino, B. M. (1979). Review and conceptual analysis of the employee turnover process. *Psychological Bulletin, 86*, 493-522.

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

Reilly, R. R., and Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. PERSONNEL PSYCHOLOGY, *35*, 1- 62.

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

Schmidt, F. L., and 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., and Caplan, J. (1981). Validity generalization results for two job groups in the petroleum industry. *Journal of Applied Psychology, 66*, 261-273.

Schmidt, F. L., Hunter, J. E., Pearlman, K., and Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. PERSONNEL PSYCHOLOGY, *32*, 257-281.

Schmitt, N., and Schneider, B. (1983). Current issues in personnel selection. In K. M. Rowland and J. Ferris (Eds.), *Research in personnel and human resources management, 1*. Greenwich, Connecticut: JAI Press.

Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.

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

Wexley, K. N., and Klimoski, R. (1984). Performance appraisal: An update. In K. M. Rowland and G. D. Ferris (Eds.), *Research in Personnel and Human Resources Management, 2*. Greenwich, Connecticut: JAI Press.