Journal of Applied Psychology
1993, Vol. 78, No. 1, 3-12

0021-9010/93/$3.00

Refinements in Validity Generalization Methods: Implications for the Situational Specificity Hypothesis

Frank L. Schmidt, Kenneth Law, John E. Hunter, Hannah R. Rothstein, Kenneth Pearlman, and Michael McDaniel

Using a large database, this study examined three refinements of validity generalization procedures: (a) a more accurate procedure for correcting the residual *SD* for range restriction to estimate SD_ρ, (b) use of $\bar{r}$ instead of study-observed *r*s in the formula for sampling error variance, and (c) removal of non-Pearson *r*s. The first procedure does not affect the amount of variance accounted for by artifacts. The addition of the second and third procedures increased the mean percentage of validity variance accounted for by artifacts from 70% to 82%, a 17% increase. The cumulative addition of all three procedures decreased the mean SD_ρ estimate from .150 to .106, a 29% decrease. Six additional variance-producing artifacts were identified that could not be corrected for. In light of these, we concluded that the obtained estimates of mean SD_ρ and mean validity variance accounted for were consistent with the hypothesis that the true mean SD_ρ value is close to zero. These findings provide further evidence against the situational specificity hypothesis.

The first published validity generalization research study (Schmidt & Hunter, 1977) hypothesized that if all sources of artifactual variance in cognitive test validities could be controlled methodologically through study design (e.g., construct validity of tests and criterion measures, computational errors) or corrected for (e.g., sampling error, measurement error), there might be no remaining variance in validities across settings. That is, not only would validity be generalizable based on 90% credibility values in the estimated true validity distributions, but all observed variance in validities would be shown to be artifactual and the situational specificity hypothesis would be shown to be false even in its limited form. However, subsequent validity generalization research (e.g., Pearlman, Schmidt, & Hunter, 1980; Schmidt, Gast-Rosenberg, & Hunter, 1980; Schmidt, Hunter, Pearlman, & Shane, 1979) was based on data drawn from the general published and unpublished research literature, and therefore it was not possible to control or correct for the sources of artifactual variance that can generally be controlled for only through study design and execution (e.g., computational and typographical errors, study differences in criterion contamination). Not unexpectedly, many of these meta-analyses accounted for less than 100% of observed validity variance, and the average across studies was also less than 100% (e.g., see Pearlman et al., 1980; Schmidt et al., 1979).

The conclusion that the validity of cognitive abilities tests in employment is generalizable is now widely accepted (e.g., see American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1985; Anastasi, 1988; Linn & Dunbar, 1985; Sackett, Schmitt, Tenopyr, Kehoe, & Zedeck, 1985; Society for Industrial and Organizational Psychology, 1987). Even when less than 100% of observed variance in validities is demonstrated to be due to artifacts, validity generalization findings usually show that the remaining variance in validities is small enough that the test will have some useful degree of validity for that job in new employment settings. Any of the several methods of estimating the generalizability of validities can be used to show this, including the simplest method based on correcting observed validity variance only for sampling error (Pearlman et al., 1980; Schmidt, Gast-Rosenberg, & Hunter, 1980). The stronger hypothesis that all variation in validities for any given job and ability construct (test type) is artifactual (i.e., the hypothesis that situational specificity has been disproven) has been less widely accepted (Sackett et al., 1985). The basis for reluctance to accept this hypothesis appears to be that most validity generalization analyses do not account for all the variance in validities across studies; some would suggest that perhaps the remaining variation of validities is due to true (but limited) situational differences.

However, there is evidence from five meta-analyses supporting the stronger hypothesis that there is no real situational specificity at all. In three of these meta-analyses, studies were carefully conducted by the same research team in multiple organizations in different geographical areas using the same tests and criterion measures in all studies and carefully controlling data quality and computational, typographical, and other errors (Dunnette et al., 1982; Dye, 1982; Peterson, 1982). In all three meta-analyses, it was found that, on average, all variance across settings was accounted for by artifacts, chiefly sampling error. These studies indicate that when it is possible to control for sources of variance that cannot be corrected for statistically, the findings contraindicate the hypothesis of situational specificity.

The remaining two meta-analyses are based on a different

Frank L. Schmidt and Kenneth Law (now at the Australian Graduate School of Management), Department of Management and Organizations, University of Iowa; John E. Hunter, Department of Psychology, Michigan State University; Hannah R. Rothstein, Department of Management, Baruch College, City University of New York; Kenneth Pearlman, AT&T; Michael McDaniel, Department of Psychology, University of Akron.

Correspondence concerning this article should be addressed to Frank L. Schmidt, Department of Management and Organizations, 651 Phillips Hall, University of Iowa, Iowa City, Iowa 52242-1323.

approach. The traditional situational specificity hypothesis makes two predictions, one well-known and obvious and one that is less well-known and more subtle. The well-known and obvious prediction is that there are real differences between jobs that cause validity coefficients to vary from setting to setting. Validity generalization research tests this prediction by examining the extent to which the variance of observed validities is due to artifacts.

The second prediction made by the situational specificity hypothesis is that if the setting (job, test, criterion, organization, and applicant pool) remains constant, validity will not vary. That is, if the factors hypothesized to cause variation in validity do not vary, validity will not vary. That this prediction is indeed made can be seen by consideration of the older assumption that local criterion-related validity studies provide a solution to the problem of situational specificity, that is, the belief that a local validity study will accurately calibrate the validity in that setting. This belief indicates that advocates of situational specificity did not understand the operation of sampling error; that is, they did not realize that sampling error alone would cause wide variation in observable validities, making it impossible to accurately calibrate validities in small sample situational studies. Furthermore, they appear to have falsely believed that use of statistical significance tests "controlled" for any problems that sampling error might create (see Hunter & Schmidt, 1990b; Schmidt, Hunter, & Pearlman, 1980; Schmidt, Ocasio, Hillery, & Hunter, 1985). This is clearly not a sophisticated form of situational specificity hypothesis, but it was the predominant hypothesis in personnel psychology for more than 50 years and is still held by some. The principles of validity generalization and meta-analysis, on the other hand, predict that small sample validities observed in the same setting will vary substantially, mostly because of sampling error. These opposing predictions were tested in two studies (Schmidt & Hunter, 1984; Schmidt, Ocasio et al., 1985). In both studies, validities observed within the same setting were found to vary markedly across studies, both in magnitude and in significance levels. Of particular significance was the fact that the level of variation was comparable to that typically observed across studies conducted in different settings, where the observed variance would include not only sampling variance and other artifactual variance, but also variance due to the hypothesized situational moderators. This finding casts doubt on the existence of such hypothesized situational moderators. Sampling error was shown to account for all of the observed variation. These findings strongly undercut the situational specificity hypothesis.

These findings led us to reexamine current validity generalization methods to determine whether they can be improved as procedures for testing the situational specificity hypothesis. These methods are adequate for demonstrating validity generalizability, as noted earlier. But successfully addressing the question of whether the small amount of variance remaining in the results of current validity generalization studies is attributable to artifacts or to real but small moderators requires more precise methods of estimating artifactual variance. We have identified three correctable imperfections in current validity generalization methods: (a) use of an approximation that overestimates the standard deviation of true validities (i.e., treating the correction for range restriction as if it were linear); (b) use of observed *r*s from individual studies in the formula for sampling error, rather than the mean observed *r* for the set of studies (McDaniel & Hirsh, 1986; Hunter & Schmidt, 1992); and (c) inclusion of non-Pearson correlations (with their larger sampling error variances) in validity generalization data sets.

The present article explains each of these imperfections, presents improved procedures, and illustrates the impact of the improved procedures by applying them to data from an existing large validity generalization database and comparing the original results with the results yielded by the improved procedures.

Nonlinearity in the Range Correction

In artifact distribution–based methods of meta-analysis, the mean ($\bar{\rho}$) and standard deviation (SD_ρ) of true correlations are estimated from the mean ($\bar{r}_{res}$) and standard deviation (SD_{res}) of the residual distribution. The residual distribution is the distribution of observed correlations expected across studies if N were always infinite (i.e., no sampling error) and reliability and range restriction were always constant at their respective mean values (Law, Schmidt, & Hunter, 1991b). To correct the residual distribution for unreliability, we could divide every value in that distribution by the mean square root of the reliabilities. But because that value is a constant, we can instead just divide both $\bar{r}_{res}$ and SD_{res} by that constant and get the same result. This is what our artifact distribution–based meta-analysis procedures do. But these procedures do exactly the same thing in correcting the residual distribution for the effects of mean range restriction, and here this approach does not work as accurately.

Current procedures use the mean level of range restriction (in the form of the ratio of the restricted to the unrestricted predictor standard deviations) to correct $\bar{r}_{res}$. This increases $\bar{r}_{res}$ by some factor, say 1.50. Then SD_{res} is multiplied by this same factor to estimate the SD of a distribution in which each r has been corrected for the mean level of range restriction. However, unlike the reliability correction, the range restriction correction is not linear in r. The correction is not the same for every value of r in the residual distribution. Instead, it is larger for smaller *r*s and smaller for larger *r*s. Thus, the approximation based on the assumption of linearity in artifact distribution–based meta-analysis procedures leads to overestimates of SD_ρ.

Simulation studies (Callender & Osburn, 1980; Raju & Burke, 1983) have demonstrated that the interactive procedure—theoretically, our most sophisticated method (see Schmidt, Gast-Rosenberg, & Hunter, 1980)—yields estimates of SD_ρ that are too large by .02 or more. This overestimation occurs in simulated data in which sample sizes are infinite (eliminating sampling error) and sources of artifactual variance such as computational errors, outliers, and non–Pearson *r*s do not exist. This overestimation stems from failure to take into account the nonlinearity of range restriction corrections. This nonlinearity can be taken into account by correcting each value in the residual distribution separately for the mean level of range restriction. To take this nonlinearity into account, the following method can be used (Law et al., 1991b). After determining the mean and SD of the residual distribution, one specifies 60 additional values in that distribution by moving out

from the mean in .1 *SD* units to 3 *SD* above and below the mean. Then one corrects each of these values individually for range restriction, using the mean of the *s*/*S* ratio. The formula used to correct each value is

$$R_i = \frac{r_i(S/s)}{\{[(S/s)^2 - 1]r_i^2 + 1\}^{1/2}},$$

where r_i = the value of the correlation in the residual distribution, R_i = the corrected value, S = the unrestricted standard deviation, and s = the mean-restricted standard deviation.

Each range-corrected r is then corrected for the mean effect of unreliability. The relative frequency of each value of r is indexed by the normal curve ordinate associated with its z score in the residual distribution. These frequencies are applied to the corresponding corrected correlations ($\hat{\rho}_i$). The frequency-weighted mean of the distribution of the corrected correlations ($\bar{\rho}$) is then determined, and the following frequency-weighted variance formula is used to find S_ρ^2:

$$S_\rho^2 = \frac{\Sigma f_i(\hat{\rho}_i - \bar{\rho})^2}{\Sigma f_i},$$

where f_i is the frequency associated with $\hat{\rho}_i$. The square root of this value (i.e., SD_ρ) is a more accurate estimate of the standard deviation of true validities. This procedure is discussed in more detail in Law et al. (1991b). That study used computer simulation to compare the accuracy of this procedure with that of the older linear procedure and found the new procedure to be considerably more accurate.

Use of $\bar{r}$ Instead of r in the Sampling Error Formula

The well-known formula for the sampling error variance of sample correlation coefficients is

$$S_e^2 = \frac{(1 - \rho_{xy}^2)^2}{N - 1},$$

where N is the sample size and ρ_{xy} is the population (uncorrected) correlation. But ρ_{xy} is unknown, and to use this formula, some method must be found to estimate it. In single studies, the estimate of ρ_{xy} typically used—because it is the only one available—is the observed correlation in the study at hand. In our early meta-analyses of employment test validities, we followed this tradition: The value used to estimate the sampling error variance in every study was the observed correlation in that study. We have since found that this procedure is not optimal. The mean observed correlation ($\bar{r}_{obs}$)—a good estimate of ρ_{xy}—is typically about .20 in this literature. Sample sizes are usually small, so there are substantial departures in both directions from ρ_{xy}. When the sampling error is large and positive (e.g., +.20, so that $r = .40$), the estimated S_e^2 is substantially reduced (by 23% in this example). But this effect is not symmetrical. When sampling error is large and negative (e.g., −.20, so that $r = .00$), estimated S_e^2 is increased by only a small amount (by 9% in this example). Thus, on balance, the sampling error in a set of correlations is substantially underestimated. The smaller the sample size in the studies analyzed, the greater this underestimation will be. Also, the smaller the (attenuated) population correlation, the greater the underestimation will be (because smaller ρ_{xy} values yield larger sampling error variances if sample sizes are equal). The result is underestimation of the amount of variance accounted for by sampling error and overestimation of SD_ρ. This distortion can be eliminated by using the $\bar{r}$ for the set of studies rather than individual rs in the formula for sampling error. The $\bar{r}$ contains little sampling error, and extreme values are unlikely. The result should be more accurate estimates of SD_ρ.

Law, Schmidt, and Hunter (1991a) tested this hypotheses for realistic data combinations using computer simulation. They found that for both the homogeneous case (no variation in population correlations) and the heterogeneous case (variable population correlations), use of $\bar{r}$ yielded more accurate estimates of sampling error variance than use of individual study rs.

Millsap (1988), in a Monte Carlo study, used r rather than $\bar{r}$ in the formula for sampling error variance. In his study, all ρ were equal so S_ρ^2 was zero, and the variance of the observed rs was solely sampling error variance, that is, $S_r^2 = S_e^2$. However, he found that his formula-derived estimates of S_e^2 were slightly smaller than the observed S_r^2 figures, and this discrepancy was larger for smaller sample sizes. He attributed this finding to inaccuracy in the formula (the formula is an approximation), but the phenomenon described here is in large part the explanation for his findings. He also found that the negative bias in his formula-derived estimates of sampling error variance was larger when measures had lower reliability. This finding is explained by the fact that lower reliability leads to lower values of ρ_{xy}, the operative population correlation. Lower ρ_i values have larger sampling error variances for any fixed sample size, thus intensifying the process described earlier. Thus, it was not unreliability (measurement error) per se that caused the increase in the underestimation, but rather the reduced value of the population correlation.

Presence of Non–Pearson rs

It is well-known that commonly used non–Pearson correlation coefficients, such as the biserial and tetrachoric, have larger standard errors than do Pearson rs. Thus, the formula for the sampling error variance of the Pearson correlation underestimates the amount of sampling error variance in these correlations. When such correlations are included in a meta-analysis, they are treated as if their standard errors were those of Pearson rs. This treatment deflates the estimate of variance accounted for by artifacts and inflates the estimate of SD_ρ in any distribution of correlations in which biserial, triserial, or tetrachoric correlations are present. If the standard formula for the sampling error variance of a correlation coefficient is used, then more accurate results can be obtained if non–Pearson rs are deleted prior to the meta-analysis. Biserial and tetrachoric correlation can be included in meta-analysis, but more complicated formulas are required (Hunter & Schmidt, 1990a, 1990b), and their use requires information that is sometimes not presented in individual studies.

Other Considerations Related to Accuracy of Estimates

An additional issue that bears on tests of the hypothesis of situational specificity is that of second-order sampling error.

This terms refers to the sampling problem created by using small numbers of studies in a validity generalization analysis. The outcome of any such analysis depends to some extent on which studies randomly happen to be available; that is, the outcome depends in part on study properties that vary randomly across studies. This is true even if all relevant, currently existing studies are included. It affects estimates of the standard deviation more than it affects estimates of the mean. (This is also the case with ordinary sampling error and ordinary statistics.) This phenomenon has been examined in depth elsewhere (Hunter & Schmidt, 1990b, chap. 9; Schmidt, Hunter, Pearlman & Hirsh, 1985). The major point is that whereas the formula for sampling error variance correctly predicts the amount of variance sampling error will produce on the average, in specific sets of studies, sampling error randomly produces more than the predicted amount of variance sometimes and randomly less variance other times. The larger the number of studies (other things equal), the smaller are the deviations expected from observed variance. However, if the number of studies is small, these deviations can be quite large on a percentage basis (although absolute deviations are usually small, even in such cases).

In cases in which the same theoretical considerations apply to a number of meta-analyses, the problem of second-order sampling error can be addressed using a meta-analysis of meta-analyses, also known as a second-order meta-analysis. Validity generalization research on cognitive ability tests is one example. Under the situational specificity hypothesis, the hypothesized situational moderators would be essentially the same for different abilities. The traditional situational specificity hypothesis is very vague; it states merely that unknown and unobservable variables that differ across situations cause observed validities to vary. What these variables were was never specified and it was never stated that these factors would be different for different abilities, other things being equal. Therefore, it appears appropriate to assume that these vague factors were not considered to be different for different abilities, considering that the situational specificity hypothesis was applied in apparently the same manner to all abilities (and, indeed, all predictors). Under the alternate hypothesis, all variance would be hypothesized to be artifactual for all abilities. The second-order meta-analysis consists of computing the average percentage of variance accounted for across the several meta-analyses. It should be clear that in conducting a second-order meta-analysis, figures greater than 100% should not be rounded down to 100%. Doing so would obviously bias the mean for these figures, given that those that are randomly lower than 100% are not rounded upward. Technical considerations in conducting second-order meta-analyses are discussed in more detail later. The key point is that it is the reciprocal of the percentage variance accounted for that must be averaged across meta-analyses; that is, the relevant mean is the harmonic mean (for an example, see Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990; see also Callender & Osburn, 1988). The methods of second-order meta-analysis are used later in the present study.

Spector and Levine (1987) conducted a computer simulation study aimed at evaluating the accuracy of the formula for the sampling error variance of r. In their study, the value of ρ was always zero, so the formula for the sampling error variance of observed rs was $S_e^2 = 1/(N - 1)$. They conducted simulation studies for various values of N, ranging from 30 to 500, and the number of observed rs per meta-analysis was varied from 6 to 100. For each combination of N and number of rs, they replicated the meta-analysis 1,000 times and then evaluated the average value of S_e^2/S_r^2 across 1,000 meta-analyses. That is, they focused their attention on the average ratio of variance predicted from the sampling error formula to the average observed variance of the rs across studies. They did not look at $S_r^2 - S_e^2$, the difference between predicted and observed variances. They found that for all numbers of rs less than 100, the ratio S_e^2/S_r^2 averaged greater than 1.00. For example, when there were 10 rs per meta-analysis and $N = 75$ in each study, the average ratio was 1.25. The smaller the number of rs per meta-analysis, the more the ratio exceeded 1.00. Kemery, Mossholder, and Roth (1987) obtained similar results in their simulation study. Both studies interpreted these figures as demonstrating that the formula for S_e^2 overestimates sampling variance when the number of correlations in a meta-analysis is less than 100. Their assumption was that if the S_e^2 formula were accurate, the ratio S_e^2/S_r^2 would average 1.00.

The conclusion that one of the most basic formulas in all of statistics—a formula that had been accepted by statisticians for more than 80 years—was erroneous was surprising. The Spector and Levine (1987) study was critiqued by Callender and Osburn (1988), who showed that if one assessed accuracy by the difference $S_r^2 - S_e^2$, the sampling error variance formula was shown to be extremely accurate, as had also been demonstrated in their previous simulation studies. There was no bias. They also demonstrated why the average ratio S_e^2/S_r^2 is greater than 1.00 despite the fact that S_e^2 accurately estimates sampling variance. When the number of correlations in a meta-analysis is small, then by chance the S_r^2 will sometimes be very small; that is, by chance all observed rs will be very similar to each other. Because S_r^2 is the denominator of the ratio, these tiny S_r^2 values lead to very large values for S_e^2/S_r^2, sometimes as large as 30 or more, and if S_r^2 should by chance be zero, the ratio is infinitely large. These extreme values pull the mean ratio above 1.00; the median ratio is very close to 1.00. The analysis by Callender and Osburn (1988) explains the surprising conclusion of Spector and Levine (1987) and demonstrates that the fundamental sampling variance formula for the correlation is unbiased.

Spector and Levine (1987) would have reached a correct conclusion had they used the reciprocal of their ratio. That is, if they had used S_r^2/S_e^2 instead of S_e^2/S_r^2, they would have found that the mean ratio was 1.00. With this reversed ratio, the most extreme possible value is zero (rather than infinity), and the distribution of ratios is much less skewed. This point has important implications for second-order meta-analyses. Second-order meta-analyses may be conducted by averaging the percentage of variance accounted for by artifacts over similar meta-analyses. In any given meta-analysis, this percentage is the ratio of artifact-predicted variance (sampling variance plus that due to other artifacts) to the observed variance. One over this ratio is the reversed ratio, S_r^2/S_e^2. In second-order meta-analysis, this reversed ratio should be averaged across studies, and then the reciprocal of that average should be taken. Another way to state this is that, in second-order meta-analysis, the harmonic mean should be computed, not the arithmetic mean.

This procedure prevents the upward bias that appeared in the Spector and Levine study and results in an unbiased estimate of the average percentage of variance in the meta-analyses that is due to artifacts (for an example, see Rothstein et al., 1990).

Method

Database

The large database for validities against measures of performance on the job in clerical occupations compiled by Pearlman et al. (1980; see their Tables 4, 5, and 7) was used in the present study. Studies in this database are highly heterogeneous, spanning the time period from the 1920s through the 1970s and including both published and unpublished data (68% unpublished). Studies were conducted in all parts of the United States and in many different types of organizations, both public and private. This database included 32 distributions of validities representing various combinations of 10 test types and five individual (true) clerical job families, as defined by the *Dictionary of Occupational Titles* (U.S. Department of Labor, 1977). These 32 distributions of validities were the data used in this study. To ensure homogeneity of criterion measures, only studies in which the criterion measure was supervisory ratings of overall job performance were used. The number of validity coefficients in each distribution and their associated total sample sizes are shown in columns 1 and 2 of Table 1. The 10 test types in Table 1 can be divided into 5 classic ability constructs (Predictors 1–5: general mental ability, verbal ability, quantitative ability, reasoning ability, and perceptual speed) and 5 test types that would generally be regarded as less homogeneous (Predictors 6–10: memory tests, spatial/mechanical tests, motor tests, performance tests, and clerical aptitude tests). In this research, analyses were conducted separately for the classical ability constructs, in the expectation that validity findings would prove more homogeneous for these constructs.

Analysis

The published validity generalization analyses by Pearlman et al. were based on the noninteractive method (Pearlman et al., 1980), whereas this study used the interactive method (Hunter & Schmidt, 1990b, chap. 4; Schmidt, Gast-Rosenberg, & Hunter, 1980), the method now typically used in our validity generalization research. A recent computer simulation study (Law et al. 1991b) indicates that, under most conditions, the interactive procedure can be expected to be slightly more accurate than the noninteractive. The artifact distributions used were the same as those used by Pearlman et al. (1980). These distributions were carefully developed based on examination of numerous validity studies and data sets in which predictor and criterion reliabilities and range restriction values were either given or could be computed. They were later checked against figures from large empirical databases, yielding considerable empirical evidence that these distributions closely match those found in real data (Alexander, Carson, Alliger, & Cronshaw, 1989; Rothstein, 1990; Schmidt, Hunter, et al., 1985, question and answer 26; see also Hunter & Schmidt, 1990b, pp. 224–226).

Meta-analysis was applied four times to the clerical data set. The first analysis (Procedure 1) used the traditional method: the older linear range correction method was used and the observed study correlations (r_i) were used in the sampling error variance formula for r. No data were removed. In the second meta-analysis (Procedure 2), the new range correction procedure was introduced; otherwise this analysis was the same as Procedure 1. The third meta-analysis (Procedure 3) was the same as the second except that the $\bar{r}$ was used in the sampling variance formula instead of the observed study r_i. The fourth meta-analysis (Procedure 4) was the same as the third except that non-Pearson

Table 1
Total Sample Sizes and Number of Validity Coefficients in Original Distributions and After Removal of Non–Pearson Correlation Coefficients

Test type/ job family[a]	Original distribution		Non–Pearson rs removed	
	N	No. of rs	N	No. of rs
General mental ability				
A	3,683	57	3,462	52
B	3,510	32	3,263	29
C	289	7	289	7
Verbal ability				
A	15,056	168	13,565	147
B	5,321	71	4,101	60
C	1,303	27	1,193	25
Quantitative ability				
A	11,584	123	10,582	107
B	6,810	84	6,003	75
C	1,259	30	1,098	25
E	1,271	17	1,236	16
Reasoning ability				
A	2,794	34	2,794	34
B	911	21	834	20
C	739	10	739	10
Perceptual speed				
A	23,100	285	20,963	251
B	11,435	146	9,581	127
C	2,775	44	2,404	38
D	878	10	556	8
E	1,665	23	1,630	22
Memory				
A	2,421	33	2,393	32
B	1,486	27	1,486	27
C	726	11	709	10
Spatial and mechanical ability				
A	2,724	18	2,724	18
B	2,694	26	2,152	20
C	537	12	405	11
Motor ability[b]				
A	5,081	56	5,081	56
B	6,046	50	5,767	47
C	1,511	21	1,511	21
D	834	12	834	12
E	1,360	21	1,325	20
Performance tests[c]				
A	3,748	45	3,422	40
Clerical aptitude tests[d]				
A	4,062	60	3,893	56
B	1,354	21	1,303	20

[a] A = stenography, typing, filing, and related occupations (*Dictionary of Occupational Titles* [DOT] Occupational Groups 201–209); B = computing and account-recording occupations (DOT Occupational Groups 210–219); C = production and stock clerks and related occupations (DOT Occupational Groups 221–229); D = information and message distribution occupations (DOT Occupational Groups 230–239); and E = public contact and clerical service occupations (DOT Occupational Groups 240–248). [b] Finger, hand, and arm dexterity tests and motor coordination tests. [c] Typing, dictation, and similar clerical performance tests. [d] Tests consist of verbal, quantitative, and perceptual speed components.

correlations were removed prior to the analysis. That is, we removed all biserial, triserial, and tetrachoric correlations and left the ordinary Pearson *r*s, phi coefficients, point-biserial coefficients, and rhos, all of which are forms of the Pearson correlation coefficient. It was observed that the non-Pearson *r*s were frequently unusually large or unusually small, as would be expected from their larger sampling errors. Although not reported here, the removal of non–Pearson *r*s had essentially no effect on mean validities. The number of coefficients remaining in each distribution and associated remaining total sample sizes are presented in columns 3 and 4 of Table 1. Introduction of the improved range correction procedure in Procedure 2 does not change the figures for percentage variance accounted for, because these figures are calculated using the residual variance prior to its transformation to estimate SD_ρ. But the new range correction procedure should affect (reduce) estimates of SD_ρ. These four meta-analysis procedures were applied to each of the 32 validity distributions described above for the clerical job families, for a total of 128 meta-analyses. The methods of second-order meta-analysis described above were used to summarize the findings. For each of the four meta-analyses procedures examined, the harmonic mean of percentage variance accounted for and the arithmetic mean of the SD_ρ estimates were computed across the 32 validity distributions for the all-predictor analysis. This same procedure was followed for the analysis of the five classic constructs; the difference was that there were only 18 validity distributions for the classic constructs. This procedure weights all first-order meta-analyses entering the second-order meta-analysis equally, as has traditionally been done. However, we also examined a second weighting method: We computed the appropriate averages of the means for each construct. (These means are shown in Table 2.) This method weights each job family equally in determining the mean for each predictor, and then weights each predictor equally in determining the overall means. This weighting procedure yielded very similar results and identical conclusions.

Results

Complete results for each individual distribution of validities are presented in Table 2. Results are presented as averages in Tables 3 and 4. Table 3 presents the mean percentage of the variance of observed validities accounted for by artifacts, averaged across job families. These figures were computed using reciprocals, as described earlier. Looking first at the results for all predictor types, the mean percentage is 70% using the original interactive meta-analysis procedure with no refinements (Procedure 1). As expected, introduction of the new range correction procedure (Procedure 2) does not affect this percentage. In Procedure 3, when $\bar{r}$ instead of r_i from the individual studies is used in the formula for sampling error variance, the mean percentage variance accounted for rises to 74%. When, in addition, non–Pearson *r*s are removed (Procedure 4), the mean percentage rises to 82%. The change from 70% to 82% is a 17% increase. These figures are consistently larger for the five classic constructs. In Procedures 1 and 2, an average of 75% of the observed validity variance is accounted for by the classic constructs. Use of $\bar{r}$ in the sampling error variance formula increases this figure to 79% (Procedure 3). Finally, when non-Pearson *r*s are removed (Procedure 4), the mean percentage variance accounted for rises to 87%. The change from 75% to 87% is a 16% increase. Thus, on average, only 13% of the observed variance is unaccounted for and therefore could possibly be due to moderators. However, as discussed below, some of this remaining variance is artifactual.

Table 4 presents the mean values for the standard deviations of true validity (SD_ρ). For the entire data set (all 10 predictors), the mean SD_ρ estimate declines from .150 to .106 (a 29% decrease) as successive refinements are added to the meta-analysis procedures. These figures are all slightly lower when only the five classic ability constructs are included. When all refinements are included, the mean SD_ρ estimate for the five classic constructs is only .097. Introducing successive data-analytic refinements causes the mean SD_ρ estimate to decline from .142 to .097, a 32% reduction. Thus the prediction that these procedural refinements would decrease estimates of SD_ρ is confirmed.

Discussion

In summary, the most accurate available estimates of the average percentage variance accounted for are 82% for all 10 predictor types and 87% for the five classic constructs. The corresponding most accurate available estimates for mean SD_ρ are .106 and .097, respectively. As predicted, improvements in the accuracy of data-analytic methods substantially increased the validity variance accounted for and substantially reduced SD_ρ estimates. However, it might appear that if there were indeed no situational true validity variance at all, the mean percentage variance accounted for would be 100% and the mean SD_ρ estimate would be zero. Yet this is logically not the case, because there remain a number of variance-producing artifacts that the refined procedures examined in this study cannot correct for. First, there are likely to be numerically erroneous validity coefficients (bad data), due to errors in the original data, computational errors, and transcription and other clerical errors. Some, but not all, of these data errors would be expected to be outliers. The use of least squares statistical methods to estimate the mean and variance of the distribution of correlations (or any distribution) is based on the assumption that the data contain no aberrant values (i.e., outliers). When this assumption does not hold, the statistically optimal properties (efficiency and unbiasedness) of least squares estimates do not hold. Under these circumstances, least squares estimates become inaccurate because of their extreme sensitivity to outliers (Huber, 1980; Tukey, 1960; see also Barnett & Lewis, 1978; Grubbs, 1969). The presence of even a single outlier can produce a substantial increase in the observed standard deviation and a somewhat smaller distortion (in either direction) of the mean. Data sets in any research area are likely to contain data points that are erroneous due to computational, transcriptional, and other errors (Gulliksen, 1986; Wolins, 1962).

On the basis of his extensive experience with data sets of all kinds, Tukey (1960) judged that virtually all data sets contain outliers and other errors. A well-known psychometrician recently expressed the following statement:

> I believe that it is essential to check the data for errors before running my computations. I always wrote an error-checking program and ran the data through it before computing. I find it very interesting that in every set of data I have run, either for myself or someone else, there have always been errors, necessitating going back to the questionnaires and repunching some cards, or perhaps discarding some subjects. (Gulliksen, 1986, p. 4)

Unfortunately, the failure to conduct such checks when conducting primary studies is very widespread. In the physical

Table 2
Results for Individual Validity Distributions: Percentage of Variance Accounted for by Artifacts and SD_ρ Estimates Produced by Different Validity Generalization Estimation Procedures

	% variance accounted for				SD_ρ			
Test type/job family[a]	Proc. 1	Proc. 2	Proc. 3	Proc. 4	Proc. 1	Proc. 2	Proc. 3	Proc. 4
General mental ability								
A	65	65	71	76	.20	.18	.16	.14
B	52	52	54	58	.22	.19	.19	.17
C	377	377	377	377	.00	.00	.00	.00
M	81	81	85	91	.14	.12	.12	.10
Verbal ability								
A	49	49	51	49	.24	.22	.21	.22
B	49	49	52	61	.26	.23	.23	.20
C	119	119	123	131	.00	.00	.00	.00
M	61	61	64	68	.17	.15	.15	.14
Quantitative ability								
A	86	86	88	89	.10	.08	.08	.07
B	83	83	85	111	.12	.10	.10	.00
C	83	83	86	80	.16	.14	.12	.15
E	233	233	233	220	.00	.00	.00	.00
M	100	100	102	107	.09	.08	.07	.05
Reasoning ability								
A	94	94	98	98	.07	.07	.04	.04
B	74	74	78	91	.20	.18	.16	.11
C	62	62	68	68	.22	.20	.19	.19
M	75	75	80	84	.16	.15	.13	.11
Perceptual speed								
A	53	53	56	56	.23	.21	.20	.20
B	80	80	83	120	.13	.11	.10	.00
C	110	110	117	157	.00	.00	.00	.00
D	51	51	52	84	.26	.23	.23	.15
E	77	77	80	86	.15	.14	.13	.11
M	69	69	71	89	.15	.14	.13	.09
Memory								
A	81	81	87	91	.13	.12	.10	.08
B	130	130	136	136	.00	.00	.00	.00
C	71	71	77	73	.19	.17	.16	.17
M	88	88	94	94	.11	.10	.09	.08
Spatial and mechanical ability								
A	83	83	83	83	.10	.09	.09	.09
B	49	49	52	71	.23	.21	.20	.14
C	89	89	96	263	.15	.13	.10	.00
M	69	69	72	101	.16	.14	.13	.08
Motor ability[b]								
A	62	62	63	63	.19	.17	.17	.17
B	78	78	82	80	.11	.11	.10	.10
C	109	109	115	115	.00	.00	.00	.00
D	55	55	58	58	.26	.24	.24	.24
E	116	116	120	119	.00	.00	.00	.00
M	77	77	80	79	.11	.10	.10	.10
Performance tests[c]								
A	23	23	26	33	.43	.37	.36	.32
Clerical aptitude tests[d]								
A	64	64	69	79	.20	.18	.16	.13
B	58	58	61	59	.24	.20	.20	.21
M	61	61	65	67	.22	.19	.18	.17

Note. Proc. = procedure; Proc. 1 = r_i used in sampling variance formula and no data removed; Proc. 2 = new range correction procedure used, otherwise same as Proc. 1; Proc. 3 = $\bar{r}$ used in sampling variance formula and new range correction procedure used; Proc. 4 = non–Pearson correlations removed, otherwise same as Proc. 3.

[a] A = stenography, typing, filing, and related occupations (*Dictionary of Occupational Titles* [DOT] Occupational Groups 201–209); B = computing and account-recording occupations (DOT Occupational Groups 210–219); C = production and stock clerks and related occupations (DOT Occupational Groups 221–229); D = information and message distribution occupations (DOT Occupational Groups 230–239); and E = public contact and clerical service occupations (DOT Occupational Groups 240–248).

[b] Finger, hand, and arm dexterity tests and motor coordination tests.

[c] Typing, dictation, and similar clerical performance tests.

[d] Tests consisting of verbal, quantitative, and perceptual speed components.

Table 3
Mean Values of Percentage of Variance Accounted for (Averaged Across Job Families) for Different Validity Generalization Estimation Procedures

Procedure	Mean % variance accounted for
All 10 predictor types	
1: No data removed and r_i used in S_e^2 formula	70
2: New range correction procedure used	70
3: Same as in Procedure 2, but $\bar{r}$ used in S_e^2 formula	74
4: Same as in Procedure 3, but non-Pearson correlations removed	82
Five classic constructs only	
1: No data removed and r_i used in S_e^2 formula	75
2: New range correction procedure used	75
3: Same as in Procedure 2, but $\bar{r}$ used in S_e^2 formula	79
4: Same as in Procedure 3, but non-Pearson correlations removed	87

sciences (e.g., physics and chemistry) extreme values are routinely eliminated. For example, Hedges (1987) found that in the area of particle physics, roughly 40% of the available studies are omitted from meta-analysis for one reason or another. In the social sciences, it is rare for even the 10% of studies with the most extreme findings to be discarded. The psychological and social sciences have recently begun to recognize the need for such "trimming" prior to data analysis. Tukey (1960) and Huber (1980) recommended deletion of the most extreme 10% of data points—the largest 5% and the smallest 5% of values. Because most validity generalization analyses have been conducted using studies of imperfect methodological quality, the presence of outliers is highly probable. Deletion of outliers does not remove all numerically erroneous coefficients; outlier procedures detect and remove only data errors that are large enough to produce extremely deviant values. Thus data errors not detected remain to produce validity variance above that predicted by the sampling variance formula and other artifact corrections. Identification of true outliers is a somewhat complicated process, because when sample sizes are small, extreme values can occur simply because of large sampling errors. Such values are not true outliers. We are currently developing procedures for identifying outliers in meta-analytic and validity generalization data.

The second reason for not expecting all observed validity variance to be accounted for in validity generalization studies is the fact that the correction for sampling error variance is an undercorrection if there is range restriction. The formula for sampling error variance assumes that the independent and dependent variables are at least approximately normally distributed. Where there is direct range restriction (truncation) on one or both variables, this assumption is violated. In personnel selection, there may be direct restriction on the test (the independent variable). For example, job offers may be made only to those applicants above the mean test score. Millsap (1989), using computer simulation studies, found that under such conditions the sample (or study) correlations have larger sampling error variances than indicated by the sampling error variance formula. That is, the formula underestimates the true amount of sampling error variance, leading to undercorrection for sampling variance and, therefore, overestimation of the residual variance and SD_ρ. The undercorrection is largest when sample sizes are 60 or less. As an example, if $N = 60$ and $\rho = .40$ in all studies, and all variance is in fact due only to sampling error, then the estimated residual SD (SD_{res}) will on average be .046 across the levels of range restriction studied by Millsap (1989). The estimated SD_ρ value will typically be about .09, close to the mean value of .097 obtained here for the five classical constructs. The correct value for both SD_{res} and SD_ρ is, of course, zero. Thus, many nonzero estimates of SD_ρ in the published literature could be due in large part to this effect. In many studies, range restriction is indirect rather than direct. For example, employees may be selected based on a different test that might correlate, for example, .60 to .70 with the independent variable test. Studies are needed to determine what the effect of such indirect range restriction is on the accuracy of the formula for sampling error variance. This question was not addressed by Millsap (1989).

Third, observed validities may vary depending on the nature of the job performance ratings. Some studies used ratings of job performance that had earlier been made for administrative purposes (e.g., pay raises, promotions), whereas other studies were based on special ratings that were used solely for research purposes in that study. Administrative ratings are known to be strongly influenced by nonperformance considerations and to yield smaller observed correlations with selection procedures than research ratings (Whetzel, McDaniel, & Schmidt, 1985). This difference is a source of artifactual variance in the observed correlations that could not be corrected for, causing SD_ρ to be an overestimate.

Fourth, the inclusion of phi coefficients and point-biseral coefficients contributes to artifactual variance (Hunter & Schmidt, 1990a). As noted earlier, such validity coefficients were retained in the present meta-analyses because they are forms of the Pearson correlation coefficient (and also because

Table 4
Mean Values of Estimates of the Standard Deviation of True Validities for Different Validity Generalization Estimation Procedures (Averaged Across Job Families)

Procedure	SD_ρ
All 10 predictor types	
1: No data removed and r_i used in S_e^2 formula	.150
2: New range correction procedure used	.133
3: Same as in Procedure 2, but $\bar{r}$ used in S_e^2 formula	.126
4: Same as in Procedure 3, but non-Pearson correlations removed	.106
Five classic constructs only	
1: No data removed and r_i used in S_e^2 formula	.142
2: New range correction procedure used	.127
3: Same as in Procedure 2, but $\bar{r}$ used in S_e^2 formula	.119
4: Same as in Procedure 3, but non-Pearson correlations removed	.097

their removal would have caused substantial data loss). As forms of the Pearson *r*, their sampling error variances are accurately estimated by the standard formula. However, these coefficients are produced by dichotomizing continuous measures of either job performance or test score or both. The point-biserial results when only one of these measures is dichotomized; the phi coefficients results when both are dichotomized. Dichotomization reduces correlations in comparison to what they would have been in the absence of dichotomization. The correlations that are artificially reduced thus differ from those that are not. The effect of this is to create variation in correlations beyond that attributable to sampling error or any of the other artifacts that are corrected for (Hunter & Schmidt, 1990a). As a result, the final effect is an artifactual reduction in the percentage variance accounted for and an artifactual inflation of estimates of SD_ρ.

The fifth factor causing SD_ρ to be overestimated is inclusion of two or more correlations from the same studies whenever the study contained two different tests measuring the same ability in the same sample (e.g., two different tests measuring spatial ability). These correlations are not independent, and the result is inflation of both the observed SD (SD_r) and SD_ρ (Hunter & Schmidt, 1990b, chap. 10).

Sixth, it has been shown that differences between employees in amount of job experience reduce the observed validities of employment tests (McDaniel, Schmidt, & Hunter, 1988; Schmidt, Hunter, Outerbridge, & Trattner, 1986). Thus, studies in which employees vary widely in job experience can be expected to report smaller correlations on average than studies in which employees vary little in time on the job. The result is additional variation in correlations across studies that is not corrected for. Again, the effect is inflate the estimate of SD_ρ.

The six considerations discussed here make it clear that it is important to always bear in mind that all estimates of SD_ρ, in this study and in all others, are likely to be overestimates. Even after a meta-analysis is completed, there is still less real variation across studies than there appears to be.

In view of the fact that these six sources of uncorrected artifactual variance exist in the present data (as in many other data sets), our obtained results appear to be fully consistent with the hypothesis that there is no real (i.e., nonartifactual) variance in true validities. Thus, the findings of this study are further evidence against the situational specificity hypothesis. Of course, the findings in this study are also consistent with the conclusion that some situational variance in true validities does exist but that the amount is extremely small. We regard this as substantively the same hypothesis, given that both hypothesis have essentially identical theoretical and practical implications. Even with large data sets, it is very difficult to distinguish definitively between these two very similar hypotheses. However, the cumulative pattern of findings from the present meta-analyses, taken together with the five previous meta-analyses discussed earlier, provides strong support for the hypothesis that there is essentially no situational variance in true validities for classic ability constructs used for selection on similar jobs.

References

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

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data.* New York: Wiley.

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

Callender, J. C., & Osburn, H. G. (1988). Unbiased estimation of the sampling variance of correlations. *Journal of Applied Psychology, 73,* 312–315.

Dunnette, M. D., Houston, J. S., Hough, L. M., Touquam, J., Lamnstein, S., King, K., Bosshardt, M. J., & Keys, M. (1982). *Development and validation of an industry-wide electric power plant operator selection system.* Minneapolis, MN: Personnel Decisions Research Institute.

Dye, D. (1982). *Validity generalization analysis for data from 16 studies participating in a consortium study.* Unpublished manuscript, George Washington University, Department of Psychology, Washington, DC.

Grubbs, F. E. (1969). Procedures for detecting outliers. *Technometrics, 11,* 1–21.

Gulliksen, H. (1986). The increasing importance of mathematics in psychological research (Pt. 3). *The Score, 9,* 1–5.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science: The empirical cumulativeness of research. *American Psychologist, 42,* 443–455.

Huber, P. J. (1980). *Robust statistics.* New York: Wiley.

Hunter, J. E. (1991). *Improvements in accuracy from using the mean correlation in the sampling error formula in meta-analysis.* Unpublished manuscript, Michigan State University, Department of Psychology, East Lansing.

Hunter, J. E., & Schmidt, F. L. (1990a). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75,* 334–349.

Hunter, J. E., & Schmidt, F. L. (1990b). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1992). *The estimation of sampling error variance in the meta-analysis of correlations: Use of mean r* in the homogenous case. Manuscript submitted for publication.

Kemery, E. R., Mossholder, K. W., & Roth, L. (1987). The power of the Schmidt and Hunter additive model of validity generalization. *Journal of Applied Psychology, 72,* 30–37.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1991a). *A Monte Carlo test of two refinements in meta-analysis procedures.* Unpublished manuscript, University of Iowa, College of Business, Iowa City.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1991b). Nonlinearity of range corrections in meta-analysis: A test of an improved procedure. Manuscript submitted for publication.

Linn, R. L., & Dunbar, S. B. (1985). Validity generalization and predictive bias. In R. A. Burk (Ed.), *Performance assessment: State of the art.* Baltimore: Johns Hopkins University Press.

McDaniel, M. A., & Hirsh, H. R. (1986, April). Methods of moderator detection in meta-analysis. In M. A. McDaniel (Chair), *An overview and new directions in the Hunter-Schmidt-Jackson meta-analysis technique.* Symposium conducted at the Annual Conference of the Society for Industrial/Organizational Psychology, Chicago.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of the validity of training and experience ratings in personnel selection. *Personnel Psychology, 41,* 283–314.

Millsap, R. E. (1988). Sampling variance in attenuated correlation coefficients: A Monte Carlo study. *Journal of Applied Psychology, 73,* 316–319.

Millsap, R. E. (1989). Sampling variance in the correlation coefficient under range restriction: A Monte Carlo study. *Journal of Applied Psychology, 74,* 456–461.

Pearlman, K., Schmidt, F. L., & 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.

Peterson, N. G. (1982, October). *Investigation of validity generalization in clerical and technical/professional occupations in the insurance industry.* Paper presented at the Conference on Validity Generalization, Personnel Testing Council of Southern California, Newport Beach, CA.

Raju, N. S., & Burke, M. J. (1983). Two new procedures for studying validity generalization. *Journal of Applied Psychology, 68,* 382–395.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote with increasing opportunity to observe. *Journal of Applied Psychology, 75,* 322–327.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, P. P. (1990). Biographical data in employment selection: Can the validities be made generalizable? *Journal of Applied Psychology, 75,* 175–184.

Sackett, P. R., Schmitt, N., Tenopyr, M. L., Kehoe, J., & Zedeck, S. (1985). Commentary on "Forty questions about validity generalization and meta-analysis." *Personnel Psychology, 38,* 697–798.

Schmidt, F. L., Gast-Rosenberg, I. F., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology, 65,* 463–661.

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. (1984). A within-setting test of the situational specificity hypothesis in personnel selection. *Personnel Psychology, 37,* 317–326.

Schmidt, F. L., Hunter, J. E., Outerbridge, A. M., & Trattner, M. H. (1986). The economic impact of job selection methods on the size, productivity, and payroll costs of the federal work force: An empirical demonstration. *Personnel Psychology, 39,* 1–29.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1980). Task differences and the validity of aptitude tests in selection: A red herring. *Journal of Applied Psychology, 66,* 166–185.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38,* 697–798.

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,* 257–381.

Schmidt, F. L., Ocasio, B. P., Hillery, J. M., & Hunter, J. E. (1985). Further within-setting empirical tests of the situational specificity hypothesis in personnel selection. *Personnel Psychology, 38,* 509–524.

Society for Industrial and Organizational Psychology. (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.

Spector, P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology, 72,* 3–9.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, J. G. Ghurye, W. Hoeffding, W. G. Madoo, & H. B. Mann (Eds.), *Contributions to probability and statistics* (pp. 448–485). Stanford, CA: Stanford University Press.

U.S. Department of Labor. (1977). *Dictionary of occupational titles* (4th ed.). Washington, DC: U.S. Government Printing Office.

Whetzel, D. L., McDaniel, M. A., & Schmidt, F. L. (1985, August). The validity of employment interviews: A review and meta-analysis. In H. R. Hirsh (Chair), *Meta-analysis of alternative predictors of job performance.* Symposium conducted at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Wolins, L. (1962). Responsibility for raw data. *American Psychologist, 17,* 657–658.

Received April 1, 1991
Revision received June 8, 1992
Accepted June 15, 1992 ■