Journal of Management
1995, Vol. 21, No. 4, 813-831

# Why Meta-Analysis Doesn't Tell Us What the Data Really Mean: Distinguishing between Moderator Effects and Moderator Processes

Craig J. Russell
*Louisiana State University*

Stephen W. Gilliland
*Louisiana State University*

*Traditional approaches to detecting the presence of moderators in meta-analyses involve inferences drawn from the residual variance in criterion-related validities* ($r_i$) *after correcting for sampling error and statistical artifacts. James, Demaree, Mulaik, and Ladd (1992) argued that these residualized interpretations of meta-analytic results may be spurious when statistical artifacts covary with true moderators. We extend their model to suggest that situational moderators might also covary with sample size and content (i.e., nonrandom sample selection error), causing meta-analysis to be uninterpretable and a significant correlation between criterion-related validities and* $n_i$. *We investigate this possibility on studies examining criterion-related validities of peer nominations originally reported by Kane and Lawler (1978). Application of residualized meta-analysis suggests the presence of moderator effects, but a significant correlation between* $r_i$ *and* $n_i$ *precludes interpretation of the moderator process behind these effects. More generally, we argue that the nature of true contingencies cannot be inferred from meta-analytic summaries of traditional criterion-related validity studies. Primary research with appropriate controls is the only means of identifying true moderator effects* and *processes on criterion-related validity.*

Schmidt (1992) recently argued that inadequate research methods were the primary reasons for an apparent decline in the advance of psychological theory over the last 92 years. Statistical tests of significance used to analyze primary research data were criticized for placing too little emphasis on Type II error. When a number of small sample studies capture true underlying regularities in data, these true regularities may be hidden by variation caused by sampling

Direct all correspondence to: Craig J. Russell, Louisiana State University, Department of Management, Baton Rouge, LA 70803-6312.

error, and hence incorrect inferences about the true regularities are drawn (a Type II error). Further, Schmidt demonstrated how estimates of effect size in studies reporting "significant" findings may be distorted. Citing the potential for erroneous conclusions being drawn from tests of statistical significance in primary research studies, Schmidt (1992) questioned the merit of current research paradigms, suggesting that scientific discovery would be better served by de-emphasizing conclusions drawn from primary research studies, and emphasizing meta-analytic data summaries.

Meta-analyses of criterion-related validities in personnel selection typically correct distributions of criterion-related validities ($r_{xy}$) for variance attributable to sampling error and statistical artifacts (Schmidt & Hunter, 1990). Proponents of meta-analysis contend that these corrections provide an objective, methodologically rigorous means of advancing science by cumulating knowledge and reconciling findings across necessarily flawed primary research studies (Hunter & Schmidt, 1990). The meta-analytic model has claimed an elevated status over narrative compilations as the technique of choice for drawing inferences from cumulative research in many substantive areas (Guzzo, Jackson, & Katzell, 1987). Schmidt (1992) went so far as to posit an alternate research paradigm, where major original discoveries are made by mining the information found in existing research literatures and not by conducting original research (see McCall & Bobko, 1990; Mumford, Stokes & Owens, 1990, for different perspectives).

Nonetheless, a number of potential problems have been identified that prevent immediate adoption of Schmidt's (1992) views. Most of these concerns focus on the common "residualization" approach to meta-analysis (James, Demaree, Mulaik, & Ladd, 1992). Specifically, traditional approaches to meta-analysis involve inferences drawn from $\sigma_\rho^2$, the residual variance remaining in $r_i$ (the correlation between predictor X and criterion Y in Study I) observed across primary research studies after controlling for sampling error and statistical artifacts. If $\sigma_\rho^2$ is large, the possibility of some unknown moderator is said to exist. Conversely, if $\sigma_\rho^2$ is small, validity is said to generalize, i.e., there is an absence of moderators that might cause $r_i$ to vary from one situation to the next and $\bar{r}$ (corrected for statistical artifacts) is considered to be an accurate estimate of $\rho$.

Problems originally identified with this residualization approach focus on subjective interpretations of how small $\sigma_\rho^2$ had to be before "validity generalization" across situational moderators could be concluded (James, Demaree & Mulaik, 1986; James et al., 1988; Kemery, Mossholder & Roth, 1987; McDaniel, Hirsh, Schmidt, Raju & Hunter, 1986; Schmidt, Hunter & Raju, 1988; Schmitt, Gooding, Noe & Kirsch, 1984; Schmitt & Noe, 1986; Thomas, 1988). Hunter, Schmidt, and Jackson (1982) suggested that if the ratio of $r_i$ variance due to sampling error divided by total $r_i$ variance ($\sigma_e^2/\sigma_r^2$, where $\sigma_r^2 = \sigma_e^2 + \sigma_\rho^2$) is greater than or equal to 75%, validity generalization could be concluded. Recent concerns have examined how various circumstances might bias estimates of $\sigma_\rho^2$ (James et al., 1992). We examine the possibility that meaningful "moderators" might covary with nonrandom sampling error (i.e.,

sample selection error) and $r_i$ obtained in primary research studies. We first use two hypothetical data sets to describe how nonrandom sample selection error might impact meta-analysis results. We then demonstrate how nonrandom sample selection error impacts actual meta-analysis results using primary research studies reported by Kane and Lawler (1978). The possibility of nonrandom sampling error and the inability of meta-analysis to detect this error preclude interpretations of moderator processes in meta-analytic summaries.

## Consequences of Nonrandom Sample Selection Error in Meta-Analysis

James et al. (1992) mounted compelling logical arguments suggesting that a typically unmeasured moderator, organizational climate, covaries with true criterion-related validities ($\rho_{xy}$), criterion reliability, and range restriction. James et al. (1992) developed a model of these relationships presented in Figure 1a.

In this model, this situation moderator (M) influences both the true level of criterion validity ($\rho_k$) and statistical artifacts ($\alpha_k$ = criterion reliability; $\phi_k$ = predictor reliability; and, $\xi_k$ = range restriction in the predictor). The result is that when meta-analytic corrections are made for these artifacts, variance in $r_i$ due to true moderator effects is lost and estimates of $\sigma_\rho^2$ will be biased downward (see James et al., 1992, for a detailed explanation concerning why meta-analysis is not "meaningful or possible" [p. 9] under these conditions). Meta-analytic researchers may inappropriately conclude that situational moderators are not present and that validity generalizes (e.g., Schmidt, 1992).

We present an extension of the James et al. (1992) model incorporating covariation between nonrandom sampling error and situational moderators. James et al.'s (1992) approach did not include the possibility that variation in sample size and content, drawn from the "population" of field settings in which personnel selection validity studies could be conducted, is not random. Using Campbell and Stanley's (1963) convention, we will call this nonrandom sample selection error or "[b]iases resulting from differential *selection* of respondents" (p. 5, original emphasis retained). As in the case of covariation between moderators and statistical artifacts, residualized interpretations of meta-analytic results after correcting for sampling error may be incorrect. While James et al. (1992) suggested that situational moderators may go undetected in typical meta-analyses, we suggest that nonrandom sample selection error may cause erroneous situational moderators to be identified in typical meta-analyses. Figure 1b specifies this circumstance.

Specifically, Hunter and Schmidt (1990) assumed that one source of variation in observed criterion-related validities is what might be expected when multiple samples are randomly drawn from a single population. Sampling theory predicts that random events will cause statistics derived from the samples to vary around the population parameter of interest. The variation in observed criterion-related validity coefficients ($r_i$) around the true population correlation ($\rho_{xy}$) will increase as sample size decreases. $E(\sigma_\rho^2) = 0$ when there is only one population value of $\rho$, while $E(\sigma_\rho^2) > 0$ when a situational moderator causes

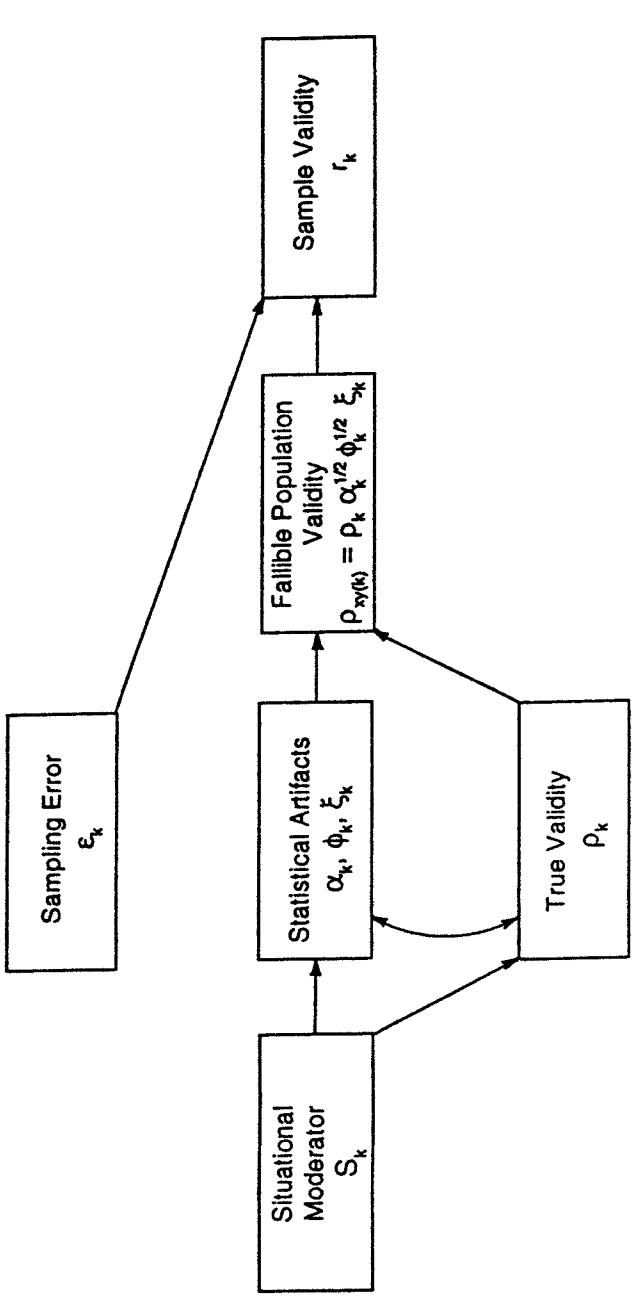


**Figure 1A.** James et al. (1992) model of moderator effects, statistical artifacts, and sampling error on criterion-related validity.

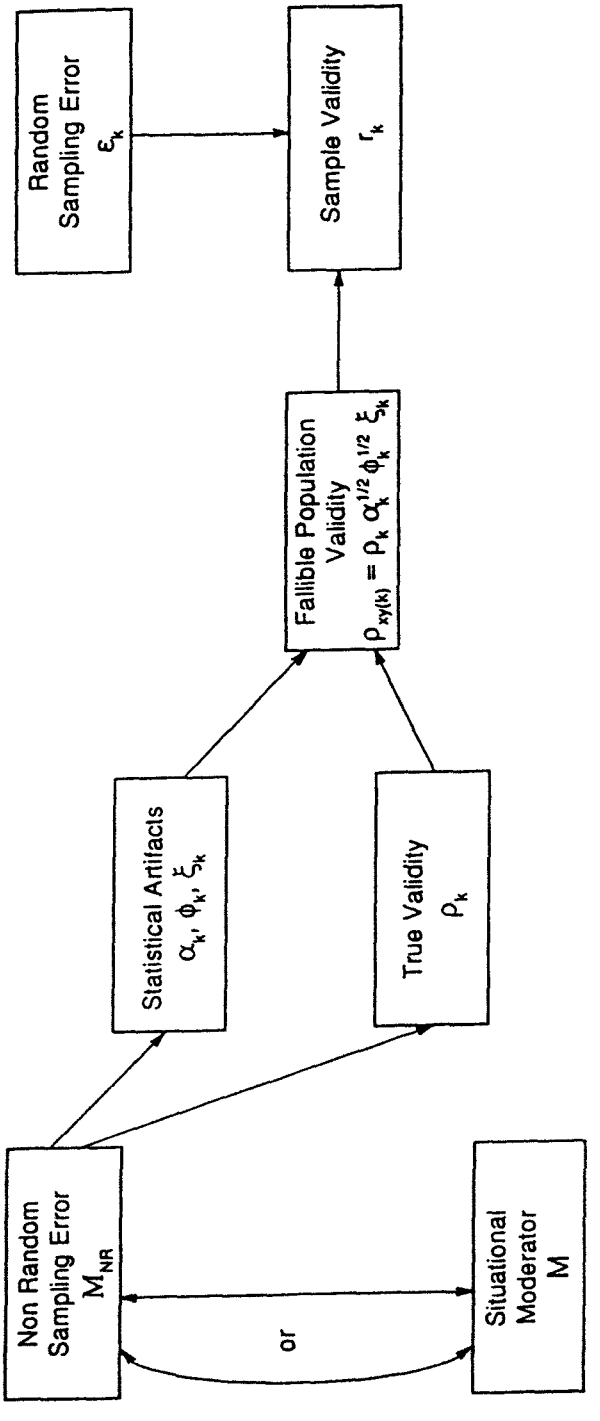


**Figure 1B.** Extension of James et al. (1992) model showing how a true moderator effect due to nonrandom sample selection error (risk aversion) is undetected when a spurious covariate (organizational climate) is the only moderator coded in a meta-analysis.

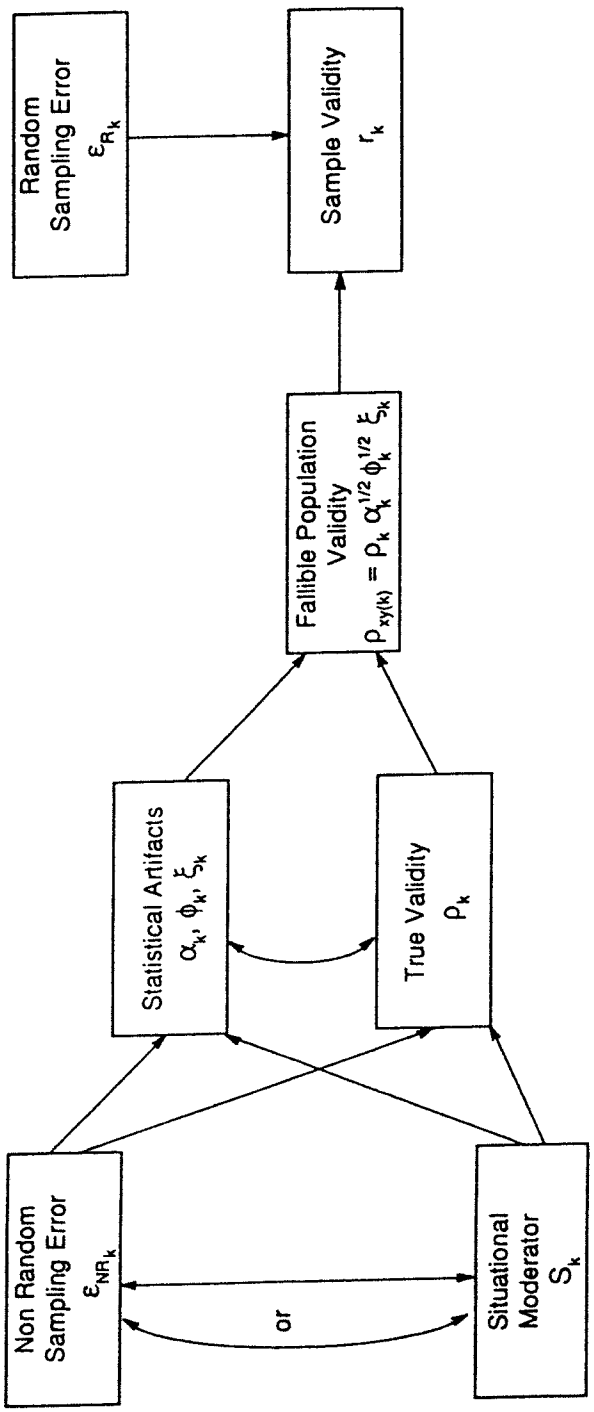


**Figure 1C.** Extension of James et al. (1992) model to situation where both nonrandom sample selection error (risk aversion) and traditional moderator effect (organization climate) simultaneously operate.

$\rho$ to be large for some moderator conditions and small for others. One moderator process previously unexamined in meta-analyses applied to the personnel selection literature is nonrandom sample selection error. We demonstrate below that nonrandom sample selection error may cause residualized interpretations of meta-analytic results to identify the wrong situational moderator. There are *at least* two ways in which nonrandom sample selection error can occur.

## Nonrandom Sample Sizes

First, nonrandom sample selection error occurs when a potential moderator variable is related to sample size. It is still assumed that observations within each sample were drawn at random, but the sample size was influenced by some moderator variable. This type of sample selection error leads to lower levels of precision in estimating $\rho$ for some levels of the moderator, for example, $\sigma_e^2$ is larger for some moderator conditions than others even though the moderator is not related to $r_i$. A hypothetical example of primary research findings that might yield this result is presented in Table 1. For consistency with James et al. (1992), we use organizational climate as a hypothetical moderator, though there are many possible variables that affect sample size in a study. We will assume managers in participative-organic organizations are less averse to taking risks (more likely to see the value of innovative social science applications) and cooperate with primary investigators, providing larger sample sizes than managerial subgroups who agree to participate from autocratic-mechanistic organizations (see James et al., 1992, for evidence supporting this possibility). Note that the data is configured so that variation around $\rho = .30$ is twice as large when $n_i = 100$ versus $n_i = 200$ (doubling the sample size is expected to yield half the sampling error variance).

The average effect size $\bar{r}$ is the same for both autocratic-mechanistic and participative-organic organizations in this hypothetical data. Further, if a meta-analysis had been conducted on all 20 observations, Hunter et al. (1982) would have concluded that there is no evidence of a moderator effect on $r_i$ because more than 75% of the observed variance in effect sizes ($\sigma_r^2$) was due to sampling error ($\sigma_e^2/\sigma_r^2 = 83.6\%$). As can be seen from results reported in Table 1, differences in the precision of the $\bar{r}$ estimate at each level of the moderator will cause estimates of $\sigma_e^2$ to vary across levels of the moderator, though a residualized interpretation of meta-analysis results using the 75% rule is correct (i.e., one value of $\rho$ is constant across levels of the moderator).[1] This example of nonrandom sampling error should pose no problem for traditional residualized meta-analytic interpretations: if $\bar{r}$ varies across levels of some true moderator, $\sigma_e^2/\sigma_r^2$ decreases in the total sample.

## Nonrandom Sample Size and Content

Sample selection may also be nonrandom if a moderator variable is related to both the number and type of observations included in a sample, for example, the moderator variable may cause samples to be selected of different sizes as

**Table 1.** Nonrandom Sample Sizes and Meta-analysis Results: $\rho = .30$

| $n_i$ | *Moderator* | $r_{xy}$ | *Meta-analysis Computations* |
|---|---|---|---|
| 100 | autocratic/mechanistic | .12 | Moderator = autocratic/mechanistic |
| 100 | autocratic/mechanistic | .16 | |
| 100 | autocratic/mechanistic | .20 | $\bar{r} = .30$, $\sigma_r^2 = .0132$ |
| 100 | autocratic/mechanistic | .24 | |
| 100 | autocratic/mechanistic | .28 | |
| 100 | autocratic/mechanistic | .32 | $\sigma_e^2 = [(1 - \bar{r}^2)^2 k/\Sigma n_i] = [(1 - .30^2)^2 10/1000] = .00828$; $\sigma_e^2 \div \sigma_r^2 = 62.7\%$ |
| 100 | autocratic/mechanistic | .36 | |
| 100 | autocratic/mechanistic | .40 | |
| 100 | autocratic/mechanistic | .44 | Moderator = participative/organic |
| 100 | autocratic/mechanistic | .48 | |
| | | | $\bar{r} = .30$, $\sigma_r^2 = .0033$ |
| 200 | participative/organic | .21 | |
| 200 | participative/organic | .23 | |
| 200 | participative/organic | .25 | $\sigma_e^2 = [(1 - .30^2)^2 10/2000] = .00414$; $\sigma_e^2 \div \sigma_r^2 = 125.5\%$ |
| 200 | participative/organic | .27 | |
| 200 | participative/organic | .29 | |
| 200 | participative/organic | .31 | Total ($k = 20$ studies) |
| 200 | participative/organic | .33 | |
| 200 | participative/organic | .35 | $\bar{r} = .30$, $\sigma_r^2 = .0066$ |
| 200 | participative/organic | .37 | |
| 200 | participative/organic | .39 | $\sigma_e^2 = [(1 - .30^2)^2 20/3000] = .005521$; $\sigma_e^2 \div \sigma_r^2 = 83.6\%$ |

well as content. This effect occurs when the observed relationship between X and Y is related to the level of the moderator $M_{NR}$; however, $M_{NR}$ effects the X-Y relationship only through its influence on which observations are likely to occur in a given sample.

For example, we are aware of a research project conducted on negative affectivity in which a large number of blank questionnaires were returned because certain questions relating to irregularities in sexual relations were deemed too personal by some respondents. Assume: (1) the purpose of this study was to examine the relationship between negative affectivity and job performance; (2) respondents who judged questions regarding their sex lives to be too personal were highly averse to taking risks (i.e., sharing personal information); and (3) risk averse people made more conservative choices on the job that resulted in fewer "big wins" and "big losses" in various aspects of job performance. In this case the moderator *process* occurs when the variable $M_{NR}$ acts to screen some subjects in and other subjects out of a given sample. This nonrandom (hence the subscript "NR" on variable $M_{NR}$) sample selection error would impact both sample size and content, resulting in two different $\bar{r}$ due to differences in the content of the samples on which they were derived. Individual differences in risk aversion (as a result of sampling bias) are the true cause of differences in $\bar{r}$.

This nonrandom sample selection effect is illustrated in Figure 2, where ellipse A contains data points associated with the high risk averse employees (who did not respond to the survey) and ellipse B contains data points for low risk averse employees (who did respond to the survey). Data points in ellipse A yield a substantially higher $r_i$ than data points in ellipse B.[2] Note that the range of X and Y values is exactly the same for both high and low risk averse individuals—we will also assume all other statistical artifacts (e.g., criterion and predictor reliability) are equal for both groups. That is, the differences in validities are not due to range restriction, predictor reliability, criterion reliability, or other artifacts which could be corrected, but rather are due to differences in the type of respondents. For any given level of negative affectivity (X), low risk averse employees exhibit a wider dispersion of true job performance (reflecting extremes occurring because of their "big wins" and "big losses").

When this nonrandom sample selection error exists in some studies but not others, $n_i$ across all studies will be related to the observed effect size, $r_i$. Specifically, negative affectivity studies asking questions about sexual relations might result in small samples with low risk averse subjects being selected and yield smaller $r_i$. Studies that measure negative affectivity without these questions will result in large samples that contain both high and low risk averse subjects and yield larger $r_i$. In traditional personnel selection settings, recruiting is designed to take "nonrandom" samples of the available labor pool in order to maximize applicant quality. Firms engaging in systematic recruiting efforts are more likely to sample from the upper tail of the labor pool performance distribution (i.e., the distribution of performance obtained if the entire labor pool had been hired). Firms engaging in less systemtatic or non-optimal

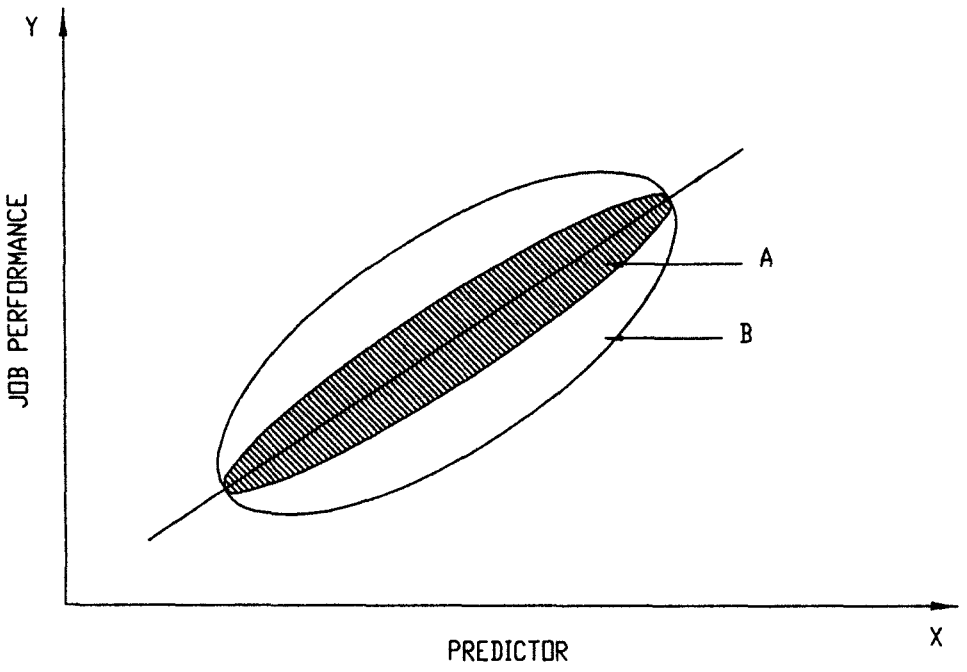


**Figure 2.** Graph of relationship between predictor and criterion for high risk averse (ellipse A) and low risk averse (ellipse B) individuals.

recruiting efforts will obtain a substantively different sample of the labor pool. Sample size cannot "cause" an effect, though clearly sample content can. Any observed relationship between $n_i$ and $r_i$ must be due to some moderator that influences or "biases" the number and type of observations contained in the sample, subsequently inflating or deflating $r_i$ (Berk, 1983).

A problem occurs when a moderator that causes nonrandom sample selection error is perfectly confounded with another variable, or what Campbell and Stanley (1963) labeled a "selection by treatment interaction" (p. 6). Simply put, if: (1) a researcher has measures of three variables (M, X, and Y); (2) a source of nonrandom sample selection error ($M_{NR}$) is the true cause of differences in rxy observed across studies; and (3) M is highly correlated with $M_{NR}$, then the researcher could falsely conclude that M is the true cause of an observed moderator effect. Note, this will be true for *any* inference drawn on the basis of $\sigma_\rho^2$, regardless of whether it is based on a decision heuristic (e.g., Hunter & Schmidt's 75% rule) or some statistical test (e.g., Hedges & Oklin's, 1985, chi-square tests).

To illustrate this circumstance, we use organizational climate (M) and risk aversion ($M_{NR}$) as hypothetical causes of a moderator effect. We expand our initial example to make the predictor X = peer assessments and the criterion Y = supervisor performance ratings in a single job occurring in a wide variety

of organizations. We will make three assumptions in this example. First, we assume that applicants who decide to apply and accept job offers (i.e., exercise self-selection) into autocratic-mechanistic organizations are more risk averse, preferring the comfort of a highly structured work environment. Similarly, we assume that applicants who self-select themselves in more participative-organic organizations desire less structure in their work and are less risk averse. For purposes of the example, it does not matter how the nonrandom sample selection occured—in actuality it is probably some combination of differential "approach" behavior by both the firm (i.e., variance in recruiting practices) and applicants (i.e., self selection). Second, we assume that two population values of $\rho$ characterize applicants to these positions, with a larger value for the high risk averse employees (who would comprise ellipse A in Figure 2). If risk aversion could be held constant across different organizational climates, $E(\sigma_\rho^2) = 0$. Third, as before we also assume that managers in participative-organic organizations will be more cooperative with the researcher (because they are also low risk averse) and provide greater access to their employees. Hence, in this example, participative-organic organizations attract low risk averse employees, resulting in lower criterion-related validities *and* larger sample sizes than autocratic-mechanistic organizations. Please note that this example and these assumptions are not based on prior theory but rather are offered to illustrate potential consequences of nonrandom sample selection error.

The hypothetical example of primary research findings presented in Table 2 suggest that a residualized interpretation of meta-analysis results could conclude that organizational climate moderates the X-Y relationship ($\sigma_e^2/\sigma_r^2 = 56.4\%$, which is less than the 75% rule proposed by Hunter et al., 1982). In fact, moderation due to organizational climate does not exist—instead, nonrandom sample selection error due to the self-selection of high and low risk averse people to these climates is causing the observed effect. Nonrandom sample selection error in this hypothetical data causes a correlation between $n_i$ and $r_i$ ($r_{n_i r_i} = .482$, $p \leq .05$). This is not a simple decrease in estimation precision for some levels of a moderator. Instead, it is due to differences in sample *size and content*, causing variation in effect size ($r_i$) due to the content of samples obtained from different organizational climates. Residualized meta-analysis would inappropriately conclude that organizational climate operates as a moderator, when applicant risk aversion is the true cause of differences in predictor-criterion relationships.

The findings presented in Table 2 highlight a subtle distinction between moderator *effects* and moderator *processes*. Meta-analytic results in Table 2 correctly demonstrate the presence of a moderator effect. However, the confounding of two moderator variables—"organizational climate" and nonrandom sample selection error caused by uneven representation of "risk aversion" across studies—could cause a meta-analytic researcher to draw incorrect conclusions about the moderator process. Meta-analytic implications for theory development and practice would be incorrect.

Nonrandom sample selection error as used herein is what Sackett, Tenopyr, Schmitt, and Kehoe (1985) called "first order" sampling error in that it occurs

**Table 2.** Nonrandom Sample Size, Sample Content, and Meta-analysis Results: $\rho_1 = .40$, $\rho_2 = .30$

| $n_i$ | *Moderator* | $r_{xy}$ | *Meta-analysis Computations* |
|---|---|---|---|
| 100 | autocratic/mechanistic | .22 | Moderator = autocratic/mechanistic |
| 100 | autocratic/mechanistic | .26 | |
| 100 | autocratic/mechanistic | .30 | $\bar{r} = .40$, $\sigma_r^2 = .0132$ |
| 100 | autocratic/mechanistic | .34 | |
| 100 | autocratic/mechanistic | .38 | $\sigma_e^2 = [(1 - \bar{r}^2)^2 k/\Sigma n_i] = [(1 - .40^2)^2 10/1000] = .007056$; $\sigma_e^2 \div \sigma_r^2 = 53.4\%$ |
| 100 | autocratic/mechanistic | .42 | |
| 100 | autocratic/mechanistic | .46 | |
| 100 | autocratic/mechanistic | .50 | |
| 100 | autocratic/mechanistic | .54 | Moderator = participative/organic |
| 100 | autocratic/mechanistic | .58 | |
| | | | $\bar{r} = .30$, $\sigma_r^2 = .0033$ |
| 200 | participative/organic | .21 | |
| 200 | participative/organic | .23 | |
| 200 | participative/organic | .25 | $\sigma_e^2 = [(1 - .30^2)^2 10/2000] = .00414$; $\sigma_e^2 \div \sigma_r^2 = 125.5\%$ |
| 200 | participative/organic | .27 | |
| 200 | participative/organic | .29 | Total ($k = 20$ studies) |
| 200 | participative/organic | .31 | |
| 200 | participative/organic | .33 | $\bar{r} = .35$, $\sigma_r^2 = .0091$ |
| 200 | participative/organic | .35 | |
| 200 | participative/organic | .37 | |
| 200 | participative/organic | .39 | $\sigma_e^2 = [(1 - .35^2)^2 20/3000] = .00513$; $\sigma_e^2 \div \sigma_r^2 = 56.4\%$ |

in original research studies where some "screen" operates to select certain types of subjects in or out of samples used in each study. First order sample selection error is captured in the traditional distinction between: (1) fixed effect, controlled experimental design with true random assignment of subjects; and (2) random effects, quasi-experimental designs that select subjects from naturally occurring groups in settings where people have *not* been randomly assigned to those groups. If each primary research study chosen for inclusion in a meta-analysis employed a design where levels and manipulations of the independent variable (predictor) were known and subjects from the population of interest were randomly assigned to each "treatment" group (i.e., predictor levels in a fixed effects experimental design), then meta-analytic partitioning of total variance into portions due to treatment and portions due to random sampling error ($\sigma_r^2 = \sigma_\rho^2 + \sigma_e^2$) could be considered the analogue of the traditional partitioning of sums of squares in analyses of variance ($\Sigma\Sigma(X_{ij} - \bar{X})^2 = \Sigma\Sigma(X_{ij} - \bar{X}_j)^2 + \Sigma n_i(\bar{X}_j - \bar{X})^2$) (cf. McClelland & Judd, 1993). For the hypothetical studies in Table 2, this would require *random assignment* of applicants to participative-organic versus autocratic-mechanistic firms *as well as* stratified random samples of high and low risk averse employees from both types of organizations.

Unfortunately, the experimental control required to achieve these sample characteristics is rare in personnel selection research. Typical field settings in personnel selection involve a certain amount of self selection by the subjects (i.e., to the applicant pool in predictive designs and to continued employment in concurrent designs) *as well as* other screening by firms' recruiting efforts or other environmental events (cf. Russell, Settoon, McGrath, Blanton, Kidwell, Lohrke, Scifries & Danforth, 1994). It seems highly probable that screening processes operate in field settings to cause subjects with certain latent predictor-criterion relationships ($\rho_{xy}$) to appear with greater or lesser frequency.

Consequently, at least three models portray how situational moderators might operate on $\bar{r}$ and $\sigma_\rho^2$ obtained using the Hunter et al. (1982) derivations. In Model I, situational moderators are unrelated to sample size and content. This is the implicitly assumed state in most meta-analytic research. In Model II, a situational moderator is related to sample size and *not* $r_i$. Hunter et al.'s (1982) residualization approach to meta-analysis should correctly conclude that a moderator does not impact $r_i$ (see Table 1). In Model III, a situational moderator may be related to both $n_i$ and $r_i$ (see Figure 1b). In this case, $n_i$ and $r_i$ are correlated due to the "bias" of the situational moderator. This bias causes traditional residualized meta-analysis interpretations to incorrectly conclude that one moderator process (e.g., involving organizational climate) exists when in fact another moderator process (e.g., involving risk aversion) is at work through nonrandom sample selection error.

Of course, combinations of these models might also exist, for example, moderators may exist that influence both the nature of the sample drawn (nonrandom sample selection error) and the nature of the X-Y relationship (Figure 1c). In the example described in Table 2, this would involve a combination of a moderator effect due to organizational climate and

nonrandom sample selection error due to subject self-selection into each climate. Regardless, investigators using the residualization approach to meta-analysis will never know whether they face Models I, II, or III unless they have determined whether $n_i$, $r_i$, and a moderator M are related.

The purpose of this study was to examine this extension of the James et al. (1992) model that permits moderators to covary with nonrandom sampling error. We cannot explore the nature of any existing sampling bias—that cannot be done when access is limited to archival summary statistics (e.g., $r_i$). Instead, we meta-analyze data originally compiled by Kane and Lawler (1978) on the validity of peer assessment, examining whether evidence of sample selection error exists (i.e., whether $n_i$ and $r_i$ are correlated) and what (if any) moderators might account for this bias.

## Method

### *Sample*

Analyses reported below use studies obtained from Table 1 originally reported in Kane and Lawler (1978). Fourteen studies using peer nominations as predictors and a variety of job performance criteria contained 61 criterion-related validities.

### *Procedure*

Validities of measures within single predictor and criterion categories were averaged to produce a "summary" validity coefficient within each independent sample reported in a study (as per Hunter et al.'s, 1982, p. 128, recommendation). Following Schmitt et al.'s (1984) procedures, no corrections in summary validities ($r_i$) were made for predictor/criterion reliability or range restriction.

### *Moderators*

Kane and Lawler (1978) reported information concerning four situational variables: type of sample (e.g., student, military, insurance agents, etc.), type of criterion, design (concurrent versus predictive), and purpose for conducting the study (administrative versus research). For the "type of sample" moderator, 69.6% of the $r_i$ were derived from military samples, 30.4% were derived from non-military samples. Given the nature of military versus non-military settings, one might label military samples as being characterized by a more autocratic-mechanistic climate relative to the non-military samples. Consistent with our earlier example in Table 2, one might expect $r_i$ to be negatively correlated with $n_i$ in the total data set and $\bar{r}$ to be higher for studies conducted on military samples. However, when a study is given approval in the military, superior officers' orders to participate tend to yield large sample sizes and high participation rates. Consequently, under these "military" autocratic-mechanistic conditions, one might expect $r_i$ to be positively correlated with $n_i$ in the total data set. In the absence of some a priori explanation of how a particular sample selection error operates, any number of logical arguments can be mounted to

explain either positive or negative correlations between $r_i$ and $n_i$ of any size. Clashing arguments can also be constructed to explain why the type of criterion used, purpose for which the study was conducted, and type of design might cause nonrandom sample selection error (cf. Russell et al., 1994).

Consequently, any number of a priori explanations could be constructed to describe how a moderator might cause nonrandom sample selection error that leads to $r_i$ being significantly correlated with $n_i$. While we still expect $r_i$ to be correlated with $n_i$ in the Kane and Lawler data, an argument can be made that $\bar{r}$ should be larger or smaller in a particular type of sample or with a particular purpose for conducting the study. Hence, we predict $r_i$ to be correlated with $n_i$ indicating the presence of sampling bias and support for Model III in the Kane and Lawler (1978) data. No clear prediction can be made for which moderator might be the cause of this sample selection error.

*Analyses*

Three analyses were conducted on the peer nomination data. First, a traditional residualized meta-analysis was conducted to estimate $\bar{r}$, $\hat{\sigma}^2_r$, $\hat{\sigma}^2_e$, and $\hat{\sigma}^2_\rho$. Second, a product moment correlation coefficient was derived for $r_i$, $n_i$ pairs. Third, traditional residualized meta-analyses were conducted in which $\bar{r}$, $\hat{\sigma}^2_r$, $\hat{\sigma}^2_e$, and $\hat{\sigma}^2_\rho$ were derived for each moderator level.

## Results

Table 3 contains results obtained from a residualized meta-analysis using procedures described by Hunter et al. (1982). The average criterion-related validity for peer nominations was $\bar{r} = .467$ (uncorrected for range restriction or measurement error) and the percentage of variance explained by random sampling error was $\hat{\sigma}^2_e/\hat{\sigma}^2_r = 4.0\%$, substantially below Hunter et al.'s (1982) 75% heuristic. Hence, Hunter et al.'s interpretation of initial residualized meta-analysis results would suggest the presence of one or more moderators.

The product moment correlation coefficients between $r_i$ and $n_i$ for peer nominations was $r = .49$, $p \leq .01$ ($N = 23$). Hence, strong evidence exists that nonrandom sample selection error has occurred in these studies. Meta-analysis results presented in Table 3 also suggest a number of potential moderation processes. Specifically, military samples yielded higher validities than nonmilitary samples ($\bar{r}$ is .19 higher) and studies conducted for administrative purposes yielded higher validities than those conducted for research reasons ($\bar{r}$ is .12 larger). Validities were much higher with a promotion criterion compared to training success or performance ratings. These moderators *may* be a source of nonrandom sample selection error that caused $r_i$ and $n_i$ to be significantly correlated (e.g., studies using peer nominations in military settings yield $\bar{n} = 495$ compared to $\bar{n} = 90$ in nonmilitary samples). If they are *not* the source of nonrandom sample selection error, some unknown variable is the true cause of the moderator effect.

**Table 3.** Meta-analyses of Kane and Lawler (1978) Peer Nomination Data

| | $k$ | $\Sigma n_i$ | $\bar{n}$ | $\bar{r}$ | $\sigma_r^2$ | $\sigma_e^2$ | $\sigma_\rho^2$ | *% Random Error* |
|---|---|---|---|---|---|---|---|---|
| All Studies | 23 | 8555 | 372 | .467 | .04122 | .00164 | .03962 | 4.0% |
| Military Samples | 16 | 7924 | 495 | .523 | .03145 | .00107 | .03039 | 29.5% |
| Non-military Samples | 7 | 631 | 90 | .339 | .01534 | .00869 | .00665 | 56.6% |
| Concurrent Designs | 5 | 471 | 94 | .524 | .02166 | .00479 | .01687 | 22.1% |
| Predictive Designs | 18 | 8084 | 449 | .451 | .04619 | .00141 | .04478 | 3.1% |
| Training Success | 6 | 1646 | 274 | .395 | .00347 | .00260 | .00087 | 74.9% |
| Job Performance Ratings | 10 | 1506 | 151 | .425 | .01363 | .00446 | .00917 | 32.7% |
| Promotion | 5 | 5152 | 1030 | .637 | .01473 | .00034 | .01439 | 2.3% |
| Research Reason | 15 | 7312 | 487 | .452 | .04842 | .00130 | .04712 | 2.7% |
| Administrative Reason | 5 | 470 | 94 | .574 | .03412 | .00478 | .02934 | 14.0% |

*Notes:* $\sigma_r^2 - \sigma_e^2 = \sigma_\rho^2$ may not be exact due to rounding.

## Discussion

Initial residualized meta-analysis results strongly suggested situational moderator variables operate when peer nominations are used as predictors. Investigators pursuing traditional meta-analysis procedures (Hunter et al., 1982) would have then examined whether $\hat{\sigma}_\rho^2$ substantially decreased and whether $\bar{r}$ 's were different when studies were grouped according to their level on some moderator. This was done for a number of moderators and differences in $\bar{r}$ were found.

Specifically, when M = "purpose for which the research was conducted," meaningful differences in $\bar{r}$ appeared for peer nominations. Research conducted for administrative reasons yielded higher average criterion-related validities (consistent with findings reported by Russell et al., 1994). For the "type of sample" moderator, military samples generated *higher* validities for peer nominations.

In spite of these findings, evidence of moderator effects cannot be interpreted without first examining $r_{n_i r_i}$. A significant product moment correlation between $r_i$ and $n_i$ for peer nominations casts doubt on the underlying nature of any observed moderators. Differences in $\bar{r}$ across military versus nonmilitary samples may be due to differences in the way samples were comprised in the military and nonmilitary settings (e.g., differences in subjects' risk aversion). At the same time, these differences may be due to the direct impact of some moderator on the true relationship between peer nominations and performance criteria. With access limited to archival data, one can only speculate concerning how these differences may be caused. Further, any one or more of these variables may be exerting more than one moderator *process*—operating simultaneously through nonrandom sample selection error (i.e., Model III) *and* changing the nature of the X-Y relationship (i.e., Model I).

Hence, while these residualized meta-analysis results indicate the presence of a moderator *effect*, the moderator *process* causing the effect is unknown. Evidence suggests Model I, Model III, or both will provide correct explanations for the moderator process, though meta-analysis partitioning of observed variance in $r_i$ obtained from these field settings cannot tell us which is most likely. Meta-analyses that base their examination of moderator effects on estimates of $\sigma_e^2$ and $\sigma_\rho^2$ (i.e., Hunter & Schmidt's 75% rule or Hedges & Olkins' chi-square test) can conclude that a moderator effect is occurring, though the process behind that effect cannot be determined.

Of perhaps greater concern is a way nonrandom sample selection error can occur *and* go undetected when conducting a meta-analyses and deriving $r_{n_i r_i}$ as described above. Specifically, our Model II described a situation where sample size covaries with levels of some moderator, while Model III described a situation where sample size and content covaried with levels of some moderator. What if a moderator operates to screen samples in a way that effects only sample content and not sample size? Because $E(r_{n_i r_i}) = 0$, traditional residualized meta-analysis would yield results suggesting the presence of a traditional "moderator" effect (i.e., $\bar{r}$ differ across levels of the moderator and

$\hat{\sigma}_e^2/\hat{\sigma}_r^2$ is small within moderator levels) when in fact these results are due to nonrandom sample selection error. This circumstance cannot be determined from meta-analyzing findings reported across original investigations. Campbell and Stanley (1963) maintained that *only* primary research using experimental or quasi-experimental designs with random assignment of subjects will address this problem. Most disconcerting, however, is the fact that without benefit of such controlled studies, this type of nonrandom sample selection error may be the true cause of any moderator effects reported in meta-analyses since the late 1970s.

## Conclusion

In sum, investigators conducting meta-analyses of primary research in which controls do not permit random selection cannot simply examine the residual variance left unexplained by random sampling error and statistical artifacts (i.e., $\sigma_\rho^2$ or $\sigma_e^2/\rho_r^2$) and draw theoretical implications about differences in criterion-related validity across situations. As is evident in the actual data presented in Table 3 and the hypothetical data presented in Table 2, residualized meta-analysis could conclude that the wrong moderator process is operating.

When reviewing existing research, meta-analytically derived estimates of $\bar{r}$ will provide accurate summaries. Indeed, meta-analysis should remain the preferred means of literature compilation relative to narrative reviews. However, we have demonstrated that Schmidt (1992) is not correct—meta-analysis is not a substitute for well designed primary research in theory development. Just as we cannot infer causation from correlational designs in primary research conducted in field settings, we cannot infer causal contingencies from meta-analyses of correlational field designs. Scientific discovery will be better served by not interpreting meta-analytic results in the absence of additional primary research that controls for nonrandom sample selection error and covariation between statistical artifacts and moderators. A shift away from current research paradigms is not called for, we simply need to do a better job of conducting programmatic primary research designed to sequentially eliminate competing (artifactual or nonartifactual) explanations.

**Acknowledgment**: We would like to thank A. Elyssa Blanton, Philip Bobko, Jose Cortina, Paul Jarley, and Randall Settoon for their comments.

## Notes

1. Technically, this variable is not a moderator since $\rho$ is a constant. However, we use the term "moderator" for simplicity and consistency in our discussion.
2. We will assume that X and Y are unstandardized. If X and Y were standardized, for both A and B to reflect the sample $r_i$, ellipse A would be need to be tilted slightly counter clockwise to account for the effects of regression toward the mean on the true underlying value of $\rho_A$.

## References

Campbell, D.T. & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Berk, R.A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review, 48*: 386-398.

Guzzo, R.A., Jackson, S.E. & Katzell, R.A. (1987). Meta-analysis. In B.M. Staw & L.L. Cummings (Eds.), *Research in organizational behavior* (pp. 407-442). Greenwich, CT: JAI Press.

Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

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

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

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

James, L.R., Demaree, R.G. & Mulaik, S.A. (1986). A note on validity generalization procedures. *Journal of Applied Psychology, 71*: 440-450.

James, L.R., Demaree, R.G., Mulaik, S.A. & Mumford, M.D. (1988). Validity generalization: Rejoinder to Schmidt, Hunter, and Raju (1988). *Journal of Applied Psychology, 73*: 673-678.

James, L.R., Demaree, R.G., Mulaik, S.A. & Ladd, R.T. (1992). Validity generalization in the context of situational models. *Journal of Applied Psychology, 77*: 3-14.

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

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*: 39-37.

McCall, M.W. Jr. & Bobko, P. (1990). Research methods in the service of discovery. Pp. 381-418 in M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology*, Vol. 1, 2nd ed. Palo Alto, CA: Consulting Psychologist Press.

McClelland, G.H. & Judd, C.M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114*: 376-390.

McDaniel, M.A., Hirsh, H.R., Schmidt, F.L., Raju, N.S. & Hunter, J.E. (1986). Interpreting the results of meta analytic research: A comment on Schmitt, Gooding, Noe, and Kirsch (1984). *Personnel Psychology, 39*: 141-148.

Mumford, M.D., Stokes, G.S. & Owens, W.A. (1990). *Patterns of life history: The ecology of human individuality*. Hillsdale, NJ: Erlbaum.

Russell, C.J., Settoon, R.P., McGrath, R., Blanton, E. A., Kidwell, R., Lohrke, F.T., Scifries, E.L. & Danforth, G.W. (1994). Investigator characteristics as moderators of personnel selection research: A meta-analysis. *Journal of Applied Psychology, 79*: 163-170.

Sackett, P.R., Tenopyr, M.L., Schmitt, N. & Kehoe, J. (1985). Commentary on forty questions about validity generalizations and meta-analysis. *Personnel Psychology, 38*: 697-798.

Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*: 1173-1181.

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. & Raju, N.S. (1988). Validity generalization and situational specificity: A second look at the 75% rule and Fisher's *z* transformation. *Journal of Applied Psychology, 73*: 665-672.

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

Schmitt, N. & Noe, R.A. (1986). On shifting standards for conclusions regarding validity generalization. *Personnel Psychology, 39*: 849-851.

Thomas, H. (1988). What is the interpretation of the validity generalization estimate. *Journal of Applied Psychology, 73*: 679-682.