Journal of Applied Psychology
1997, Vol. 82, No. 4, 528–538

0021-9010/97/$3.00

Sampling Variance in the Correlation Coefficient Under Indirect Range Restriction: Implications for Validity Generalization

Herman Aguinis and Roger Whitehead
University of Colorado at Denver

The authors conducted Monte Carlo simulations to investigate whether indirect range restriction (IRR) on 2 variables X and Y increases the sampling error variability in the correlation coefficient between them. The manipulated parameters were (a) IRR on X and Y (i.e., direct restriction on a third variable Z), (b) population correlations ρ_{XY}, ρ_{XZ}, and ρ_{YZ}, and (c) sample size. IRR increased the sampling error variance in r_{XY} to values as high as 8.50% larger than the analytically derived expected values. Thus, in the presence of IRR, validity generalization users need to make theory-based decisions to ascertain whether the effects of IRR are artifactual or caused by situational-specific moderating effects.

Meta-analysis constitutes a set of procedures used to quantitatively integrate a body of literature. Validity generalization (VG) is one of the most commonly used meta-analytic techniques in industrial and organizational (I&O) psychology, management, and numerous other disciplines (e.g., Hunter & Schmidt, 1990; Hunter, Schmidt, & Jackson, 1982; Mendoza & Reinhardt, 1991; Schmidt, 1992). For example, Hunter and Schmidt estimated that VG methods have been used in over 500 studies to investigate the relationships between preemployment tests and job performance measures. Because of its frequent implementation, VG has recently been characterized as one of the three major meta-analytic approaches (Johnson, Mullen, & Salas, 1995).

VG extends arguments from psychometric theory to assert that a substantial portion of the variability observed in an X–Y relationship across individual studies is the result of sources of variance not explicitly considered in a study design. Consequently, to better estimate an X–Y relationship in the population, researchers should (a) attempt to control the impact of these extraneous sources of variance by implementing sound research designs and (b) correct for the extraneous across-study variability by subtracting it from the total observed variance in study-level effect size estimates (Aguinis & Pierce, in press). There is consensus that these extraneous sources of variance are typically not explicitly considered in a study's design. However, there is an ongoing debate regarding which sources of variance are artifactual in nature and which are theoretically meaningful (James, Demaree, & Mulaik, 1986; James, Demaree, Mulaik, & Mumford, 1988).

The sources known to increase across-study variability in effect size estimates are (a) sampling error, (b) error of measurement in the dependent variable, (c) dichotomization of a continuous dependent variable, (d) dichotomization of a continuous independent variable, (e) range variation in the independent variable, (f) range variation in the dependent variable due to attrition artifacts, (g) deviation from perfect construct validity in the independent variable, (h) deviation from perfect construct validity in the dependent variable, (i) reporting or transcriptional error, and (j) variance due to extraneous factors (Hunter & Schmidt, 1990; Hunter et al., 1982; Schmidt et al., 1993). However, despite that these factors have been identified, their contribution to overall variability of rs across studies usually accounts for not more than approximately 80.00% to 90.00% of the total variance (e.g., Pearlman, Schmidt, & Hunter, 1980; Schmidt, Hunter, Pearlman, & Shane, 1979). In consequence, in recent investigations of VG methodology, researchers have hypothesized that (a)

Herman Aguinis, College of Business and Administration, University of Colorado at Denver; Roger Whitehead, Department of Psychology, University of Colorado at Denver. Both authors contributed equally to this study.

A preliminary version of this article was presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, Florida, in May 1995. We are grateful to Charles A. Pierce (Montana State University), Chockalingam Viswesvaran (Florida International University), and the members of the Behavioral Science Research Group for their helpful feedback on earlier drafts.

Correspondence concerning this article should be addressed to Herman Aguinis, College of Business and Administration, University of Colorado at Denver, Campus Box 165, P.O. Box 173364, Denver, Colorado 80217-3364. Electronic mail may be sent via Internet to haguinis@castle.cudenver.edu. Herman Aguinis's World Wide Web address is http://www.cudenver.edu/~haguinis.

already identified sources of extraneous variance may cause more variability than is recognized (Law, Schmidt, & Hunter, 1994b); and (b) there may be additional factors, not yet identified, which are possibly causing effect size estimates (e.g., rs) to erratically fluctuate across studies (Schmidt et al., 1993). As an example of the former, a recent Monte Carlo (MC) study conducted by Millsap (1989) demonstrated that the sampling error variance of rs across studies is typically much larger than is suspected. More specifically, the results of Millsap's simulation revealed that the sampling error variance of rs affected by direct range restriction is larger than is estimated by the traditional sampling error variance formula (i.e., $S_r^2 = [1 - r^2]^2/[N - 1]$). Consequently, the sampling error variance is underestimated when rs are affected by direct range restriction. Therefore, if direct range restriction is caused by artifacts and not by theoretically meaningful moderator variables, a conclusion may be reached erroneously regarding the existence of variability above and beyond sampling error, whereas in actuality, across-study variability of rs is caused by (artifactual) predictor variable range restriction. Stated differently, the unexpected increase in r variability across studies may artificially inflate Type I error rates regarding the null hypothesis of a no-moderating effect. Moreover, researchers may incorrectly conclude that validities vary across various specific contexts and situations and, thus, are not generalizable.

More recently, Schmidt et al. (1993) identified additional artifacts hypothesized to inflate the across-study variability of rs. One of the factors identified by Schmidt et al. is indirect range restriction (IRR). IRR is a common phenomenon in I&O psychological research. In personnel selection research, for example, applicants for a job are initially selected on the basis of a cognitive abilities test (predictor Z), such that only those with a score above cutoff score z are selected. Then, as part of a criterion-related validity study, the validity of a new Test X is evaluated as a predictor of job performance (Y) and X is administered to a sample of current employees (i.e., incumbents). Note that incumbents constitute a range-restricted sample because they have already been selected on the basis of Z scores. Thus, to the extent that Z scores are correlated with the new predictor (X) and with job performance (Y), direct selection on Z leads to IRR (also called implicit, induced, or incidental range restriction) on both X and Y. It deserves noting that direct range restriction (e.g., on Z) and, consequently, IRR (e.g., on X and Y) occur very frequently in contexts in which samples are selected from larger pools of applicants (e.g., educational and business organizations; Ghiselli, Campbell, & Zedeck, 1981, p. 295; Hunter & Schmidt, 1990, p. 209; Linn, 1983a, 1983b). Accordingly, Thorndike (1949) stated that range restriction, "imposed by indirect selection on the basis of some variable other than the ones being compared, appears by far the most common and most important one for any personnel selection research program" (p. 175).

Despite that IRR occurs frequently in I&O psychological research, especially in such research areas as personnel selection and validation, there is no empirical evidence to support Schmidt et al.'s (1993) contention that the variability of effect sizes is larger than is estimated using the traditional sampling error formula when rs are affected by IRR. In addition, if such an increase exists, there is a need to know its magnitude and practical significance regarding the implementation of VG procedures. Accordingly, the purpose of our study was to use an MC strategy (Hartley & Harris, 1963; Noreen, 1989; Rubinstein, 1981) to examine (a) whether IRR increases the sampling error variance in the correlation coefficient above its analytically expected value and (b) the extent to which the sampling error variance estimator used in VG studies underestimates sampling error variance in the presence of IRR.

The MC strategy was used because it allows researchers to overcome difficulties and complexities imposed by the concurrent manipulation of several parameters, which often make the investigation of sampling distributions mathematically difficult or even intractable.

Method

Overview

MC simulations were conducted following a method similar to that implemented by Millsap (1989). In the simulation, we generated trivariate (X, Y, Z) arrays from multivariate normal populations and assessed the impact of (a) severity of IRR on X and Y (i.e., SR, the selection ratio on Z), (b) size of sample (i.e., n), and (c) size of population correlations (ρ_{XY}, ρ_{XZ}, and ρ_{YZ}), on the observed variance of empirically derived sampling distributions of r_{XY}s (i.e., $S^2_{or_{xy}}$). Then, we computed the difference between the empirically derived or observed sampling error variance $S^2_{or_{xy}}$ and the analytically derived or expected sampling error variance $S^2_{er_{xy}}$.

Manipulated Parameters

The following parameters were manipulated in the simulation.

IRR. We conducted an IRR on X and Y by restricting the range on Z using a top-down procedure (see *Simulation procedure* below). Range restriction on variable Z (i.e., SR) can be easily converted to v, the restricted to unrestricted population standard deviations (SDs) ratio (this mathematical equivalence is possible because Z is normally distributed; see Table 1). Because Z is correlated with X and Y, direct restriction on Z causes IRR on X and Y. In the simulation, SR (see Table 1) took on values ranging from 0.10 to 10.00, in increments of 0.10, to represent conditions ranging from very severe range restriction (i.e., SR = .1, sample scores represent the top 10.00% of the

Table 1
Sample Sizes Corresponding to Selection Ratios for Each of the Truncation Values n and Restricted to Unrestricted Standard Deviations Ratios v

		n		
SR	*v*	25	60	100
0.1	0.408	250	600	1,000
0.2	0.467	125	300	500
0.3	0.518	83	200	300
0.4	0.558	63	150	250
0.5	0.603	50	120	200
0.6	0.649	42	100	167
0.7	0.697	36	86	143
0.8	0.765	31	75	125
0.9	0.844	28	67	111
1.0	1.000	25	60	100

Note. $v = 1 + [(z)(f/SR)] - (f/SR)^2$, where z is the standard normal deviate corresponding to the selection ratio (*SR*) and f is the ordinate of the standard normal density function at z (Schmidt, Hunter, & Urry, 1976).

population distribution of Z scores; $v = .408$, sample *SD* is 40.80% as large as the *SD* of the population distribution of Z scores) to no range restriction (i.e., $SR = 1.0$, $v = 1.000$; sample scores represent the full population range).

Note that all the study-level correlation coefficients are affected by the same degree of IRR in each cell of the design. This is not typical in VG studies in which the severity of IRR is likely to vary from correlation coefficient to correlation coefficient. However, our study is not intended to mirror the typical VG study. To assess the estimation accuracy of sampling error variance in the presence of IRR, we needed to hold the population validity constant. Otherwise, error would be introduced in the final results because the correction for the real variance in true validities would not be perfectly accurate.

Sample size. Sample size n was set at values of 60, 100, and 140. These values cover a fairly typical range in several I&O psychology specialities, especially in personnel selection research. For example, Lent, Aurbach, and Levin (1971) found that the median sample size in 1,500 validation studies was 68. More recently, Russell et al. (1994) conducted a meta-analysis that included the 138 validation studies of personnel selection systems published in the *Journal of Applied Psychology* and *Personnel Psychology* between 1964 and 1992; he ascertained that the median sample size was 103 (C. J. Russell, personal communication, February 21, 1996).

Population intercorrelations. The correlations ρ_{XY}, ρ_{XZ}, and ρ_{YZ} were set at values between 0.10 and 0.90, in increments of 0.10, to represent varying degrees of effect size.

Summary. The manipulation of the independent variables led to a full factorial design with a total of 21,870 cells or conditions, that is, 10 (*SR*) × 3 (n) × 9 (ρ_{XY}) × 9 (ρ_{XZ}) × 9 (ρ_{YZ}). Note, however, that it is not possible to generate all possible combinations of correlations ranging from 0.10 to 0.90 among three variables. After two correlations are specified, the value of the third correlation has a limited range, as indicated by the following equation (McNemar, 1962, p. 167):

$$r_1 = (r_2 r_3) +/- \sqrt{1 - r_2^2 - r_3^2 + (r_2^2 r_3^2)}. \quad (1)$$

For example, when the correlation between X and Z is 0.80 (e.g., r_2) and the correlation between Y and Z is also 0.80 (e.g., r_3), the correlation between X and Y (i.e., r_1) can only take on values between 0.28 and 1.00. Because of this design consideration, the resulting number of combinations of parameter values (i.e., cells) in our study was 19,422 (88.81% of the cells in the hypothetical full factorial design).

Procedure and Dependent Variable

Computer programs. The simulation was performed using FORTRAN programs incorporating the International Mathematical and Statistical Libraries (1989) subroutine RNMVN that generates random normal scores under a user-supplied covariance matrix (cf. Aguinis, 1994).[1]

Simulation procedure. Five thousand samples were generated for each of the 19,422 cells (i.e., combination of parameter values) of the design. The simulation involved the following three steps.

1. Trivariate (X, Y, Z) arrays of size N were generated from multivariate normal populations with a mean of zero (i.e., $\mu_X = \mu_Y = \mu_Z = 0.00$), unit variance (i.e., $\sigma_X^2 = \sigma_Y^2 = \sigma_Z^2 = 1.00$), and correlations ρ_{XY}, ρ_{XZ}, and ρ_{YZ}.

2. The N generated trivariate arrays were sorted in descending order on Z and truncated at the nth value. The ratio n/N provides the *SR*. Identical to Millsap's (1989) procedure, the value of N was systematically manipulated to give desired values of *SR* for fixed values of n. Table 1, equivalent to Millsap's Table 1 (p. 457), shows arrays of size N corresponding to *SR*s for each truncation value N, together with corresponding values of v (i.e., ratio of restricted to unrestricted *SD*s).

3. Correlations r_{XY}, r_{XZ}, and r_{YZ} were calculated from each of the 5,000 samples generated for each cell in the design.

Dependent variable. To assess whether IRR on X and Y spuriously inflates the analytically derived expected variance in the sampling distribution of r_{XY}s, we followed Millsap's (1989) procedure and computed (a) the *observed* variance $S^2_{or_{xy}}$ from observed (i.e., generated) sampling distributions of r_{XY}s for each cell in the design and (b) Fisher's (1921, 1954) *expected* estimator $S^2_{er_{xy}}$ (shown in Equation 2) based on the average of 5,000 estimators. The 5,000 estimators were computed based on each of 5,000 r_{XY}s generated for each cell in the design:

$$S^2_{er_{xy}} = \frac{(1 - r^2_{XY})^2}{n - 1}. \quad (2)$$

Subsequently, the difference between the observed and the expected variances (d) was computed for each of the 19,422 parameter value combinations:

$$d = S^2_{or_{xy}} - S^2_{er_{xy}}. \quad (3)$$

Key Accuracy Checks

To assess the key accuracy (i.e., validity) of the computer programs, we first replicated Millsap's (1989) simulation and

[1] Source code versions of the programs can be obtained from Herman Aguinis.

compared $S^2_{or_{xy}}$s and $S^2_{er_{xy}}$s reported in his study with those generated using our newly developed computer programs. Values were generated for each cell of his 9 (ρ_{XY}) × 10 (SR) × 3 (n) full factorial design. Note that Millsap used n values of 25, 60, and 100, which differ from the values of 60, 100, and 140 used in our study. Thus, we also generated values for $n = 25$ to be able to fully compare our results with Millsap's. Subsequently, we computed the mean observed and expected variance for each sample size condition. Then, we formally compared our results with those reported by Millsap by conducting six independent-samples t tests. Observed and expected variance means, discrepancies, and t statistics are reported in Table 2. Values for the t statistics were very small and in no case approached the .05 statistical significance level. Thus, we concluded that our computer programs were valid, and we proceeded to the IRR–trivariate investigation.

Results and Discussion

In our simulation, we examined the effects of IRR on the difference between observed and expected error variances (i.e., d) in the sampling distribution of rs under various conditions of sample size and variable intercorrelations. Tables 3–5 show d values for factorial combinations of ρ_{XY} and SR collapsing across values of ρ_{XZ} and ρ_{YZ}, together with the percentage by which IRR increases the sampling error variance in r above its analytically derived expected value.

Table 3 shows the differences between observed and expected variances (ds) for a sample size of 60. An examination of this table shows that (a) ds are positive for all conditions, (b) values of d increase as ρ_{XY} decreases, and (c) values of d increase as SR changes from no restriction ($SR = 1.0$) to any level of restriction ($SR \neq 1.0$). Table 4, listing ds for $n = 100$, shows a similar pattern of results. However, the impact of IRR is not as strong as when $n = 60$. The same pattern of results is observed in Table 5 ($n = 140$), with an even further overall decrease in the values of d. Taken together, the findings presented in these tables indicate that (a) the sampling error variance estimator used in VG studies underestimates the true variance in r, even in the absence of IRR (i.e., $SR = 1.0$ conditions; Hunter & Schmidt, 1994); and (b) IRR accentuates this underestimation even further, especially in situations of small sample and effect size.

Comparison of Effects of IRR and Direct Range Restriction

Next, we compared the amount of sampling error variance underestimation in r under IRR with that under direct range restriction. Note that IRR and direct range restriction are independent processes. Direct range restriction on X or Y can be present in the absence or presence of IRR (i.e., direct range restriction on Z). Thus, both IRR and direct range restriction may have an impact on the variances of X and Y and, consequently, their correlation. However, they can operate in isolation or concurrently.

To compare the relative effects of these two types of range restriction, we contrasted our results with those reported by Millsap (1989). Table 6 presents means for d when $SR < 1.0$ (mean obtained from all levels of $SR \neq 1.0$) and $SR = 1.0$ (no range restriction), collapsed across the other manipulated parameter values for IRR and direct range restriction situations.

Results reported in Table 6 show that our findings regarding the impact of IRR on the sampling variance in r are virtually as strong as those reported by Millsap (1989) regarding the effect of direct range restriction. Similarly, both Millsap's simulation and ours demonstrate that (a) the analytically derived expected sampling variance given in Equation 2 is negatively biased and (b) this bias is larger in restricted than unrestricted data. Likewise, we also found that this bias increases as sample size decreases. Finally, note that we also generated data for $n = 25$ to compare our results with Millsap's. However, it should be noted that such an unusually small sample size deviates substantially from the typical sample size in today's validation studies (i.e., $n \approx 100$).

Table 6 indicates that, for $SR = 1.0$ (i.e., no restriction), our ds are slightly smaller than those reported by Millsap (1989). This downward trend in our results is due to the

Table 2
Variance Means, Discrepancies, and t Statistics

n	Variance M	Millsap	Replication	Discrepancy	$t(178)^a$
25	Observed	.0299363	.0304002	.0004638	.2690
25	Estimated	.0272177	.0273014	.0000838	.0524
60	Observed	.0121137	.0120363	−.0000773	−.0109
60	Estimated	.0114042	.0114131	.0000089	.0127
100	Observed	.0072179	.0071974	−.0000200	−.0478
100	Estimated	.0068450	.0068506	.0000056	.0131

Note. Millsap = results from Millsap (1989); Replication = values obtained using our study's computer programs; Discrepancy = replication variance − Millsap variance.
[a] $p > .05$, for all t statistics.

Table 3
Values of d ($S^2_{or_{xy}} - S^2_{er_{xy}}$) for n = 60

	SR										
ρ_{XY}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	*M*
0.10	.00101	.00077	.00117	.00107	.00073	.00048	.00110	.00080	.00027	.00041	.00078
	(7.71)	(5.67)	(7.88)	(7.16)	(4.87)	(3.08)	(6.83)	(4.97)	(1.64)	(2.52)	(5.23)
0.20	.00107	.00083	.00123	.00109	.00076	.00048	.00108	.00080	.00026	.00033	.00079
	(7.95)	(5.91)	(8.19)	(7.11)	(4.87)	(3.08)	(6.78)	(5.05)	(1.60)	(2.16)	(5.27)
0.30	.00096	.00072	.00111	.00093	.00063	.00039	.00095	.00072	.00020	.00023	.00068
	(6.75)	(4.97)	(7.38)	(6.03)	(4.03)	(2.51)	(6.23)	(4.85)	(1.34)	(1.70)	(4.58)
0.40	.00088	.00066	.00097	.00078	.00055	.00034	.00079	.00061	.00017	.00016	.00059
	(6.25)	(4.67)	(6.94)	(5.29)	(3.71)	(2.36)	(5.77)	(4.67)	(1.30)	(1.32)	(4.23)
0.50	.00080	.00061	.00083	.00065	.00049	.00032	.00063	.00049	.00016	.00010	.00051
	(6.57)	(4.97)	(6.84)	(4.94)	(3.76)	(2.55)	(5.35)	(4.46)	(1.45)	(1.00)	(4.19)
0.60	.00053	.00039	.00054	.00039	.00032	.00021	.00041	.00033	.00010	.00005	.00033
	(5.60)	(4.09)	(5.81)	(3.75)	(3.15)	(2.33)	(4.55)	(3.94)	(1.29)	(0.68)	(3.52)
0.70	.00039	.00028	.00035	.00024	.00022	.00015	.00025	.00020	.00006	.00002	.00022
	(4.93)	(3.44)	(4.98)	(2.82)	(2.99)	(2.35)	(3.93)	(3.38)	(1.15)	(0.44)	(3.03)
0.80	.00017	.00012	.00014	.00009	.00010	.00007	.00010	.00008	.00002	.00001	.00009
	(4.55)	(2.79)	(4.04)	(2.04)	(2.71)	(2.11)	(3.04)	(2.47)	(0.93)	(0.31)	(2.51)
0.90	.00004	.00003	.00003	.00002	.00003	.00002	.00002	.00001	.00000	.00000	.00002
	(5.07)	(2.76)	(3.70)	(1.60)	(3.20)	(1.91)	(2.30)	(1.38)	(0.65)	(0.26)	(2.28)
M	.00065	.00049	.00071	.00058	.00042	.00027	.00059	.00045	.00014	.00014	.00045
	(6.25)	(4.36)	(6.19)	(4.52)	(3.69)	(2.49)	(4.97)	(3.91)	(1.26)	(1.15)	(4.01)

Note. $S^2_{or_{xy}}$ = observed sampling error variance in r_{XY}; $S^2_{er_{xy}}$ = expected sampling error variance in r_{XY}; *SR* = selection ratio on *Z*; *n* = sample size. Values in parentheses show the percentage by which $S^2_{or_{xy}}$ is larger than $S^2_{er_{xy}}$.

fact that our simulation design consisted of trivariate (as compared to Millsap's bivariate) arrays and, thus, was not a full factorial. Because, after two correlation values are specified, values for the third correlation are limited, our design was slightly unbalanced in that it included more higher (i.e., closer to 0.90) than lower (closer to 0.10) values for ρ_{XY} (cf. Equation 1). Consequently, because *d* decreases as *r* increases, our values for *d* in the

Table 4
Values of d ($S^2_{or_{xy}} - S^2_{er_{xy}}$) for n = 100

	SR										
ρ_{XY}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	*M*
0.10	.00065	.00050	.00034	.00031	.00068	.00064	.00063	.00018	.00041	.00001	.00043
	(8.38)	(6.18)	(4.23)	(3.58)	(7.27)	(6.70)	(6.56)	(1.89)	(4.16)	(0.09)	(4.90)
0.20	.00068	.00053	.00035	.00031	.00066	.00064	.00060	.00018	.00040	.00002	.00044
	(8.55)	(6.35)	(4.07)	(3.43)	(7.05)	(6.70)	(6.20)	(1.83)	(4.23)	(0.27)	(4.87)
0.30	.00061	.00046	.00028	.00022	.00055	.00055	.00048	.00014	.00036	.00004	.00037
	(7.09)	(5.07)	(2.93)	(2.31)	(5.98)	(6.05)	(5.22)	(1.48)	(4.12)	(0.54)	(4.08)
0.40	.00056	.00040	.00024	.00017	.00044	.00046	.00037	.00011	.00030	.00005	.00031
	(6.38)	(4.40)	(2.43)	(1.60)	(5.22)	(5.61)	(4.49)	(1.40)	(3.94)	(0.76)	(3.62)
0.50	.00051	.00036	.00022	.00014	.00034	.00037	.00027	.00010	.00023	.00005	.00026
	(6.63)	(4.37)	(2.52)	(1.33)	(4.71)	(5.41)	(3.84)	(1.54)	(3.64)	(0.94)	(3.49)
0.60	.00034	.00021	.00013	.00005	.00020	.00024	.00015	.00006	.00015	.00004	.00016
	(5.42)	(3.17)	(1.83)	(0.41)	(3.59)	(4.70)	(2.84)	(1.46)	(3.08)	(0.99)	(2.75)
0.70	.00025	.00014	.00009	.00003	.00011	.00015	.00008	.00004	.00008	.00002	.00010
	(4.59)	(2.26)	(1.48)	(−0.14)	(2.73)	(4.40)	(2.14)	(1.55)	(2.50)	(0.91)	(2.24)
0.80	.00011	.00005	.00004	.00000	.00004	.00006	.00003	.00002	.00003	.00001	.00004
	(4.05)	(1.79)	(1.65)	(−0.31)	(1.87)	(3.93)	(1.51)	(1.66)	(1.73)	(0.55)	(1.84)
0.90	.00002	.00001	.00001	.00000	.00001	.00002	.00001	.00001	.00001	.00000	.00001
	(4.90)	(0.92)	(2.82)	(0.74)	(1.88)	(3.99)	(1.94)	(2.05)	(1.13)	(0.45)	(2.08)
M	.00041	.00029	.00019	.00013	.00034	.00035	.00029	.00009	.00022	.00003	.00024
	(6.22)	(3.83)	(2.66)	(1.44)	(4.48)	(5.28)	(3.86)	(1.65)	(3.17)	(0.61)	(3.32)

Note. $S^2_{or_{xy}}$ = observed sampling error variance in r_{XY}; $S^2_{er_{xy}}$ = expected sampling error variance in r_{XY}; *SR* = selection ratio on *Z*; *n* = sample size. Values in parentheses show the percentage by which $S^2_{or_{xy}}$ is larger than $S^2_{er_{xy}}$.

Table 5
Values of d ($S^2_{or_{xy}} - S^2_{er_{xy}}$) for n = 140

	SR										
ρ_{xy}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	*M*
0.10	.00019	.00028	.00033	.00027	.00024	.00022	.00042	.00021	.00027	.00018	.00026
	(3.92)	(4.95)	(5.43)	(4.29)	(3.78)	(3.26)	(6.10)	(2.93)	(3.89)	(2.59)	(4.11)
0.20	.00022	.00030	.00034	.00030	.00024	.00022	.00038	.00019	.00026	.00016	.00026
	(4.34)	(5.17)	(5.39)	(4.62)	(3.50)	(3.17)	(5.57)	(2.65)	(3.83)	(2.45)	(4.07)
0.30	.00020	.00026	.00027	.00026	.00018	.00018	.00030	.00013	.00023	.00014	.00021
	(3.43)	(4.10)	(4.13)	(3.94)	(2.51)	(2.59)	(4.51)	(1.92)	(3.57)	(2.29)	(3.30)
0.40	.00020	.00023	.00022	.00024	.00014	.00016	.00022	.00008	.00018	.00010	.00018
	(3.34)	(3.64)	(3.49)	(3.82)	(1.97)	(2.42)	(3.59)	(1.34)	(3.31)	(1.95)	(2.89)
0.50	.00021	.00022	.00020	.00022	.00012	.00015	.00015	.00005	.00014	.00006	.00015
	(4.27)	(3.83)	(3.42)	(4.10)	(1.85)	(2.66)	(2.83)	(0.96)	(3.14)	(1.56)	(2.86)
0.60	.00015	.00013	.00011	.00012	.00006	.00010	.00007	.00001	.00009	.00003	.00009
	(3.99)	(2.83)	(2.40)	(3.65)	(1.15)	(2.42)	(1.73)	(0.34)	(2.65)	(1.04)	(2.22)
0.70	.00012	.00009	.00007	.00010	.00004	.00007	.00003	.00000	.00005	.00001	.00006
	(3.93)	(2.00)	(1.75)	(3.55)	(0.85)	(2.50)	(0.95)	(0.08)	(2.26)	(0.42)	(1.83)
0.80	.00006	.00003	.00003	.00005	.00002	.00003	.00001	−.00000	.00002	.00000	.00002
	(4.42)	(1.46)	(1.55)	(3.63)	(0.83)	(2.56)	(0.39)	(−0.19)	(1.81)	(0.05)	(1.65)
0.90	.00001	.00001	.00001	.00001	.00001	.00001	.00000	−.00000	.00000	−.00000	.00001
	(4.79)	(1.10)	(2.80)	(4.07)	(1.32)	(3.04)	(0.99)	(−0.02)	(1.40)	(−1.54)	(1.80)
M	.00015	.00017	.00017	.00018	.00012	.00013	.00018	.00007	.00014	.00007	.00014
	(4.05)	(3.23)	(3.37)	(3.96)	(1.97)	(2.74)	(2.96)	(1.11)	(2.87)	(1.20)	(2.75)

Note. $S^2_{or_{xy}}$ = observed sampling error variance in r_{xy}; $S^2_{er_{xy}}$ = expected sampling error variance in r_{xy}; *SR* = selection ratio on *Z*; *n* = sample size. Values in parentheses show the percentage by which $S^2_{or_{xy}}$ is larger than $S^2_{er_{xy}}$.

absence of restriction suffer a slight negative bias as compared with Millsap's results. This design consideration strengthens the relevance of our study's results because our estimates regarding the degree of negative bias in the sampling error variance estimate in *r* obtained using Equation 2 should be considered somewhat conservative.

Effects of Variable Intercorrelations

We also investigated the degree to which *d* values were affected by the level of shared variance between the variable pairs *Z*–*X* and *Z*–*Y*. This shared variance can be precisely expressed as $R^2_{Z.XY}$, namely, the proportion of variance in *Z* accounted for by variables *X* and *Y* (Pedhazur, 1982, p. 107):

$$R^2_{Z.XY} = \frac{r^2_{XZ} + r^2_{YZ} - (2r_{XZ}r_{YZ}r_{XY})}{1 - r^2_{XY}}. \qquad (4)$$

Under the condition of direct range restriction on *Z*, IRR on *X* and *Y* increases to the extent that $R^2_{Z.XY}$ increases and, consequently, values of *d* should increase. To examine this prediction, we plotted representative values of *SR* (i.e., 0.1, 0.4, 0.7, and 1.0) with values of $R^2_{Z.XY}$, ρ_{XY}, *n*, and *d* in Figures 1–3.

A perusal of Figures 1–3 indicates that *d* values increase as (a) $R^2_{Z.XY}$ (i.e., shared variance) increases, (b) ρ_{XY} decreases, and (c) *SR* shifts from 1.0 to any other value (i.e., from no IRR to any level of IRR). These effects are strongest in Figure 1 ($n = 60$) and smaller in magnitude, yet still noticeable, as sample size increases to 100 (Figure 2) and to 140 (Figure 3).

It should be noted that the three peaks for *d* values shown in Figures 1–3, and more noticeable for $n = 60$, are due to the facts that (a) larger values of $R^2_{Z.XY}$ were found to increase *d*, (b) small values for ρ_{XY} were found to increase *d*, and (c) our design did not include all possible combinations of correlations between variables *X*, *Y*, and *Z* (cf. Equation 1). For example, Figure 1a shows a value for *d* virtually reaching .005 for $\rho_{XY} = .30$,

Table 6
Values of d ($S^2_{or_{xy}} - S^2_{er_{xy}}$) Under Direct and Indirect Range Restriction

Range restriction and *n*	Selection ratio	
	<1.0	1.0
Direct (Millsap, 1989)		
25	.00283	.00174
60	.00075	.00035
100	.00040	.00016
140[a]	—	—
Indirect (our study)		
25	.00224	.00146
60	.00048	.00014
100	.00027	.00003
140[a]	.00015	.00007

[a] Millsap's investigation of the effects of direct range restriction included sample size of only 25, 60, and 100.

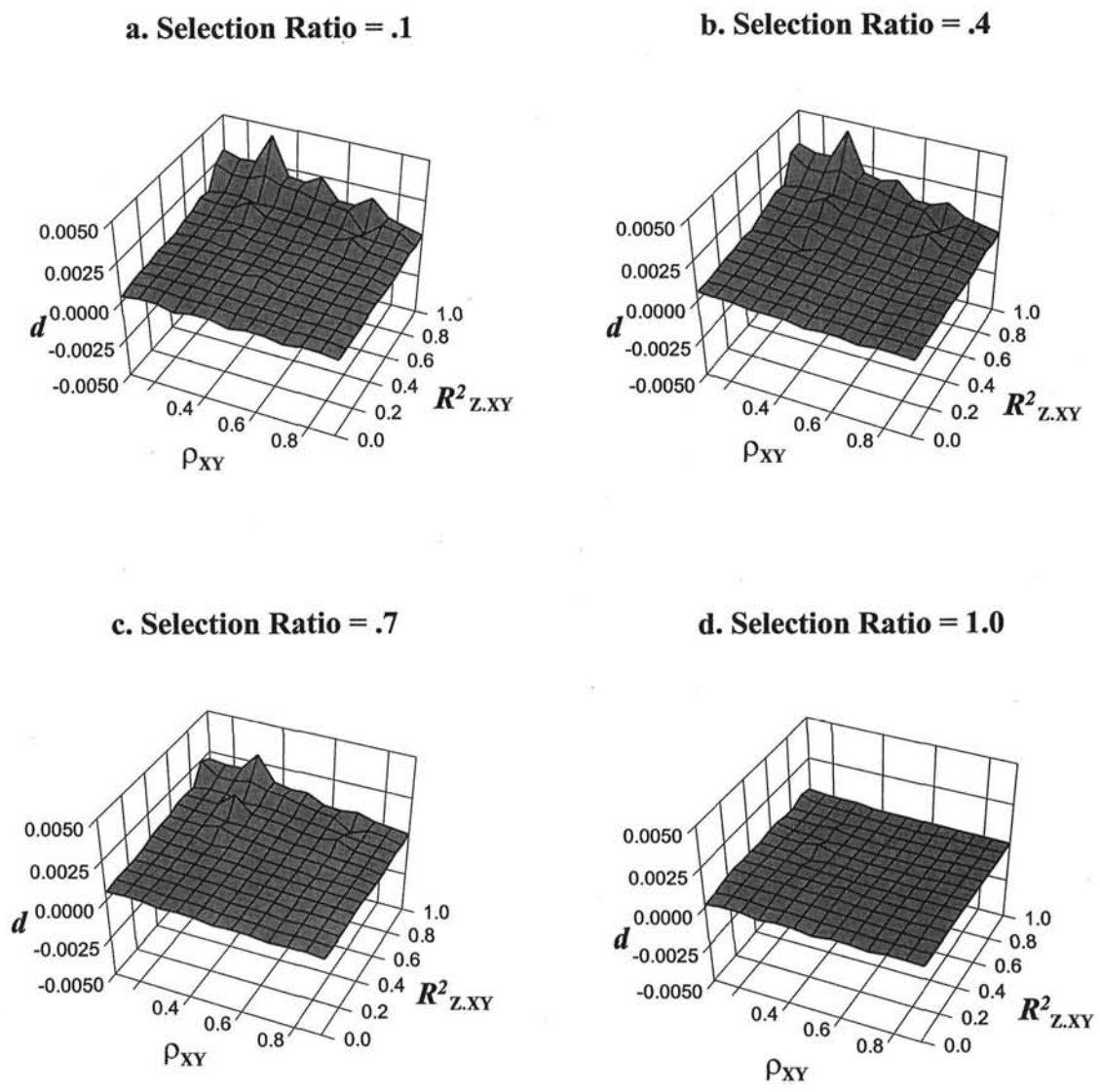

Figure 1. Values for d as a function of selection ratio, validity coefficient (ρ_{XY}), and shared variance between the directly restricted variable (Z) and the two indirectly restricted variables (X, Y; $R^2_{Z.XY}$) for a sample size of 60.

and $R^2_{Z.XY} \approx .97$, correspond to $\rho_{XZ} = \rho_{YZ} = .80$. No other d value in the graph is close to reaching .005 because no other combination of values for ρ_{XZ} and ρ_{YZ} can yield a comparably high $R^2_{Z.XY} \approx .97$, given the constraint shown in Equation 1 of $\rho_{XY} \approx .30$ or lower. Consequently, the next possible highest combination of values for $R^2_{Z.XY}$ and ρ_{XY} in the design are considerably lower and so is the resulting d.

Conclusions and Implications

Primary-level as well as meta-analytic researchers are concerned with the estimation of moderating effects (e.g., Aguinis, 1995; Aguinis, Bommer, & Pierce, 1996; Aguinis, Pierce, & Stone-Romero, 1994; Aguinis & Stone-Romero, 1997; Stone-Romero, Alliger, & Aguinis, 1994). The results of our study lead to several meaningful conclusions regarding the estimation of moderating effects in the context of VG procedures.

First, as Schmidt, Hunter, and their colleagues have advocated for over 1 decade (Hunter & Schmidt, 1994; Law, Schmidt, & Hunter, 1994a; Schmidt & Hunter, 1978), sampling error variance in r computed using Equation 2 is systematically underestimated in VG studies. Our MC simulations confirm that, even in the absence of IRR, the analytically estimated sampling variance computed using

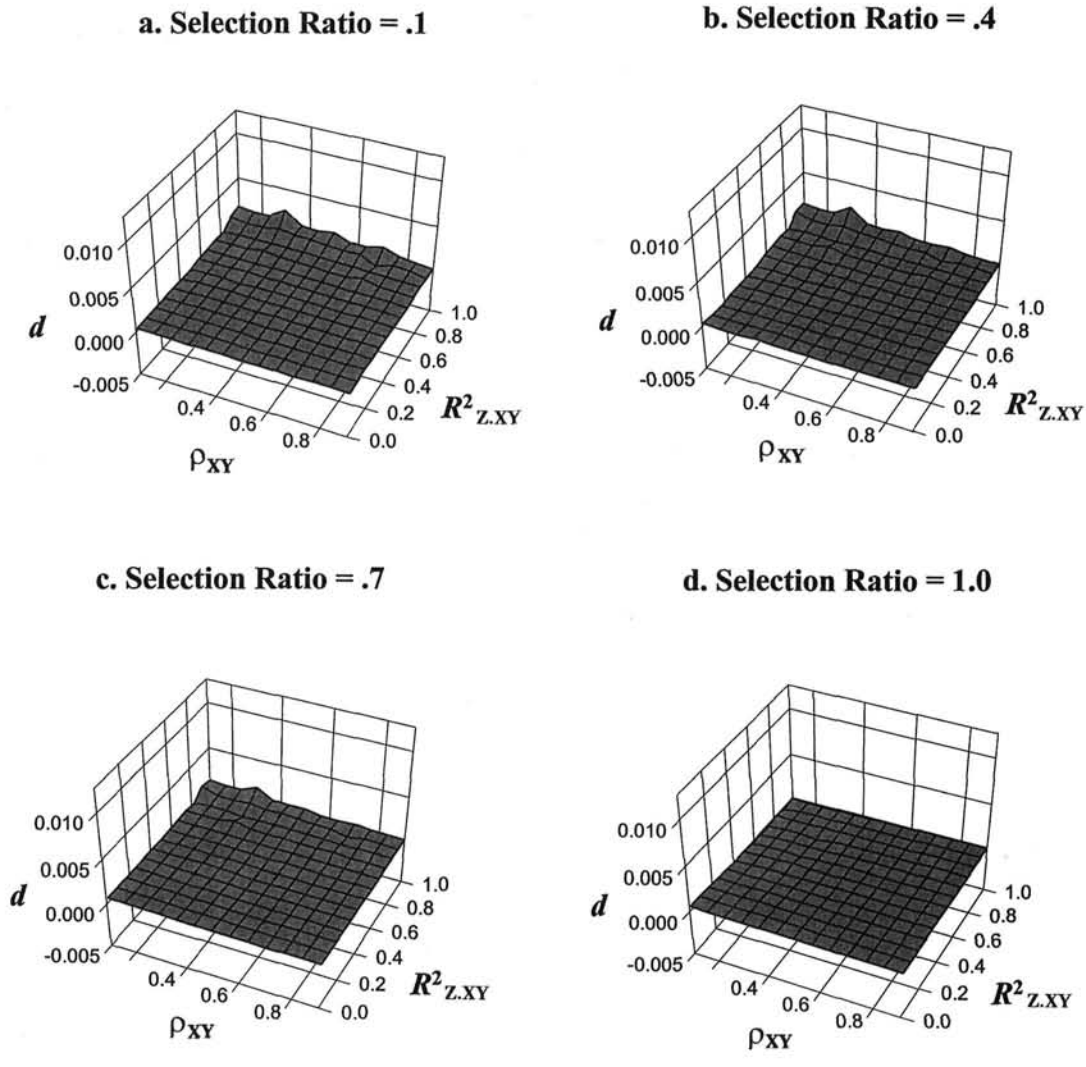

Figure 2. Values for d as a function of selection ratio, validity coefficient (ρ_{XY}), and shared variance between the directly restricted variable (Z) and the two indirectly restricted variables (X, Y; $R^2_{Z.XY}$) for a sample size of 100.

Equation 2 has a systematic negative bias as compared with the empirically computed sampling variance based on the actual generated data. This underestimation may be quite large under some conditions, especially as sample sizes approach smaller values (e.g., 60) and effect sizes are small (i.e., .50 or smaller, which can be considered to be the typical range for validity coefficients in personnel selection research). For instance, when $n = 60$, in the absence of IRR, and when collapsing across all levels of variable intercorrelations, the observed sampling error variance is 1.15% larger than the analytically derived (i.e., expected) variance. When $\rho_{XY} = .20$, in the absence of IRR, and when collapsing across all sample sizes (i.e., 60, 100, and 140), the observed sampling error variance is 1.63% larger than the expected value computed in VG studies.

Second, a new finding and unique contribution of our study is the conclusion that IRR worsens the underestimation of the VG sampling error variance estimator significantly. In situations of IRR, the actual variance in r across studies is underestimated even more radically when a (a) sample size is smaller than 100 and approaches 60 (see Figures 1–3), (b) true validity is .60 or smaller (see Tables 3–5), and (c) shared variance between the directly restricted variable and the two IRR variables (i.e., $R^2_{Z.XY}$) is approximately .75 (e.g., $\rho_{XY} = .60$, $\rho_{XZ} = .20$, and $\rho_{YZ} = .80$; see Figure 1). Also, the effects of IRR were found to be comparable in magnitude with the effects of direct range restriction. Finally, it should be noted that the severity of IRR is not as important as the mere presence of any degree of IRR.

To use a meaningful example, the concurrent presence

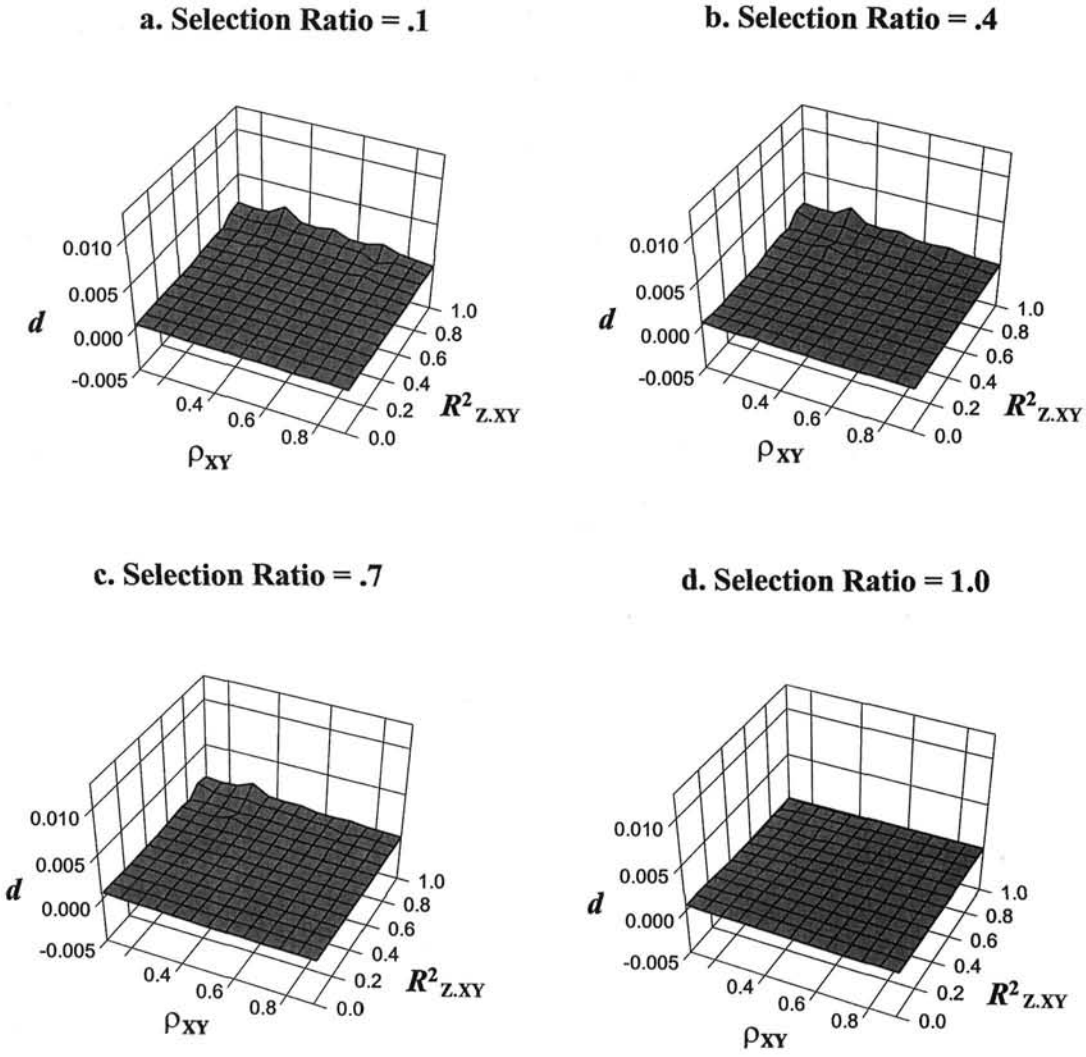

Figure 3. Values for d as a function of selection ratio, validity coefficient (ρ_{XY}), and shared variance between the directly restricted variable (Z) and the two indirectly restricted variables (X, Y; $R^2_{Z.XY}$) for a sample size of 140.

of IRR ($SR < 1.0$), sample size of 60, true validity (ρ_{XY}) of .60, ρ_{XZ} = .60, and ρ_{YZ} = .60 yields an observed sampling error variance in r 3.16% larger than the value computed using Fisher's (1921, 1954) estimator. For the same parameters and a sample size of 100, this difference decreases to 2.34%; for a sample size of 140, it is further reduced to 1.85%. Thus, the proportion of variance in r due to IRR is always above zero, and this proportion increases substantially as sample size approaches 60. As a consequence of the increased sampling error variance in the correlation coefficient caused by IRR and in the absence of theory-based hypotheses regarding the impact of IRR, researchers may incorrectly assume that this artifactual variance is due to potential moderator variables, and, hence, false moderators may be "discovered."

In summary, Tables 3–5 show that when collapsing across values of ρ_{XY}, ρ_{XZ}, and ρ_{YZ}, IRR artificially inflates the variance in r up to a high percentage ($\approx$ 8.50%). Also, Tables 3–5 indicate that, even though the proportion of variance increase is quite sizable for some conditions (e.g., small sample size and moderate or small effect size), it is modest for others (e.g., large sample size and large effect size). Thus, that the median sample size in the validation research literature is larger now than 20 years ago (i.e., 103 vs. 68) is encouraging. Nevertheless, because IRR-caused variability may be incorrectly attributed to nonsubstantive moderating effects and the situational specificity hypothesis may be incorrectly assumed to be valid, the presence of IRR should not be ignored in future VG endeavors. Moreover, researchers should con-

sider a priori hypotheses regarding the presence of substantive moderator variables that may cause IRR.

Limitations and Research Needs

Our MC study used a multivariate random normal generator. Thus, even though complying with the (multivariate) normality assumption is common practice in MC investigations of VG and meta-analytic methods in general (e.g., Callender & Osburn, 1981; Millsap, 1989), we acknowledge that our study's results may not be generalizable to situations in which this assumption is not tenable.

Second, our study ascertained the effects of IRR on the sampling error variance in the correlation coefficient. This is a new finding and unique contribution to the VG literature. However, the conclusions of this research leave VG researchers in a perhaps uncomfortable situation. Unless information is gathered regarding possible IRR in the primary-level studies used in a VG investigation, VG researchers cannot establish whether unexplained variance due to IRR is artifactual or caused by potential moderators. To remedy this difficulty, at present, we can only extend Hunter and Schmidt's (1990) recommendation that primary-level researchers report as much information as possible regarding their studies, so eventual quantitative reviews can be as accurate as possible. This would include not only the reporting of statistics to be used in a meta-analysis but also the reporting of detailed procedures used to collect the data, including information regarding IRR and direct range restriction processes (e.g., Hattrup & Schmitt, 1990).

We foresee at least two avenues for future research. First, given our study's results regarding the impact of IRR on the sampling variance in the correlation coefficient, it would be desirable that future researchers address possible statistical corrections to prevent the negative bias in $S^2_{er_{xy}}$ in IRR situations. Because direct restriction (prior to IRR) can occur on more than one variable (i.e., a cognitive abilities test, Z_1, and a personality test, Z_2), future researchers should address the question of whether these corrections should be (a) performed individually by assessing the impact of each restricted Z variable or (b) computed only once on the basis of the compound effect of all Z variables. Research on range restriction corrections by Ree, Carretta, Earles, and Albert (1994) demonstrated that a multivariate correction does not yield the same results as a series of univariate corrections. More specifically, Lawley's (1943) multivariate correction results in corrected correlations that are closer to the specified population parameters as compared with correcting a matrix one correlation at a time. Thus, Ree et al.'s conclusion suggests that a multivariate correction is preferred. However, this recommendation relies on the fairly restrictive assumption that a researcher has all the information needed to implement it (e.g., restricted and unrestricted *SD*s for all variables involved).

Second, if feasible, in the future researchers should examine the extent to which IRR may have increased the across-study variance in already published VG investigations that accounted for less than 100.00% of this variability. Our study's results suggest that the presence of IRR may have led researchers who did not have substantive a priori moderating effect hypotheses regarding the effect of IRR to the erroneous conclusion of the possible presence of unexplained moderating effects.

A Closing Comment

In closing, we urge researchers to consider the implications of IRR for the conduct of quantitative summaries of research literatures. In the presence of IRR, variability across study-level rs can be underestimated by as much as 8.50%. In such IRR situations, researchers need to make theory-based decisions to ascertain whether the effects of IRR are artifactual or caused by situational-specific moderating effects.

References

Aguinis, H. (1994). A QuickBASIC program for generating correlated multivariate random normal scores. *Educational and Psychological Measurement, 54,* 687–689.

Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management, 21,* 1141–1158.

Aguinis, H., Bommer, W. H., & Pierce, C. A. (1996). Improving the estimation of moderating effects by using computer-administered questionnaires. *Educational and Psychological Measurement, 56,* 1043–1047.

Aguinis, H., & Pierce, C. A. (in press). Testing moderator variable hypotheses meta-analytically. *Journal of Management.*

Aguinis, H., Pierce, C. A., & Stone-Romero, E. F. (1994). Estimating the power to detect dichotomous moderators with moderated multiple regression. *Educational and Psychological Measurement, 54,* 690–692.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82,* 192–206.

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

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1,* 1–32.

Fisher, R. A. (1954). *Statistical methods for research workers.* Edinburgh, Scotland: Oliver & Boyd.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: Freeman.

Hartley, H. O., & Harris, D. L. (1963). Monte Carlo computa-

tions in normal correlation problems. *Journal of the Association for Computing Machinery, 10,* 302–306.

Hattrup, K., & Schmitt, N. (1990). Prediction of trades apprentices' performance on job sample criteria. *Personnel Psychology, 43,* 453–466.

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. (1994). Estimation of sampling error variance in the meta-analysis of correlations: Use of average correlation in the homogeneous case. *Journal of Applied Psychology, 79,* 171–177.

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

International Mathematical and Statistical Libraries. (1989). *IMSL library reference manual* (10th ed.). Houston, TX: Author.

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.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology, 80,* 94–106.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994a). Nonlinearity of range corrections in meta-analysis: Test of an improved procedure. *Journal of Applied Psychology, 79,* 425–438.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994b). A test of two refinements in procedures for meta-analysis. *Journal of Applied Psychology, 79,* 978–986.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh, 62*(Section A, Pt. 1), 28–30.

Lent, R. H., Aurbach, H. A., & Levin, L. S. (1971). Research design and validity assessment. *Personnel Psychology, 24,* 247–274.

Linn, R. L. (1983a). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement, 20,* 1–15.

Linn, R. L. (1983b). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 27–40). Hillsdale, NJ: Erlbaum.

McNemar, Q. (1962). *Psychological statistics* (3rd ed.). New York: Wiley.

Mendoza, J. G., & Reinhardt, R. N. (1991). Validity generalization procedures using sample-based estimates: A comparison of six procedures. *Psychological Bulletin, 110,* 596–610.

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

Noreen, E. W. (1989). *Computer intensive methods for testing hypotheses.* New York: Wiley–Interscience.

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.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: Holt Reinhart Winston.

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology, 79,* 298–301.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo method.* New York: Wiley.

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

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. (1978). Moderator research and the law of small numbers. *Personnel Psychology, 31,* 215–232.

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., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473–485.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78,* 3–12.

Stone-Romero, E. F., Alliger, G. M., & Aguinis, H. (1994). Type II error problems in the use of moderated multiple regression for the detection of moderating effects of dichotomous variables. *Journal of Management, 20,* 167–178.

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

Received March 6, 1996
Revision received February 21, 1997
Accepted February 24, 1997 ■