Fixed Effects vs. Random Effects Meta-Analysis Models: Implications for Cumulative Research Knowledge

John E. Hunter and Frank L. Schmidt*

Research conclusions in the social sciences are increasingly based on meta-analysis, making questions of the accuracy of meta-analysis critical to the integrity of the base of cumulative knowledge. Both fixed effects (FE) and random effects (RE) meta-analysis models have been used widely in published meta-analyses. This article shows that FE models typically manifest a substantial Type I bias in significance tests for mean effect sizes and for moderator variables (interactions), while RE models do not. Likewise, FE models, but not RE models, yield confidence intervals for mean effect sizes that are narrower than their nominal width, thereby overstating the degree of precision in meta-analysis findings. This article demonstrates analytically that these biases in FE procedures are large enough to create serious distortions in conclusions about cumulative knowledge in the research literature. We therefore recommend that RE methods routinely be employed in meta-analysis in preference to FE methods.

Introduction

In the social and behavioral sciences today, conclusions about cumulative knowledge are increasingly based on the results of meta-analyses (Cooper and Hedges 1994; Hunter and Schmidt 1996). One indication of this is the large number of meta-analyses appearing in research journals in psychology and related areas, including journals that formerly published only individual empirical studies. Another indication is the fact that textbooks summarizing knowledge within fields increasingly cite meta-analyses rather than a selection of primary studies, as was the case until recently (Hunter and Schmidt 1996).

This development means that the soundness of the research endeavor is dependent on the accuracy of meta-analyis methods. Any substantial inaccuracy in widely used meta-analysis methods has fundamental implications for the quality of the resulting knowledge structures. This article focuses on an issue in meta-analysis that has such implications.

Within meta-analysis methods there is a distinction between fixed effects (FE) models (Hedges and Olkin 1985, ch. 7) and random effects (RE) models (Hedges and Olkin 1985, ch. 9). These models lead to different significance tests and confidence intervals for mean effect sizes (mean r or mean d). They also yield different significance tests for moderator variables (interactions) in meta-analysis; that is, different significance tests for the relation between study characteristics and study outcomes (effect sizes) (National Research Council 1992; Overton 1998). Hedges (1992: 284–92) provides a succinct overview of the differences between these two models in meta-analysis. Other treatments of this distinction can be found in Hedges (1994a; 1994b), National Research Council (1992), Raudenbush (1994), and Shadish and Haddock (1994).

Application of FE significance tests and confidence intervals is based on the assumption that the studies being analyzed are homogeneous at the level of study population effect sizes. For example, if the effect size index used is the d value, the FE model assumes that , the population value of d, is the same in all studies included in the meta-analysis. RE models do not make this assumption (Hedges and Olkin, ch. 9; Hedges 1992; National Research Council 1992). RE models allow for the possibility that the population parameter values vary from study to study (Becker 1996; Hedges 1992).

The methods described in Hunter, Schmidt and Jackson (1982), Hunter and Schmidt (1990a), Callender and Osburn (1980), and Raju and Burke (1983) are RE models (Hedges and Olkin 1985, ch. 9: 242; National Research Council 1992: 94-5). These methods have been extensively applied to substantive questions in the published literature (e.g., see Schmidt 1992). The methods described in Hedges (1988), Hedges and Olkin (1985, ch. 9), Raudenbush and Bryk (1985), and Rubin (1980, 1981) are also RE methods. These latter methods have been

* Address for correspondence: Frank Schmidt, College of Business, University of Iowa, Iowa City, IA 52242. E-mail: frank-schmidt@uiowa.edu

used less frequently in meta-analysis. For example, although *Psychological Bulletin*, the major review journal in psychology, has published many meta-analyses, we could locate no meta-analyses published in that journal that employed these methods. Cooper stated: 'In practice, most meta-analysts opt for the fixed effects assumption because it is analytically easier to manage' (1997: 179). The National Research Council report stated that many users of meta-analysis prefer FE models because of 'their conceptual and computational simplicity' (1992: 52).

RE formulas for statistical significance of mean effect sizes and moderator relationships have the appropriate Type I error rate (e.g., 5% for a designated alpha of .05), both when population parameter values (ρ or δ) are the same across all studies and when the population parameters vary across studies. However, when population parameters vary across studies, the FE formulas have Type I error rates that are higher than the nominal values – often much higher.[1] And if confidence intervals are used based on the FE standard errors, the confidence intervals are too narrow. For example, a nominal 95% confidence interval may actually be a 60% confidence interval, a substantial inaccuracy, and a substantial over-statement of the precision of the meta-analysis results.

The fact that FE models produce inaccurate results unless population effect sizes are constant across studies is important because it is likely that there is at least some variation in study population parameters in all research domains. Many would argue that for theoretical or substantive reasons there is always some variation in population parameter values across studies. That is, they would argue that there are always at least some real (i.e., substantive, not methodological) moderator variables that create differing values of δ_i or ρ_i across studies (National Research Council 1992). We do not argue for this position, because based on our experience some study domains do appear to be homogeneous at the level of substantive population parameters (e.g., see Schmidt, Law, Hunter, Rothstein, Pearlman and McDaniel 1993). However, whether this is true can be ascertained only by using RE models to estimate the level of heterogeneity. FE models do not allow for an estimate of S^2_ρ or S^2_δ, because they assume homogeneity *a priori*. That is, they assume $S^2_\rho = 0$ and $S^2_\delta = 0$.

Even if there are no substantive moderators causing variation in population parameters, there are methodological variations across studies that cause variation in study population ρ_i or δ_i. For example, if the amount of measurement error in the measures used varies across studies, this variation creates variation in study population parameters; studies with more measurement error will have smaller study population values of δ_i or ρ_i. So even if there is no substantive variation in population parameters, variations across studies in such methodological factors as reliability of measurement, range variation, or dichotomization of continuous variables (Hunter and Schmidt 1990b) will create variation in study population parameters (Osburn and Callender 1992). Such variation will typically exist and, hence, the assumption of homogeneous study population effect sizes or correlations will usually be false for this reason alone.

The formulas for statistical significance used in published applications of the Hedges and Olkin (1985) and Rosenthal and Rubin (1982a, b) meta-analysis methods are almost invariably FE formulas.[2] Hedges and Olkin (1985) specify that a chi square test for homogeneity should precede the test for the significance of the mean correlation or d value. This chi square test is the same for both the fixed and RE models (Hedges and Olkin 1985, ch. 9; Hedges 1992) and does not itself suffer from any Type I error bias. If this test is non-significant, then the chi square test cannot reject the hypothesis of homogeneity. However, a non-significant homogeneity test does not support a conclusion of homogeneity of study population values of ρ and δ. Unless the number of studies is large, this chi square test typically has low power to detect variation in study population parameters (Hedges and Olkin 1985; Mengersen, Tweedie, and Biggerstaff 1995; Morris and DeShon 1997; National Research Council 1992, p. 52), resulting in frequent Type II errors. That is, the chi square is often non-significant in the presence of real variation in study population parameters (Hedges and Olkin 1985). As a result, FE significance tests are often applied to heterogeneous study domains, resulting in inflated Type I error rates and confidence intervals (around mean effect sizes) that are substantially narrower than the actual confidence intervals.

In addition, even if the chi square test of homogeneity *is* significant (indicating heterogeneity of population effect sizes), users of FE methods nevertheless often apply FE formulas for statistical significance of mean effect sizes (or compute confidence intervals using the FE standard error of the mean effect size). This practice ensures distorted results and conclusions. Examples of meta-analysis that have been done this include Bettencourt and Miller (1996), Bond and Titus (1983), Burt, Zembar, and Niederehe (1995), Collins and Miller (1994), Eagly and Johnson (1990), Eagly, Karau, and Makhijani (1995), Eagly, Makhihani, and Klonsky (1992), Erel and Burman (1996), Feingold (1994), Ito, Tiffany, Miller, and Pollock (1996), Knight, Fabes, and Higgins (1996),

Newcomb and Bagwell (1995), Polich, Pollock, and Bloom (1995), Symons and Johnson (1997), Van Ijzendorn (1995), Wood (1987), and Wood, Lundgren, Ouellette, Busceme, and Blackstone (1994). These meta-analyses all focus on substantive research questions that are quite important. With these practices, it is even more likely that FE significance tests will have substantially inflated Type I error rates and will report falsely narrow confidence intervals, overestimating the precision of the findings.[3]

The National Research Council (1992: 147) stated that the use of FE models in meta-analysis is 'the rule rather than the exception' and that FE models 'tend to understate actual uncertainty' in research findings. The National Research Council recommended 'an increase in the use of RE models in preference to the current default of FE models' (ibid: 2), (see also pp. 185–7 of that report.). Others have also warned that use of FE models can lead to inflated Type I error rates and erroneously narrow confidence intervals (e.g., Hedges 1994a; Raudenbush 1994; Rosenthal 1995).[4] However, FE models have continued to be 'the rule rather than the exception' in the published literature in psychology and other disciplines.

In this article, we focus on the two most commonly used significance tests in meta-analysis: (1) the test for the mean effect size (e.g., mean r or mean d) in the domain of studies; and (2) the test for potential moderator variables. We quantitatively calibrate the extent to which FE significance tests have Type I error rates that are higher than nominal Type I error rates in heterogeneous domains. We also calibrate the extent to which confidence intervals computed using FE methods are narrower than their true values, suggesting levels of precision in meta-analysis results that do not exist. The literature currently includes no other studies that estimate the magnitude of the errors induced in these two tests by the FE model. Our demonstrations are in terms of correlations, but the same principles and conclusions apply to standardized effect sizes (d-values) and to other effect size indices (such as proportions and odds ratios).

Rationales for the Fixed Effects Model

What rationales have been offered for the FE model? The initial rationale was that scientists may sometimes not be interested in the mean effect size for the full domain of studies, but rather may be interested only in the specific effect sizes represented in the studies included in a meta-analysis (Hedges 1992; 1994a; 1994b; Raudenbush 1994; Shadish and Haddock 1994). Under this rationale, researchers do not regard the studies in their meta-analysis as a sample from a potentially larger population of studies, but rather as the entire universe of studies of interest. Under this assumption, there is no possibility of sampling error due to this sampling of study effect sizes, because all possible study effect sizes are by definition included in the meta-analysis. Overton (1998) showed that the FE model has the appropriate level of Type I error under this assumption. However, the key question is whether this assumption is ever realistic or appropriate.

The major problem with this assumption is that it is difficult (and perhaps impossible) to conceive of a situation in which a researcher would be interested only in the specific studies included in the meta-analysis and would not be interested in the broader task of estimation of the population effect sizes for the research domain as a whole. For example, suppose that 16 studies have been conducted relating the personality trait of Openness to Experience to job performance, but those studies have been published in widely varying journals. Suppose that as a result, the meta-analyst locates only 8 of the 16 studies to include in his or her meta-analysis. Consider this question in a hypothetical survey of personality researchers: Which would you as a researcher find more informative: (1) meta-analysis means and confidence intervals that generalize to the entire domain of studies in this research area (RE model results), or (2) meta-analysis means and confidence intervals that describe only the specific studies located for this meta-analysis and cannot be generalized to the research domain as a whole (FE model)? It seems clear that substantive researchers would (rightly) prefer the random effects results. Science is about generalization, and the purpose of research is the identification of generalizable conclusions (Overton 1998). Conclusions limited to a specific subset of studies are not scientifically informative.

Consider the same reasoning applied at the broader study domain level. Would a researcher rather have the results of a meta-analysis that describes only the first several studies conducted in a domain or the outcome of a meta-analysis that generalizes to all studies that could or might be conducted in the future in that domain? That is, would he/she prefer for the meta-analysis results and conclusions to generalize to future replication studies or would he/she prefer results that do not generalize to future replication? Most researchers would judge that conclusions about the broader study domain are of more scientific value. That is, the information produced by RE models is the information most researchers expect a meta-analysis to convey (Overton 1998), while the information produced by FE models is of very limited scientific value.

This rationale in support of the fixed effects model sprang from an analogy with FE models in analysis of variance (ANOVA). In ANOVA, a FE design is one in which all levels of the treatment that are of interest are included in the design, while a RE model in ANOVA is one in which only a sample of treatment levels of interest is included in the study. By analogy with this distinction in ANOVA, Hedges and Olkin (1985) labeled the two different approaches to meta-analysis as FE and RE models. Hence in FE meta-analysis models, the studies included in the meta-analysis are assumed to constitute the entire universe of relevant studies, whereas in RE models the studies are taken to be a sample of all possible studies that might be conducted or might exist on the subject. The National Research Council report (1992: 46 and 139) indicates that there are problems with this analogy. The report states:

> The manner in which the terms 'fixed effects' and 'random effects' are used in the meta-analysis literature is somewhat different from the classical definitions used in other techniques of statistics such as analysis of variance, where 'fixed effects' is the term required to deny the concept of a distribution of the true effects, $\delta_1 \ldots \delta_k$, and 'random effects' supposes that the are sampled from a population and therefore have a distribution. (ibid: 46)

As an aid in interpreting this statement, consider research on the effects of drugs on patients. A researcher might include the dosages 0 mg, 10 mg, and 20 mg. In FE ANOVA, these treatments (dosages) are fixed at these levels, the only ones considered of interest, and the idea that there is a naturally occurring distribution of dosages from which these three dosages are sampled is denied. This is different from the FE model in meta-analysis in two important ways. First, in meta-analysis the researcher does not specify (or fix) the parameter values (ρ_1 or δ_1) in the individual studies included in the FE meta-analysis. Instead, these are merely accepted as they happen to occur in the sample of studies. That is, they are merely observed and are not manipulated. The second difference flows from the first: Because the researcher does not fix the parameter values in the studies included in the meta-analysis, but rather merely accepts them as they happen to have occurred, there is no basis or rationale for postulating or assuming that these parameter values do not have a distribution across studies – the key assumption of the fixed model in ANOVA. These are the reasons why the National Research Council report (1992) rejected the analogy between FE models in ANOVA and FE models in meta-analysis. These considerations lead to the conclusion (stated earlier) that the FE model in meta-analysis is legitimate only in study sets in which S^2_ρ or $S^2_\delta) = 0$. Under these (rare) circumstances, study parameter values are indeed fixed – although all at the same value, unlike FE ANOVA designs. As discussed earlier, the National Research Council report concluded that whenever this condition is not met the FE meta-analysis model leads to elevated Type I error rates and unrealistically narrow confidence intervals.

Recently, Hedges and Vevea (1998: 488) have abandoned the rationale discussed here for the FE model in favor of the conclusion that there is no statistical rationale or justification for the FE model in meta-analysis, but that there can be a rationale based on subjective judgment by the researcher. They begin by acknowledging that FE results are of little value or interest if they cannot be generalized beyond the specific studies included in the meta-analysis, and they conclude that such generalization is 'not justified by a formal sampling argument.' However, they argue that a researcher can make a subjective 'extrastatistical' or 'extraempirical' judgment that generalization of FE estimates to the whole research domain is justified: 'Specifically, inferences may be justified if the studies are judged *a priori* to be sufficiently similar to those in the study sample' (ibid: 488). This judgment is the judgment that the new studies (to which generalization is to be extended) have study parameters (ρ_i or δ_i) that exactly reproduce, study for study, those in the original study set included in the FE meta-analysis. It is difficult to see how such a subjective judgment could be justified. What basis could a researcher have for such knowledge?

Although Hedges and Vevea provide computational examples of application of the FE model, they give no substantive example (even a hypothetical one) of a case in which such a subjective judgment would be appropriate as a rationale for use of the FE model. Nor do they provide any guidelines or suggestions for when this might be appropriate.

Hedges and Vevea recognize that the FE model in meta-analysis has no statistical justification, and this is a step forward. However, their attempt to provide a subjective rationale based on questionable judgments by researchers appears to be weak. If a procedure has no statistical justification, replacing it with one that does would seem preferable to constructing a subjective, non-statistical justification. The RE model has a clear statistical justification, requires no subjective judgments, and can be used in all applications of meta-analysis.

The Significance Test for the Mean Effect Size

Many researchers conduct significance tests on the mean effect size in meta-analysis. Fortunately, the total sample size in meta-analysis is usually large enough that the Type II error rate for the test in meta-analysis is lower than the average 50% to 60% Type II error rate in significance tests in single studies (Cohen 1962; 1994; Hunter 1997; Schmidt 1996; Sedlmeier and Gigerenzer 1989). Our focus here, however, is on Type I, rather than Type II, errors.

The Type I error rates for the RE and FE models for a variety of research domains are shown in Table 1. The methods used to compute Table 1 are presented in Appendix A. Study sample sizes in Table 1 vary from 25 to 1600. The second column in Panel B of Table 1 shows the homogeneous case ($SD_\rho = 0$). For the remaining columns, the domain is successively more heterogeneous, with SD_ρ varying from .05 to .25, values that cover most of the range of estimates of SD_ρ in the empirical research literature (Hunter and Schmidt 1990a).

Panel A in Table 1 indicates that the Type I error rate for the RE formula is always 5%, the requirement for a conventional significance test when alpha = .05. Panel B shows the error rates for the FE formula. Note first that for all sample sizes, the FE formula has a 5% error rate if the domain is homogeneous. The columns for the heterogeneous domains show higher error rates. The Type I error rate of the fixed effect formula becomes larger as the extent of heterogeneity increases and as the average sample size in the domain increases.

The first row in Table 1 represents domains in which the average sample size is N = 25; about the average sample size for studies of the effectiveness of psychotherapy (Stoffelmeyr, Dillavou, and Hunter 1983), for example. For $SD_\rho = .05$, the error rate is only 6% (i.e., only 20% larger than the nominal 5% rate). However, for $SD_\rho = .10$, the error rate climbs to 8% (60% larger than the 5% nominal rate). As heterogeneity climbs to $SD_\rho = .25$, the error rate climbs to 22% (four times larger than the 5% required for a conventional significance test).

The second row in Table 1 represents domains in which the average study sample size is N = 100; about the average for laboratory studies in psychology. In this case, the sampling error is large, but not so large that it overwhelms the variance of population effect sizes quantitatively. Even for a very small $SD_\rho = .05$, the error rate is 8% (i.e., 60% larger than the 5% required for a conventional significance test). For $SD_\rho = .10$, the error rate climbs to 16% (more than three times as large as the 5% required for a conventional significance test). As the heterogeneity climbs to $SD_\rho = .25$, the Type I error rate climbs to 46%.

The third and fourth rows of Table 1 represent domains for survey research; i.e., study sample sizes ranging from N = 400 to N = 1600. For these cases, sample size is larger than for typical studies in psychology, and many might intuitively expect any statistical procedure to work better for such large sample studies than for the more typical small sample study domain. Actually, the opposite is true. The FE formulas are less accurate for large sample study domains than for small sample study domains. When N

Table 1: Type I error rates for the random effects and the fixed effects significance test for the mean correlation in meta-analysis (nominal alpha = .05)

Panel A. The RE significance test						
Prob (Type I error) = .05 in all cases						
Panel B. The FE significance test						
Study Sample Sizes	Homogeneous Case ($SD_\rho = 0$)	Heterogenous cases ($SD_\rho > 0$) SD_ρ .05	.10	.15	.20	.25
25	.05	.06	.08	.11	.16	.22
100	.05	.08	.16	.28	.38	.46
400	.05	.17	.38	.53	.63	.70
1600	.05	.38	.63	.75	.81	.85
. . .						
∞	.05	1.00	1.00	1.00	1.00	1.00

Note: RE = Random Effects; FE = Fixed Effects. SD_ρ = the standard deviation of population correlations across the studies included in the meta-analysis.

= 400, even for the very small of SD_ρ .05, the error rate is 17%. For SD_ρ = .10, the error rate climbs to 38%. As the heterogeneity climbs to SD_ρ = .25, the Type I error rate climbs to 70%.

For average study sample sizes larger than N = 400, the Type I error rate for FE model becomes very extreme. Consider a very small degree of heterogeneity, SD_ρ = .05. When studies each have N = 1600, the Type I error rate ranges from 38% to 85%. As the average sample size becomes infinite, the Type I error rate for the FE formula climbs toward 100%.

The FE significance test for mean effect size presented by Rosenthal (1978) differs from the FE tests advanced by Hedges and Olkin (1985, ch. 7). Rosenthal (1978) suggested that 'combined p-value' methods be used to determine the significance of mean effect sizes. However, as shown in Appendix B, these methods also assume homogeneous study population δs or ρs and also have elevated Type I error rates whenever this assumption is not met. So the analysis of Type I error rates presented in Table 1 applies to combined p-value methods, as well as to other FE significance tests. (Criticism of combined p-value methods on other grounds (Becker 1987; National Research Council 1992: 178–80) has led to a recent decline in their use in published meta-analyses.)

Effect on Confidence Intervals

In applying both FE and RE meta-analysis methods, Hedges and Olkin (1985) suggest placing confidence intervals around the mean effect size rather than testing the mean effect size for statistical significance. Others have also concluded that confidence intervals are to be preferred over significance tests (Cohen 1994; Hunter 1997; Loftus 1996; Schmidt 1996). However, as in the case of significance tests, confidence intervals will typically be inaccurate when based on FE formulas for the standard error of the mean effect size. Because it omits the effects of variation in study population effect sizes (see Appendix A), the FE standard error of the mean underestimates the actual standard error, resulting in confidence intervals that are erroneously narrow. The result is overestimation of the precision of meta-analysis findings. This problem is not hypothetical; it occurs with some frequency in the meta-analysis research literature.[5]

This problem is illustrated in Table 2. This table applies to the situation in which the mean population correlation is zero (i.e., μ_ρ = 0). (Results are more extreme when μ_ρ departs from zero in either direction.) Panel A of Table 2 illustrates the underestimation of the standard error of the mean correlation (or standard error of the mean d value) by FE formulas for heterogeneous study domains. This panel presents the ratio of the FE standard error to the actual standard error (× 100) for varying values of SD_ρ and study sample size.

When SD_ρ = 0 (that is, when the study domain is homogeneous), the FE standard error is 100% of the actual standard error; that is, there is no underestimation of the standard error. However, as SD_ρ becomes larger, the FE formula increasingly underestimates the actual standard error. For example, when studies each have N = 100, the estimated standard error is only 89% as large as the actual standard error when SD_ρ = .05. As SD_ρ increases beyond .05, underestimation becomes more severe: 69%, 43%, 31%, 23%, and 19%. That is, as study heterogeneity becomes greater, the FE SE formula, which assumes no heterogeneity exists, produces increasingly inaccurate estimates of the of the mean effect size.

Likewise, holding SD_ρ constant, the FE estimates become increasingly inaccurate as study sample size increases. For example, consider the column SD_ρ = .10. As study Ns increase from 25 to 1600, the $FE_{\bar{r}}$ estimate declines from 89% of the actual value to only 23% of the actual value. As N becomes increasingly large, primary sampling error becomes progressively smaller, and the effect of heterogeneity of the study domain (SD_ρ) becomes increasingly more important.

Panel B of Table 2 shows the effect of this underestimation of the on confidence intervals placed around the mean effect size. In meta-analysis, as in other applications, construction of confidence intervals is one of the most important uses of the SE of the mean. Probably the most commonly used confidence interval (CI) is the 95% CI. Panel B of Table 2 shows the actual confidence levels of CIs that are obtained with FE methods when the nominal CI is the 95% CI. First, note that when SD_ρ = 0 (i.e., in the homogeneous case), the nominal and actual CIs are the same. That is, a researcher intending to compute the 95% CI will, in fact, obtain a 95% CI.

However, as discussed earlier, homogeneity is rare and may be essentially nonexistent. If the studies are heterogeneous, the computed ostensible 95% CI will in fact have a lower confidence level. That is, the computed CI will be too narrow. For example, if studies each have N = 100 and SD_ρ = .10, then the computed CI – believed by the researcher to be the 95% CI – will in fact be the 83% CI. That is, the meta-analyst will report that one can be 95% confident that the interval contains the actual (population) mean value – when, in fact, one can only be 83% confident of this. As study Ns become larger, the CIs become even less accurate: 60% and 35% here.

Likewise, as SD_ρ becomes larger the computed CI becomes increasingly inaccurate for a given sample size. For example, holding study

Table 2: Under-estimation of standard error of the mean ($SE_{\bar{r}}$) and of confidence interval widths by fixed effects models

Panel A. Under-statement of standard error of $\bar{r}$ ($SE_{\bar{r}}$) by the FE model when $\mu_\rho = 0$						
Study Sample Sizes	Homogeneous Case ($SD_\rho = 0$)	Heterogenous cases ($SD_\rho > 0$) SD_ρ .05	.10	.15	.20	.25
25	100	.97	.89	.79	.70	.62
100	100	.89	.69	.54	.43	.36
400	100	.69	.43	.31	.23	.19
1600	100	.43	.23	.16	.12	.10
. . .						
∞	100	0	0	0	0	0
Panel B. Actual confidence levels of CIs for nominal FE 95% confidence interval when $\mu_\rho = 0$						
Study Sample Sizes	Homogeneous Case ($SD_\rho = 0$)	Heterogenous cases ($SD_\rho > 0$) SD_ρ .05	.10	.15	.20	.25
25	95	.94	.92	.88	.83	.77
100	95	.92	.83	.71	.61	.52
400	95	.83	.60	.45	.35	.29
1600	95	.60	.35	.24	.18	.15
. . .						
∞	95	0	0	0	0	0

Notes: FE = Fixed Effects. Values shown are FE $SE_{\bar{r}}$ as a percentage of actual $SE_{\bar{r}}$.
SD_ρ = the standard deviation of population correlations across the studies included in the meta-analysis.

Ns constant at N = 100, the computed nominal 95% CI is the 92% CI when $SD_\rho = .05$, the 83% CI when $SD_\rho = .10$, on to the 52% CI when $SD_\rho = .25$. In all cases, these confidence intervals would be falsely interpreted as 95% confidence intervals, exaggerating the precision of our cumulative knowledge base.

Hence, when FE methods are used, confidence intervals, like significance tests, can be quite inaccurate when the studies in the meta-analysis are heterogeneous even to a small or moderate degree. As noted by the National Research Council (1992), some degree of heterogeneity of studies in meta-analysis is probably nearly universal. Many published meta-analyses employ FE methods to compute confidence intervals around mean effect sizes, as noted earlier. These CIs are almost certainly too narrow and almost certainly overestimate the precision of the meta-analysis findings (National Research Council 1992). On the other hand, CIs based on RE formulas do not have this problem and are accurate.

Comment on Number of Studies in the Domain

Our analysis has not mentioned of the number of studies included in the meta-analysis. Since the total amount of data in a meta-analysis is proportional to the number of studies, one might think that the fixed effect formula would work better for a meta-analysis conducted on a large number of studies than for a meta-analysis on a small number of studies. However, this is not the case. The size of the inconsistency in the fixed effect formula is the same regardless of the number of studies in the meta-analysis.

The reason for this stems from the fundamental problem with the FE formula: the underestimation of the sampling error in the mean sample correlation (see Appendix A).

The actual sampling variance for $\bar{r}$ is:

$$V_{e_{\bar{r}}} = V_r/K = (V_\rho + V_e)/K$$

But the FE estimate of this sampling variance is:

$$V_{e_{\bar{r}}} = V_e/K = (0 + V_e)/K$$

where K is the number of studies; V_ρ is the variance of population correlations; and V_e is the average sampling error variance within studies (see Appendix A).

To illustrate the extent of this under-estimation, one can look at the ratio of the FE sampling error variance estimate to the actual sampling error variance:

$$\frac{\text{Fixed effects estimate}}{\text{Actual variance}} = \frac{\frac{V_e}{K}}{\frac{V_\rho + V_e}{K}} = \frac{V_e}{V_\rho + V_e}$$

Note that in this ratio the constant K (the number of studies) cancels out. That is, the ratio is the same for any number of studies in the domain.

Assessment of Potential Moderator Variables

Effect on Significance Tests for Moderators

In addition to significance tests (and confidence intervals) for mean effect sizes, meta-analyses often report significance tests for hypothesized moderator variables. Often, for example, the full set of studies is broken out into subsets based on an hypothesized moderator variable and tests are run to determine whether these subset means are significantly different from each other. For example, one subset may be based on high school students and the other on college students. Within each subset of studies, the study population values (ρs and δs) can be either homogeneous (identical in value) or heterogeneous (differing in value).

Table 3 shows the actual Type I error rate for the RE and FE formulas for a variety of research domains. The methods used to compute Table 3 are presented in Appendix C. Again, the average study sample size in a research domain is varied from 25 to 1600, as noted in the first column of Table 3. The homogeneous domain ($SD_\rho = 0$) is illustrated in the second column of Panel B in Table 3. For the remaining columns, the domain is successively more and more heterogeneous; with SD_ρ varying from .05 to .25, the same range of values examined in Tables 1 and 2. The mean population correlation for the domain is assumed to be $\mu_\rho = .20$.

The general pattern of findings in Table 3 is identical to the pattern for Table 1. The FE formula is slightly more inaccurate for the moderator test than for the general test. However, this is merely because the mean correlation for the domain is $\mu_\rho = .20$ instead of $\mu_\rho = .00$. For higher values of the mean correlation, the FE model shows even larger Type I error rates than those in Table 3. This is because as the mean correlation becomes larger, the primary sampling error becomes smaller (i.e., larger correlations have less sampling error). Underestimation of the standard error by the FE model is greater the smaller sampling error variance is relative to the variance of population correlations. This is the same pattern observed in Table 2.

As shown in Panel A in Table 3, the Type I error rate for the RE formula is always 5%; the requirement for a conventional significance test at $\alpha = .05$. Panel B shows the error rates for the FE formula for the moderator significance test. If the domain is homogeneous, the fixed effect formula has a 5% error rate for all sample sizes. As noted earlier, the term homogeneous here refers to the study population parameter correlations across studies *within* each of the two moderator categories (e.g., the studies conducted on males vs. those conducted on females).

In general, homogeneity of population correlations within hypothesized moderator groups will occur rarely, if ever, in real data.

Table 3: Type I error rates for the random effects and fixed effects significance test for moderators in meta-analysis (nominal alpha = .05)

Panel A. The RE significance test

Prob {Type I error} = 5% in all cases

Panel B. The FE significance test

Study Sample Sizes	Homogeneous Case ($SD_\rho = 0$)	Heterogenous cases ($SD_\rho > 0$) SD_ρ .05	.10	.15	.20	.25
25	.05	.06	.08	.12	.17	.23
100	.05	.08	.17	.29	.39	.48
400	.05	.17	.40	.55	.65	.71
1600	.05	.40	.65	.76	.82	.85
. . .						
∞	.05	1.00	1.00	1.00	1.00	1.00

Note: RE = Random Effects; FE = Fixed Effects. SD_ρ = The standard deviation of population correlations across the studies within each of the two moderator categories. $\mu_\rho = .20$ in each moderator category. Hence, $\mu_{\rho 1} - \mu_{\rho 2} = 0$, and the Null Hypothesis holds.

Homogeneity of study population correlations within sub-grouped study sets requires three conditions for its occurrence. First, there must be only one substantive moderator. (If there is more than one moderator, then the subsets of studies logically must be internally heterogeneous.) Second, the researcher must correctly hypothesize and test for this moderator and not for some other potential moderator. Third, there must be no heterogeneity of study population correlations stemming from differences between studies in amount of measurement error in measures used, in range variation, in dichotomization of measures, or in other methodological factors that affect study population correlations (Osburn and Callender 1988). Thus homogeneity of study population correlations within moderator sub-grouped study sets is unlikely to occur with any frequency.

All columns for the heterogeneous domains in Table 3 show error rates higher than 5%. The inconsistency of the fixed effect formula becomes larger as the extent of heterogeneity increases within the hypothesized moderator categories and as the average study sample size in the domain of studies increases. When the average study sample size is N = 25, primary sampling error is large relative to the variance of population effect sizes. For SD_ρ = .05, the error rate is only 6%. However, for SD_ρ = .10, the error rate climbs to 8% (60% larger than the 5% nominal rate). As the heterogeneity climbs to SD_ρ = .25, the Type I error rate climbs to 23%.

When the studies in the meta-analysis each have sample size of 100, the primary sampling error is large, but not so large that it overwhelms the variance of population effects within moderator categories. Even for a very small SD_ρ = .05, the error rate is 8%. For SD_ρ = .10, the error rate climbs to 17% (more than three times as large as the 5% nominal rate). As the heterogeneity climbs to SD_ρ = .25, the Type I error rate climbs to 48%.

Again, the third and fourth rows of Table 3 represent study domains for survey research; sample sizes range from N = 400 to N = 1600. As was the case in Table 1 and 2, the fixed effect formula is less accurate for large sample study domains than for small sample study domains. Consider the row for N = 400. Even for a small SD_ρ = .05, the error rate is 17%. For SD_ρ = .10, the error rate climbs to 40%. As the heterogeneity climbs to SD_ρ = .25, the error rate climbs to 71%. If the average sample size in a domain is as high as N = 400, then the Type I error rate for the FE formula in a heterogeneous domain is always more than three times as high as the 5% required for a conventional significance test. For average study sample sizes larger than N = 400, the FE formula becomes very extreme. With study sample sizes of 1600, instead of a 5% error rate, the Type I error rate climbs from 40% to 85% as study domain heterogeneity increases. As the average sample size becomes infinite, the Type I error rate for the FE formula climbs to 100%.

The practical implication of the findings in Table 3 is that the current widespread use of FE tests may be resulting in the 'detection' of many moderator variables that do not in fact exist. Acceptance of non-existent moderators (interactions) as real leads (among other things) to the belief that the psychological phenomena studied are more complex than they in fact are. This results in misunderstanding of the meaning of research literatures and therefore retards the development of cumulative scientific knowledge. Hence this is a potentially serious problem. (Our discussion of the findings in Table 3 makes no mention of the number of studies contained in the meta-analysis. The reason is the same for the moderator test as for the test applied to mean effect sizes: The term *K* (the number of studies) cancels out in the ratio that measures the underestimation of the sampling error standard deviation by the FE model, as noted earlier.)

Effect on Confidence Intervals for Moderators

Although they are used less frequently than significance tests in the moderator literature, confidence intervals can be employed in examining moderator variables. In meta-analysis, the relevant confidence interval is the interval around the difference between the mean effect sizes in the moderator sub-groups. For example, we can determine the difference between the mean correlation for men and the mean correlation for women; we can then place a confidence interval around this difference to determine the precision of this estimate. In heterogeneous domains, all such confidence intervals computed using FE formulas will be erroneously narrow, just as they are when placed around mean effect sizes. That is, FE confidence intervals for moderator variables overstate the precision of the estimate of the difference between the sub-group mean d or mean r values. That is, they will overestimate the accuracy of the moderator findings. Hence, the inaccuracy problems associated with FE analyses of moderator variables cannot be avoided by employing confidence intervals in place of significance tests. On the other hand, confidence intervals based on the RE model do not have this deficiency and are accurate.

Conclusion

Two significance tests are frequently applied to mean correlations and mean d-values in meta-

analysis: a test of the significance of the overall mean effect size in a study domain and a moderator test comparing means across subsets of studies. The traditional RE significance tests for these purposes were derived for meta-analysis from the directly comparable tests in elementary statistics; i.e., the t test and the F test for comparing means. Hedges and Olkin (1985) and Rosenthal and Rubin (1982a,b) have also presented significance tests that derive from the FE meta-analysis model. This article has shown that in virtually all study domains the FE tests are not conventional significance tests. In a conventional significance test, the Type I error rate equals the nominal alpha level. The FE tests meet this requirement only if they are used in a homogeneous study domain. If the study domain is heterogeneous, as virtually all study domains are, then the error rate for the FE formulas is higher than 5%. This study is the first to quantitatively calibrate the *magnitude* of the increase in the Type I error rate and the bias in confidence intervals for analysis of both means and moderator variables. In both cases, these errors are considerably larger than most researchers have imagined them to be. They are large enough to produce serious distortions in research conclusions in the literature based on meta-analyses.

We conclude that FE models and procedures are rarely, if ever, appropriate for real data in meta-analyses and that therefore RE models and procedures should be used in preference to FE models and procedures. This is the same recommendation made by the National Research Council (1992). Up until now, the recommendations of the National Research Council Report have apparently had little effect on actual practice. We hope that the findings of this research will help to change that.

Appendix A: Methods and Concepts Underlying Table 1

The Homogeneous Domain

For a homogeneous domain, the population effect size is uniform across studies and observed study outcomes (rs or ds) differ only because of ordinary sampling error, which we will label primary sampling error. Consider the correlation as a measure of effect size. Let r_i = the sample correlation for study i; ρ_i = the population correlation for study i; e_i = the sampling error for study i. In general, $r_i = \rho_i + e_i$, but in the homogeneous domain, $SD_\rho = 0$ and $\rho_1 = \rho_2 = \rho_3 = \ldots = \rho$, so $r_i = \rho + e_i$.

For a homogeneous domain, the average population effect size is the same as 'the' population effect size because there is only one effect size ρ. Thus to estimate the mean population effect size is to estimate 'the' population effect size.

The mean correlation across studies. Given the sample correlations from K studies, the average sample correlation across studies is related to the average population correlation by the following formula:

$$\bar{r} = \bar{\rho} + \bar{e} = \rho + \bar{e}$$

Thus the sampling error in the average correlation is the average sampling error. This average sampling error is minimized if the average used is a frequency (sample size) weighted average of the study effect sizes. If a frequency weighted average is used, then the sampling error in $\bar{r}$ in a homogeneous domain of studies is $V_{\bar{e}} = V_e/K$, where V_e is the sample size weighted average sampling error for the K individual studies. The critical ratio significance test is $Z = \bar{r}/SD_{\bar{e}}$ (two tailed test 5% comparison value: $Z = 1.96$ (where $SD_{\bar{e}} = \sqrt{V_{\bar{e}}}$). For homogeneous domains, this critical ratio formula is both the RE formula and the FE formula. If the null hypothesis is correct, this formula will falsely register as significant only 5% of the time at alpha = .05.

The Heterogeneous Domain

In the heterogeneous domain, the study population correlations vary from study to study; i.e., $V_\rho > 0$. For example, Barrick and Mount (1991) found in a large meta-analysis that for the personality trait of Openness to Experience, the distribution of population correlations with job performance has a mean of .04 and a standard deviation of .16; i.e., $V_\rho = .16^2 = .0256$. This variance is not due to sampling error and does not depend on sample size in the studies. It is an estimate of variability in the actual underlying study population values of ρ_i. This variability exists even if N were infinite in every study.

For any study conducted in this domain, with probability .68, the population correlation for that study would fall in the range between −.12 and +.20. With probability .16, the study population correlation could be lower than −.12, and with probability .16, the study population correlation could be higher than +.20. Consider the values of the population correlation (not the observed correlation) in the first 4 studies conducted in this domain. Each study population correlation will differ randomly from the domain average of .04. Thus the mean study population average for the meta-analysis of those 4 studies would differ from the domain average of .04. The variance of the average is the variance of a single study value divided by the number of studies. In each study $V_\rho = .0256$

and $SD_\rho = .16$. For the average of 4 studies $V_{\bar\rho} = V_\rho/4$ and $SD_{\bar\rho} = SD_\rho/2$. In this case, $V_{\bar\rho} = .0064$ and $SD_\rho = .08$. That is, the average population value (not the observed value) across the first four studies will vary with a mean of .04 and a standard deviation of .08. This means that with probability .68, the mean population correlation will be between $-.04$ and $+.12$. However, with probability .16, the mean population correlation for those four studies could be smaller than $-.04$, and with probability .16, the mean population correlation for those four studies could be larger than $+.12$.

Unless the number of studies is extremely large, the average population effect size in a meta-analysis is subject to random variation because of the variation in population effect sizes in that study domain. This variation is not due to primary sampling error; it occurs even if every study in the meta-analysis has an infinite sample size. This variation must be considered and included when computing either confidence intervals or a significance test for the mean population correlation (Becker 1996; Osburn and Callender 1992; Schmidt, Hunter and Raju 1988).

Let the domain mean population correlation be symbolized μ_ρ. Then the mean for K studies is $M_\rho = \mu_\rho + e'$, where e' is a sampling error with $SD_e = SD_\rho/\sqrt{K}$.

We now consider real studies with finite sample sizes. That is, we now introduce primary sampling error. We then compute the sampling error in the mean observed (sample) study correlation as an estimate of the domain mean population correlation.

For each study i, $r_i = \rho_i + e_i$, where e_i is the primary sampling error. Averaging across studies, $\bar r = \bar\rho + \bar e = \mu_\rho + e' = \mu_\rho + (e' + \bar e)$. Sampling error in the heterogeneous case can now be seen to have two components: e' and $\bar e$. The variance of sampling error also has two components: $V_{e'} + V_{\bar e}$.

The difference between the fixed effect formulas and the RE formulas can now be simply stated: the FE formulas do not include the error component e'. The fixed effect formula has only the component $V_{\bar e}$ and does not have the component $V_{e'}$. That is, the FE formula assumes that is zero. This assumption is met for the homogeneous domain but not met for the heterogeneous domain.

Thus the FE model is a special case of the more general RE model (National Research Council 1992: 139). The special case represented by the FE model is the case in which it is known *a priori* that the study population parameters do not vary from study to study. Hence the National Research Council Report states that the FE model should be used only if one has 'significant prior information' that study population parameters are constant across studies (ibid: 143 and 147). Otherwise, one should use the more general RE model 'to reduce the likelihood of overstating precision' of findings (ibid: 143). However, it is rare that one has such prior information. Because its statistical power is frequently low, the chi-square based test of homogeneity cannot be relied on to make this determination. The National Research Council report states that the major reason for preferring RE models is the low statistical power of homogeneity tests (ibid: 52).

Suppose that the null hypothesis in our example is true and the mean correlation between Openness to Experience and job performance is not .04 but .00. With this assumption, we can now examine Type I bias in the FE model. The standard deviation is still assumed to be .16. The critical region for the significance test is determined by the 95% two-sided probability interval for the average sample correlation. The null hypothesis for openness is $\mu_\rho = 0$. For any study i, primary sampling error is then $V_{e_i} = 1/(N_i - 1)$. Let V_e be the average of these values across studies. Sampling error variance for the mean sample correlation is then:

$$\text{RE formula: } V = V_{e'} + V_{\bar e} = .16^2/K + V_e/K$$
$$\text{FE formula: } V = V_{e'} = V_e/K$$

For example, let $K = 4$ and let all $N_i = 65$.

$$V_\rho = .16^2 = 0.256 \text{ and } V_e = 1/64 = 0.15625$$

The RE formula for the sampling error variance of the mean observed correlation then yields:

$$V = .0256/4 + .015625/4 = .01030625$$
$$SD_{e_{\bar r}} = \sqrt{.00390625} = .1015$$

The FE formula for the sampling error variance of the mean observed correlation yields:

$$V = .015625/4 = .00390625$$
$$SD_{e_{\bar r}} = \sqrt{.00390625} = .0625$$

The 95% confidence intervals $[\mu_\rho \pm 1.96(SD_{e_{\bar r}}) = 0 \pm 1.96(SD_{e\bar r})]$ are then:

$$\text{RE formula: } -.20 + .20$$
$$\text{FE formula: } -.12 + .12$$

The Hedges and Olkin (1985, ch. 7) FE methods emphasize confidence intervals over significance tests performed on the mean effect size. As explained in the main text, we believe this emphasis on confidence intervals is laudable. However, as this example shows, use of FE standard errors of the mean to compute confidence intervals leads to confidence intervals that are erroneously narrow.

The fixed effect formula underestimates the

sampling error standard deviation for the mean sample correlation. In our example, SD_{e_r} is .1015 but is estimated to be .0625 by the FE model. If the FE standard error were correct, then only 5% of sample values would be more extreme than ±.12. But in our example more than 5% of values are outside the central region.

Consider the upper critical region; the region that is assumed to have only a 2.5% chance of occurring.

$$\begin{aligned}\text{Prob}\{\mu_\rho > .12\} &= \{\text{Prob } z > .12/SD_{e_r} \\ &= .12/.1015\} \\ &= \text{Prob}\{z > 1.18\} = 12\%\end{aligned}$$

That is, the effective critical value for this significance test is not 1.96 but 1.18. The upper critical region thus has a rate of 12% instead of 2.5%. As a result, the overall Type I error rate for this test is 24% rather than 5%. The resulting confidence interval is not the nominal expected 95% interval, but is instead the 76% confidence interval.

Appendix B: FE Significance Tests Based on Combined p-Values

Combined p-values: Summed z Value

Researchers using the methods presented by Rosenthal and Rubin typically use any of a variety of techniques to test the mean correlation for significance. These techniques are from Rosenthal's (1978) article advocating the combined *p* value as a way of testing the global null hypothesis for effect sizes. This appendix briefly discusses the more important of these methods. Rosenthal and Rubin use the 'z' notation in two different ways. When the effect size is measured using a correlation, they use 'z' for the Fisher z transformation of the correlation. However, in other contexts, they use 'z' for the critical ratio of a significance test. It is as a critical ratio that they use 'z' in their discussion of combined p values.

The two methods that Rosenthal (1978) recommended as having the lowest Type II error rates are the methods that use 'summed z values'. These methods were developed by Mosteller and Bush (1954) from earlier work by Stouffer, Suchman, De Vinney, Star and Williams (1949). In both methods, each effect size is used to generate a one-tailed significance test. The p value for this test is then converted to a standardized normal deviate denoted 'z' and the z values are either summed directly or used to compute a weighted sum.

Although Rosenthal and Rubin (1982b) do not note it, there is a direct connection between the summed z values and the average correlation. Consider the direct sum of z values. If the effect size is measured by a correlation, then we need not compute p values in order to convert that treatment correlation to a significance test z value. The classic z test (critical ratio) for the correlation can be written:

$$z = \sqrt{(N-1)}r$$

The summed z value is thus a weighted sum of the correlations. If we use the symbol w_i for the weight to be given to the results from study I, then:

$$w_i = \sqrt{(N_i - 1)}$$

The summed z value is then computed to be:

$$\text{Sum } z_i = \text{Sum } w_i r_i$$

Dividing by the sum of the weights converts this to a weighted average:

$$\text{WtAve}(r) = \Sigma w_i r_i / \Sigma w_i$$

The significance test on the summed z value is obtained by noting that under the global null hypothesis, the summed z value is the sum of independent unit normal deviates and thus has variance *K* where *K* is the number of studies. The critical ratio for the summed z value is thus:

$$CR = \text{Sum} z_i / \sqrt{K}$$

Exactly this same test is easily computed for the weighted average correlation using:

$$CR = \alpha \text{WtAve}(r),$$

where $\alpha = (\Sigma w_i)/\sqrt{K}$.

In summary, the simplest method of combining p values is to compute the sum of z values for each study where 'z' is the critical ratio normal deviate for a one tailed significance test. This method is identical to using a weighted average correlation where each study is weighted by the square root of sample size.

We can compare the summed z value to the conventional test on the mean correlation by asking the following questions: Does the use of square root weighting change the basic assumptions of the test? And, does it have higher or lower Type II error?

The combined p value approaches all assume the null hypothesis to be the global null hypothesis; i.e., these methods assume that the study population correlations are uniformly zero. Thus the summed z test assumes a homogeneous domain and has a Type I error rate of 5% only if the domain is homogeneous – the same problem that exists other fixed effect formulas.

Even if the study domain is homogeneous, this combined p value test has less than optimal statistical power to detect a nonzero population.

Consider the hypothesis that the study domain is homogeneous but the domain correlation is not zero. Against this hypothesis, the ideal weight for each correlation is not the square root of sample size but rather the sample size itself (for simplicity, we ignore the use of N rather than $N - 1$). Thus the summed z test will have lower power than the usual test of the significance of the frequency weighted mean r. The worst possible case occurs when there is one large sample study and many small sample studies. For example, if we have 16 studies with sample size 26 and one study with sample size 401, then the standard error for the square root weighted average will be 51% larger than the standard error for the ordinary sample size weighted average and the summed z test will have its lowest power relative to test on the frequency weighted mean r.

We know of no situation where the square root weights would be equal to or superior to the simple sample size weights in terms of sampling error. When evaluated for power, the sum of z values appears to be an inferior significance test under all conditions.

Combined p-values: Weighted Sum of z Values

Rosenthal (1978) also noted that Mosteller and Bush (1954) suggested that the simple sum of z values could be replaced by a weighted sum of z values. In particular, they recommended the use of a frequency weighted z value. That is, they recommended:

$$\text{Weighted sum of } z = \Sigma w_i z_i, \text{where } w_i = N_i$$

This weighed sum is also the numerator for a weighted average correlation. Consider the weights:

$$w_i = N_i\sqrt{(N_i - 1)}$$

which differs only trivially from

$$w_i = N_i^{1.5}$$

$$\text{Weighted sum of z } = \Sigma w_i r_i$$

The corresponding weighted average correlation is thus obtained by dividing the weighted sum of z by the sum of the weights for the correlations:

$$\text{WtAve}(r) = \Sigma w_i r_i / \Sigma w_i$$

The significance test for the weighted sum of z is obtained by dividing by its standard error. Assuming the global null hypothesis, the sampling error variance is:

$$\text{SEV of Weighted sum of } z = \Sigma N_i^2$$

Exactly that same significance test is run by dividing the weighted average correlation by its standard error. The sampling error variance of the weighted average r is:

$$\text{SEV of Weighted sum of } r = \Sigma N_i^2 / (\Sigma w_i)^2$$

Note that this technique also assumes the global null hypothesis. Thus it assumes not only that the mean population correlation is zero, but also that the domain is homogeneous. If the mean correlation is zero but the domain is not homogeneous, then this technique will not have an alpha of 5% but will have a larger Type I error rate. Hence it has the problems discussed in the text of this article for FE models.

The Type II error rate of the weighted sum of z will also be larger than the Type II error rate of the ordinary weighted average correlation. The optimal weight for the homogeneous domain is the sample size itself, not the 1.5^{th} power of the sample size. While the simple sum of z gives too little weight to large studies, the weighted sum of z gives too much weight to large studies. As in the story of Goldilocks and the three bears, it is the medium weights of the conventional frequency weighting that are 'just right'.

Others have also discussed problems with combined p-value statistics. Perhaps the most comprehensive treatments are Becker (1987) and National Research Council (1992: 174–80). As a result of these problems, the National Research Council (ibid: 182) recommended that the use of combined p-value methods in meta-analysis 'be discontinued'. A scanning of the literature does in fact suggest that there has been a decline in the use of these methods in meta-analysis.

Appendix C: Methods and Concepts Underlying Table 3

The Homogeneous Domain

The FE formula has the appropriate Type I error rate if the overall domain is homogeneous; the Type I error rate will be the 5% required for a conventional $\alpha = .05$ significance test.

The null hypothesis for a moderator variable significance test is that the potential moderator is in fact not a moderator variable. That is, the null hypothesis is that the mean effect sizes do not differ from one another across the domain of studies. Is this assumption of homogeneity reasonable?

If a domain is homogeneous, there can be no true moderator variables for that domain. Every potential or hypothesized moderator is in fact not a moderator variable. The null hypothesis would be true for every potential moderator variable. But if a meta-analyst knows that the domain is homogeneous, he or she would not consider

testing for moderator variables for that domain. This is consistent with the actual practice of many meta-analysts. Many meta-analysts conduct a chi square test for homogeneity on the complete set of studies. If that test is not significant, they conclude that the domain is homogeneous and do not test further for potential moderator variables. Only if the test is significant do they present analyses for potential moderator variables. That is, in current practice, most meta-analysts do not test a potential moderator hypotheses unless they have evidence from the homogeneity test that the domain is not homogeneous. That is, many meta-analysts who use the FE formula to test their moderator hypotheses already have evidence indicating that the domain is heterogeneous. If the FE formula is inappropriate for the heterogeneous domain, then it is inappropriate in virtually all current applications of that formula to the analysis of hypothesized moderators. We show that this is, in fact, the case.

The Heterogeneous Domain

If a study domain is heterogeneous, then study population correlations vary across the domain and there must be some explanation for the variation. Suppose that there is a true moderator variable and it is binary in nature. That is, ρ_i assumes only two values. What is the Type I error rate for the FE model? If the null hypothesis is true, then $\mu_{\rho 1} = \mu_{\rho 2}$ for any false moderator. The FE significance test for $\mu_{\rho 1} = \mu_{\rho 2}$ assumes that $SD_{\rho 1} = SD_{\rho 2} = 0$ That is, it assumes there is no variation in study population parameters within the two moderator study subsets. If this condition is not met, the FE test has an elevated Type I error rate. Whenever there is a real moderator within the study subsets, it will cause $SD_{\rho 1}$ and $SD_{\rho 2}$ to be greater than zero, creating a Type I bias in the FE test.

Because this point is important, we illustrate it with an extended (hypothetical) example. As noted in an example in the main text, there is considerable variation in the population correlation between Openness to Experience and job performance; i.e., a standard deviation of $SD_\rho = .16$ (Barrick and Mount 1991). Suppose that this variation is explained by some binary variable. For example, suppose that in autocratic organizations, inquisitiveness is regarded as questioning of authority and that high Openness workers are therefore regarded as 'attitude problems' and thus receive lower performance ratings. On the other hand, assume that in democratic organizations, inquisitiveness is positively regarded and encouraged. High Openness workers are regarded as 'showing initiative' and thus receive higher performance ratings. This means that openness will have a positive correlation with performance ratings in a democratic organization but will have a negative correlation in an autocratic organization.

Consider a hypothetical example that matches the findings for the population correlation between Openness to Experience and job performance: $\mu_\rho = .04$ and $SD_\rho = .16$. Assume for convenience that there is a 50–50 split between autocratic and democratic organizations. Now suppose that in autocratic organizations $\rho = -.12$ and in democratic organizations $\rho = +.20$. Hence we have a binary moderator variable.

The null hypothesis for the moderator significance test is that we are studying some potential moderator variable that is not in fact a true moderator. In our binary example, the null hypothesis assumes that if the studies are broken out into two groups based on the moderator hypothesis, the mean correlation will be the same in both groups. Let us consider a false moderator (i.e., one for which $\mu_{\rho 1} = \mu_{\rho 2}$). Because any potential moderator variable that is correlated with organization climate will be spuriously correlated with study outcome, a false moderator variable must be uncorrelated with the true moderator. So to illustrate the null hypothesis we need a potential moderator variable that is uncorrelated with organization climate. Let that potential moderator variable be sex of the employee. Some theorists might hypothesize that inquisitive questioning will be encouraged from men but discouraged from women (i.e., men are allowed autonomy while women are expected to be compliant). Thus a meta-analyst faced with the finding that this study domain is heterogeneous ($S^2_\rho > 0$) might predict that sex of worker will be a moderator variable for this domain. For purpose of our example, we assume that this hypothesis is false.

Suppose then that the reviewer locates a subset of studies in which the workers were all women and another subset of similar size in which the workers were all men. The reviewer predicts that sex will be a moderator variable and expects to find a positive correlation in the studies on male workers and a negative correlation in the studies on female workers. If the null hypothesis is true, then sex of worker is not a moderator variable, and the mean correlation will be the same in each sub-domain ($\mu_{\rho 1} = \mu_{\rho 2}$).

Assume then that half the studies with women workers were conducted in autocratic organizations while the other half were conducted in democratic organizations. Assume the same for studies with men workers. Then in each sub-domain, we have

$$\text{Male workers: } = +.04, SD_\rho = .16$$

$$\text{Female workers: } = +.04, SD_\rho = .16$$

We have a false potential moderator that meets the assumptions for the null hypothesis of the moderator significance test. Within the two subgroups, μ_ρ and $SD_\rho = .16$, the same values as in the total group of studies.

Up to this point, this analysis is at the level of study population parameters (ρs). That is, we have assumed that each study had an infinite sample size. We now introduce primary sampling error. For simplicity, assume that the average sample size for studies with women is the same as the average sample size for men. Suppose that in both cases, the average sample size is N = 68. Then in every study conducted in autocratic organizations, $\mu_\rho = -.12$ and $SD_e = 1 - (-.12)^2/\sqrt{67} = .1204$. Likewise, in every study conducted in democratic organizations, $\mu_r ho = .20$ and $SD_e = .1173$. Across all studies, $\bar{r} = (-.12 + .20)/2 = +.04$ and:

$$V_r = V_\rho + \text{Ave(Ve)}$$
$$= .16^2 + (.1204^2 + .1173^2)/2$$
$$= .0256 + .0141277 = .0397277$$
$$SE_r = .1993$$

Consider now the meta-analysis moderator analysis for sex of worker. If there were a large number of studies in both sub-domains, then the preliminary findings would be the same for both subsets of studies: for men, the mean sample correlation would be +.04 with a standard deviation of .1993 and for women, the mean sample correlation would be +.04 with a standard deviation of .1993. This is the null hypothesis for the moderator significance test.

In fact, the number of studies done in this domain will probably be modest. The mean sample correlation for men will differ randomly from +.04. The mean sample correlation for women will differ randomly from +.04. Since these means are from independent samples, the two deviations will be independent of each other. As a result, the two deviations will be different (if computed without rounding) and thus the two observed means will be different. This difference will be due to sampling error and will fall within the range dictated by statistical theory.

We now have a set of observed correlations (studies) with mean +.04 and standard deviation .1993. If we randomly select K such studies and compute the mean of those K studies, then the sampling variance of that mean will be:

$$V_m = V_r/K = .0397/K$$

Thus we have the statistical basis for predicting the outcome of the meta-analysis:

K studies with women workers:

$$\bar{r}_w = +.04 + \text{sampling error} = .04 + se_w$$
$$V_{se_w} = .0397/K$$

K studies with men workers:

$$\bar{r}_w = +.04 + \text{sampling error} = .04 + se_m$$
$$V_{se_m} = .0397/K$$

Difference between means:

$$\bar{r}_m - \bar{r}_w = .00 + se_m - se_w$$
$$V_{se} = \text{Var}(se_m) + \text{Var}(se_w)$$
$$= (.0397/K) + (.0397/K)$$
$$= 2(.0397/K) = .0794/K$$

The RE significance test is then:

$$z = (\bar{r}_m - \bar{r}_w)/V_{se}^{\frac{1}{2}} = (\bar{r}_m - \bar{r}_w)/SD_{se}$$
$$z = (\bar{r}_m - \bar{r}_w)/\sqrt{(.0794/K)}$$

For the FE model formula:

$$V_e = \text{variance due to primary sampling error}$$
$$= (1 - \rho^2)^2/(N - 1) = (1 - .04^2)^2/67$$
$$= .014878$$
$$SD_e = .1220$$
$$= V_{e_m}/K + V_{e_w}/K$$
$$= 2V_e = 2(.01478/K)$$
$$= .02956/K$$

The FE significance test is then:

$$z = (\bar{r}_m - \bar{r}_w)/\sqrt{(.02956/K)}$$

The two significance tests differ in their estimate of the standard error in the mean correlation for each sub-set of studies. The FE formula erroneously assumes that the standard deviation of sample correlations in each group is .1220 when it is in fact .1993. That is, the FE formula underestimates the standard error by a factor of .1220/.1993 = .61, i.e., by 39%. To state this another way, the FE formula assumes a standard deviation of .12 when the actual standard deviation is 63% larger than .12. The probability of the critical region is thus higher than the 5% assumed by the advocates of the FE significance test.

Suppose there are 25 studies in each sub-set (i.e., a total of 50 studies for the whole meta-analysis). The FE formula assumes that the critical region for its statistic 'z' is the usual ±.1.96 because it assumes that the standard deviation of 'z' is 1. But in reality, the standard deviation of this 'z' is actually 1.63. The

probability of the upper critical region for the RE formula is $Pz > 1.96) = .025$. For the FE formula this probability is:

$$P\{\text{'z'} > 1.96) = P\{z > 1.96/1.63 = 1.20\} = .115$$

Thus the overall Type I error rate is 5% for the RE formula but 23% for the FE formula. Users of FE model assume that the error rate for their significance test is 5% when it is in fact 23%, nearly 5 times larger.

Notes

1 Much of the focus of this article is on the use of significance tests in meta-analysis – a practice that we do not advocate (Hunter 1997; Schmidt 1996; Schmidt and Hunter 1997). However, this practice is widespread and therefore it is important to examine its effects, especially with respect to Type I error rates in the FE meta-analysis models.

2 The following meta-analyses, all of which appeared in *Psychological Bulletin*, the premier general review journal in psychology, are some recent examples: Bettencourt and Miller (1996), Bond and Titus (1983), Burt, Zembar and Niederehe (1995), Collins and Miller (1994), Eagly and Carli (1981), Eagly and Johnson (1990), Eagly, Karau and Makhijani (1995), Eagly, Makhihani and Klonsky (1992), Erel and Burman (1996), Feingold (1994), Herbert and Cohen (1995), Ito, Tiffany, Miller and Pollock (1996), Jorgensen, Johnson, Kolodzie0, and Scheer (1996), Knight, Fabes and Higgins (1996), Newcomb and Bagwell (1995), Polich, Pollock, and Bloom (1994), Symons and Johnson (1997), Van Ijzendorn (1995), Voyer, Voyer and Bryden (1995), Wood (1987), and Wood, Lundgren, Ouellette, Busceme and Blackstone (1994).

3 In some of these meta-analyses, the authors specify that confidence intervals for mean effect sizes are first computed and presented and only after this are tests of homogeneity conducted (e.g., Collins and Miller 1994: 462; Feingold 1994: 432)!

4 Erez, Bloom and Wells (1996) called for increased use of RE models in preference to FE models. However, they did not discuss the most widely used FE methods, those of Hedges and Olkin (1985). Nor did they discuss the FE methods of Rosenthal and Rubin (Rosenthal 1991; Rosenthal and Rubin 1982). They also misidentified the methods in Hunter, Schmidt and Jackson (1982), Hunter and Schmidt (1990a), Callendar and Osburn (1980), and Raju and Burke (1983) as FE methods; these methods are all RE methods.

5 Examples include Bettencourt and Miller (1996), Bond and Titus (1983), Burt, Zember and Niederehe (1995), Collins and Miller (1994), Eagly and Johnson (1990), Eagly, Karau and Makhijani (1995), Eagly, Makhihani and Klonsky (1992), Erel and Burman (1996), Feingold (1994), Ito *et al.* (1996), Knight, Fabes and Higgins (1996), Newcomb and Bagwell (1995), Polich, Pollock and Bloom (1995), Van Ijzendorn (1995), Wood (1987), Wood *et al.* (1994).

References

Barrick, M.R. and Mount, M.K. (1991) The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, **44**, 1–26.

Becker, B.J. (1987) Applying tests of combined significance in meta-analysis. *Psychological Bulletin*, **102**, 164–71.

Becker, B.J. (1988) Synthesizing standardized mean change measures. *British Journal of Mathematical and Statistical Psychology*, **41**, 257–78.

Becker, B.J. (1996) The generalizability of empirical research results. In C. P. Benbow and D. Lubinski (eds.), *Intellectual Talent: Psychological and Social Issues*. Baltimore: Johns Hopkins University Press, 363–83.

Bettencourt, B.A. and Miller, N. (1996) Gender differences in aggression as a function of provocation: A meta-analysis. *Psychological Bulletin*, **119**, 422–7.

Bond, C.F. and Titus, L.J. (1983) Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin*, **94**, 265–92.

Brown, S.P. (1996) A meta-analysis and review of organizational research. *Psychological Bulletin*, **120**, 235–55.

Burt, D. B., Zembar, M.J. and Niederehe, G. (1995) Depression and memory impairment: A meta-analysis of the association, it pattern, and specificity. *Psychological Bulletin*, **117**, 285–303.

Callender, J.C. and Osburn, H.G. (1980) Development and test of a new model for validity generalization. *Journal of Applied Psychology*, **65**, 543–58.

Cohen, J. (1962) The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, **65**, 145–53.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*. (2nd edn.) Hillsdale, NJ: Erlbaum.

Cohen, J. (1992) Statistical power analysis. *Current Directions in Psychological Science*, **1**, 98–101.

Cohen, J. (1994) The earth is flat ($p < .05$). *American Psychologist*, **49**, 997–1003.

Collins, N.L. and Miller, L.C. (1994) Self-disclosure and liking: A meta-analytic review. *Psychological Bulletin*, **116**, 457–75.

Cooper, H. (1997) Some finer points in meta-analysis. In M. Hunt (ed.), *How Science Takes Stock: The Story of Meta-analysis*. New York: Russell Sage Foundation, pp. 169–81.

Cooper, H. and Hedges, L.V. (eds.) (1994) *The*

Handbook of Research Synthesis. New York: Russell Sage Foundation.
Cowles, M. (1989) *Statistics in Psychology: An Historical Perspective*. Hillsdale, NJ: Erlbaum.
Eagly, A.H. and Carli, L.L. (1981) Sex of researchers and sex-typical communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies. *Psychological Bulletin*, **90**, 1–20.
Eagly, A.H. and Johnson, B.T. (1990) Gender and leadership style: A meta-analysis. *Psychological Bulletin*, **108**, 233–56.
Eagly, A.H., Karau, S.J. and Makhijani, M.G. (1995) Gender and the effectiveness of leaders: A meta-analysis. *Psychological Bulletin*, **117**, 125–45.
Eagly, A.H., Makhijani, M.G. and Klonsky, B.G. (1992) Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin*, **111**, 3–22.
Erel, O. and Burman, B. (1996) Interrelatedness of marital relations and parent-child relations: A meta-analytic review. *Psychological Bulletin*, **118**, 108–32.
Erez, A., Bloom, M.C. and Wells, M.T. (1996) Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology*, **49**, 275–306.
Feingold, A. (1994) Gender differences in personality: A meta-analysis. *Psychological Bulletin*, **116**, 429–56.
Hedges, L.V. (1983) A random effects model for effect sizes. *Psychological Bulletin*, **93**, 388–95.
Hedges, L.V. (1988) The meta-analysis of test validity studies: Some new approaches. In H. Wainer and H. Braun (eds.), *Test Validity*. Hillsdale, NJ: Erlbaum, pp. 191–212.
Hedges, L.V. (1992) Meta-analysis. *Journal of Educational Statistics*, **17**, 279–96.
Hedges, L.V. (1994a) Statistical considerations. In H. Cooper and L.V. Hedges (eds.), *Handbook of Research Synthesis*. New York: Russell Sage Foundation, pp. 29–38.
Hedges, L.V. (1994b) Fixed effects models. In H. Cooper and L.V. Hedges (eds.), *Handbook of Research Synthesis*. New York: Russell Sage Foundation, pp. 285–300.
Hedges, L.V. and Olkin, I. (1985) *Statistical Methods for Meta-analysis*. Orlando, FL: Academic Press.
Hedges, L.V. and Vevea, J.L. (1998) Fixed- and random effects models in meta-analysis. *Psychological Methods*, **3**, 486–504.
Herbert, T.B. and Cohen, S. (1995) Depression and immunity: A meta-analytic review. *Psychological Bulletin*, **113**, 472–86.
Hunter, J.E. (1997) Needed: A ban on the significance test. *Psychological Science*, **8**, 3–7.
Hunter, J.E. and Schmidt, F.L. (1990a) *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Beverly Hills, CA: Sage.
Hunter, J.E. and Schmidt, F.L. (1990b) Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology*, **75**, 334–49.
Hunter, J.E. and Schmidt, F.L. (1996) Cumulative research knowledge and social policy formulation: the critical role of meta-analysis. *Psychology, Public Policy and Law*, **2**, 324–347.
Hunter, J.E., Schmidt, F.L. and Jackson, G.B. (1982) *Meta-analysis: Cumulating Research Finding across Studies*. Beverly Hills, CA: Sage.
Ito, Tiffany A., Miller, N. and Pollock, V.E. (1996) Alcohol and aggression: A meta-analysis of the moderating effects of inhibitory cues, triggering events, and self-focused attention. *Psychological Bulletin*, **120**, 60–82.
Jorgensen, R.S., Johnson, B.T., Kolodziej, M.E. and Scheer, G.E. (1996) Elevated blood pressure and personality: A meta-analytic review. *Psychological Bulletin*, **120**, 293–320.
Knight, G.P., Fabes, R.A., and Higgins, D.A. (1996) Concerns about drawing causal inferences from meta-analyses: An example in the study of gender differences in aggression. *Psychological Bulletin*, **119**, 410–21.
Loftus, G.R. (1996) Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, **5**, 161–71.
Mengersen, K.L., Tweedie, R.L. and Biggerstaff, B. (1995) The impact of method choice on meta-analysis. *Australian Journal of Statistics*, **37**, 19–44.
Morris, S.B. and DeShon, R.P. (1997) Correcting effect sizes computed from factorial analysis of variance for use in meta-analysis. *Psychological Methods*, **2**, 192–9.
Mosteller, F.M. and Bush, R.R. (1954) Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of Social Psychology: Volume I. Theory and Method*. Cambridge, MA: Addison-Wesley.
National Research Council (1992) *Combining Information: Statistical Issues and Opportunities for Research*. Washington, DC: National Academy Press.
Newcomb, A.F. and Bagwell, C.L. (1995) Children's friendship relations: A meta-analytic review. *Psychological Bulletin*, **117**, 306–47.
Osburn, H.G. and Callender, J. (1992) A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology*, **77**, 115–22.
Overton, R.C. (1998) A comparison of fixed effects and mixed (random effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods*, **3**, 354–79.
Polich, J., Pollock, V.E., and Bloom, F.E. (1994) Meta-analysis of P300 amplitude from males at risk for alcoholism. *Psychological Bulletin*, **115**, 55–73.
Raju, N.S. and Burke, M.J. (1983) Two procedures for studying validity generalization. *Journal of Applied Psychology*, **68**, 382–95.
Raudenbush, S.W. (1994) Random effects models. In H. Cooper and L.V. Hedges (eds.), *Handbook of Research Synthesis*. New York: Russell Sage Foundation, pp. 301–22.
Raudenbush, S.W. and Bryk, A.S. (1985) Empirical Bayes meta-analysis. *Journal of Educational Statistics*, **10**, 75–98.
Rosenthal, R. (1978) Combining results of independent studies. *Psychological Bulletin*, **85**, 185–93.
Rosenthal, R. (1991) *Meta-analytic Procedures for Social Research*. Sage.
Rosenthal, R. (1995) Writing a meta-analytic review.

Psychological Bulletin, **118**, 183–92.
Rosenthal, R. and Rubin, D.B. (1982a) Further meta-analytic procedures for assessing cognitive gender differences. *Journal of Educational Psychology*, **74**, 708–12.
Rosenthal, R. and Rubin, D.B. (1982b) Comparing effect sizes of independent studies. *Psychological Bulletin*, **92**, 500–4.
Rubin, D.B. (1980) Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association*, **75** (372), 801–27.
Rubin, D.B. (1981) Estimation in parallel randomized experiments. *Journal of Educational Statistics*, **6**, 337–400.
Schmidt, F.L (1992) What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, **47**, 1173–81.
Schmidt, F.L. (1996) Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, **1**, 115–29.
Schmidt, F.L. and Hunter, J.E. (1997) Eight common but false objections to the discontinuation of statistical significance testing. In L. Harlow and S. Muliak (eds.), *What If There Were No Significance Tests?* Hillsdale, NJ: Lawrence Erlbaum.
Schmidt, F.L., Hunter, J.E. and Raju, N. S. (1988) Validity generalization and situational specificity: A second look at the 75% rule and the Fisher z transformation. *Journal of Applied Psychology*, **73**, 665–72.
Schmidt, F.L., Law, K., Hunter, J.E., Rothstein, H.R., Pearlman, K. and McDaniel, M. (1993) Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology*, **78**, 3–13.
Sedlmeier, P. and Gigerenzer, G. (1989) Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, **105**, 309–16.
Shadish, W.R., and Haddock, C.K. (1994) Combining estimates of effects size. In H. Cooper and L.V. Hedges (eds.), *Handbook of Research Synthesis*. New York: Russell Sage Foundation, pp. 261–82.
Stoffelmeyr, B.E., Dillavou, D. and Hunter, J.E. (1983) Premorbid functioning and recidivism in schizophrena: A cumulative analysis. *Journal of Consulting and Clinical Psychology*, **51**, 338–52.
Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. and Williams, R.M., Jr. (1949) *The American Soldier: Adjustment During Army Life* (Vol. 1). Princeton, NJ: Princeton University Press.
Symons, C.S. and Johnson, B.T. (1997) The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, **121**, 371–94.
Van Iizendorn, M.H. (1995) Adult attachment representations, parental responsiveness, and infant attachment: A meta-analysis on the predictive validity of the adult attachment interview. *Psychological Bulletin*, **117**, 387–403.
Voyer, D., Voyer, S. and Bryden, M.P. (1995) Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, **117**, 250–70.
Wood, W. (1987) Meta-analytic review of sex differences in group performance. *Psychological Bulletin*, **102**, 53–71.
Wood, W., Lundgren, S., Ouellette, J.A., Busceme, S., and Blackstone, T. (1994) Minority influence: A meta-analytic review of social influence processes. *Psychological Bulletin*, **115**, 323–45.