UNDERSTANDING STATISTICS, 2(1), 1–12

Meta-Analysis and Power: Some Suggestions for the Use of Power in Research Synthesis

Steven J. Muncer, Mark Craigie, and Joni Holmes
Department of Applied Psychology
University of Durham, Stockton Campus

The importance of statistical power is under recognized both in single study research and meta-analysis. The power of a study is the probability that it will lead to a statistically significant result. A simple method of establishing the adequacy of the power of meta-analysis is suggested and examples from 2 meta-analytic studies that may have produced inaccurate results are provided. Suggestions are made for changes in the protocol of meta-analytic studies that highlight the importance of power analysis.

keywords: meta-analysis, weighted mean effect, power

In the last 20 years meta-analysis has become increasingly important as a research tool both in the health and social sciences. Many medical interventions have been implemented based on the results of meta-analysis of randomized control trials. Meta-analysis has also become more sophisticated over the years with a move away from simple vote counting methods to a reliance on unbiased mean effect size, which gives the procedure a more sophisticated and scientific appearance (Bangert-Drowns, 1986).

There have, however, always been critics from both the medical (Feinstein, 1995) and social sciences (Eysenck, 1995). The criticism from the medical sciences was intensified by the discrediting of the results from a meta-analysis of magnesium as a successful intervention for myocardial infarction (Egger & Davey-Smith, 1995; Teo, Yusuf, Collins, Held, & Peto. 1991). More recently LeLorier, Gregoire,

Requests for reprints should be sent to Steven J. Muncer, Department of Applied Psychology, University of Durham University, Queen's Campus, University Boulevard, Thornaby, Stockton-on-Tees, TS17 6BH England. E-mail: s.j.muncer@durham.ac.uk

Benhaddad, La Pierre, and Derederian (1997) also found that there were large discrepancies between the results of meta-analyses and those of large scale randomized controlled trials. Specifically they found that meta-analysis would have resulted in the adoption of an ineffective treatment in 32% of cases and to the rejection of a useful treatment in 33% of the cases. The reasons for these possibly inaccurate findings are usually given as publication bias (Rosenthal, 1979), the tower of Babel effect (Gregoire, Derderian, & LeLorier, 1995), and inadequacies in experimental procedure. *Publication bias* refers to a preference for publishing significant results while penalizing other studies that, despite being methodologically sound, do not report statistically significant outcomes. Although we recognize these reasons, in this article we argue that an important weakness with both meta-analyses and the studies used in meta-analyses is the lack of attention given to statistical power.

RATIONALE FOR ADOPTION OF POWER ANALYSIS

At the present time most reviewers rely heavily on statistical significance when making decisions about acceptance or rejection of an article. Negative results, that is those that show no significant effect of a treatment, are extremely difficult to publish, so much so that they may be given little coverage in an article that reports other significant findings (Polit & Sherman, 1990). Naylor (1997) estimated that there is a "threefold difference" in the publication of randomized controlled trials that offer significant results as compared to those trials with nonsignificant results. In addition, it has also been suggested that those studies with negative results that are published have a much longer period of time "in press." The lack of, and delay in, the publication of negative results that are placed secondary to significant studies is a major issue for those interested in the establishment of a valid body of knowledge. This over reliance on significance at the expense of power is caused by a lack of knowledge about the meaning of both (Carver, 1978; Falk, 1986; Pollard & Richardson, 1987), and the effect of this over reliance on meta-analytic results is detrimental.

The power of an experiment is determined by the probability of making a Type 2 error, that is the probability of not rejecting the null hypothesis given that it is false, which is referred to as β (Oakes, 1986). The power of an experiment is thus $(1 - \beta)$. The statistical power of an experiment is affected by the size of the sample, the effect size (i.e., the impact in standard units of the independent variable on the dependent variable), and the significance level chosen. If these factors are known, then the power of the experiment can be derived from tables provided by Cohen (1988) or by computer packages, such as G*Power (Erdfelder, Faul, & Buchner, 1996). Cohen argued that the importance of Type 2 errors and power analysis has been ignored because of the overemphasis on the importance of the null hypothesis. A misunderstanding of the meaning of significant results may have been the cause of this problem.

For whatever reason, power analysis has been either seldom carried out or reported in any area of research. Studies that have analyzed the power of already published work have found that for small and medium effect sizes, the power has been lower than the accepted figure of .8. Polit and Sherman (1990) reported the findings of studies that examined power in 541 studies from reports published in psychological, education management, and medical journals; for small effect sizes the average power was .19 and for medium effect sizes .53. This type of finding has also been demonstrated in the social sciences (Brewer, 1972; Chase & Chase, 1976; Kazdin & Bass, 1989; Orme & Combes-Orme, 1986; Rossi, 1990).

It may be argued that low power in studies that report significant results is unimportant, as they have already rejected the null hypothesis. This is, however, to prejudge the results. If these studies did not produce significant results, it could be because the null hypothesis was true or it could be because the study was seriously underpowered. This has been used as an implicit reason for the rejection of studies; that is, a study will only be published if there is a significant result and the null hypothesis can be rejected. More important, it influences researchers as far as attempting to publish their research is concerned: significant results become paramount. For those involved in meta-analysis, it creates the so-called "file drawer" problem:

> The possibility that the journals are filled with the five per cent of studies that show Type I errors while the file drawers back at the lab are filled with 95 per cent of these studies that show non-significant results of $p > .05$. (Rosenthal, 1979, p. 638)

INTRODUCING POWER ANALYSIS INTO META-ANALYSIS

We believe that power analysis should be applied not just to individual studies but to meta-analysis and also to the selection of studies that can be entered into a meta-analysis. It is taken for granted that one of the benefits of meta-analysis is that it will increase the power of the research by increasing sample sizes. Although it is far from clear that power is additive, certainly two underpowered studies should not be viewed with the same confidence as one adequately powered study. The researchers who carried out the underpowered studies are putting all their eggs in one basket; the results would have to be significant for the study to be publishable, and hence a possible bias is, therefore, introduced.

Studies with low power have a serious and detrimental effect on the adequacy of any meta-analysis to which they contribute. A low powered study is far more likely to have been published if its findings were significant due to publication bias. This is, of course, also true for studies with higher power. However, a low powered study has to exhibit a much larger effect size to produce a significant find-

ing. It is likely that "freak" sampling of extreme values in the population could cause such large effect sizes; in other words, there is an inflation of the effect size due to accidental selection of nonrepresentative samples. In the particular case in which the study shows a false positive result (i.e., where a Type I error has occurred), this will have even more damaging consequences to the accuracy of a meta-analysis. In this case, low powered studies are equally likely to show positive and negative effect sizes, and in both directions the estimated effect size from the study will be inflated. Hence, low powered studies are likely to inflate effect size estimates and increase heterogeneity of effect sizes in the meta-analysis. Where low powered false positive studies are included, the increase in heterogeneity of effect sizes may mask the extent and direction of the true effect size. In contrast, studies that are adequately powered are less prone to such heterogeneity of effect sizes, as they are more likely to represent the true effect size in the populations.

One potential alternative to incorporating power analysis into meta-analysis is to use effect size estimates that explicitly take into account the variability in the estimators. One such example is the use of weighted means to measure and estimate effect sizes. We accept that such procedures do ameliorate the effects of low power to a degree; we are arguing that directly dealing with the problem by excluding low powered studies is a more effective way to remedy the problem. In fact, we would advocate the use of both weighted means and power analysis. We also accept that there are many other factors that affect the heterogeneity of effect sizes in meta-analysis; however, this fact should not preclude researchers from eliminating one of the main factors leading to heterogeneity, namely low power. It would be highly beneficial if researchers conducting meta-analyses applied a minimum power value to studies as part of their inclusion criteria.

The power of the meta-analysis itself should also be examined. It is, for example, possible to combine a number of studies and still produce a combined study that is insufficiently powerful. For example, Gøtszche, Hammarquist, and Burr (1998) published a meta-analysis of studies looking at house dust mite control measures and their effectiveness in the management of asthma. Their major result was that 41 out of 113 patients exposed to treatment interventions improved, compared with 38 out of 117 in the control groups. This suggested a nonsignificant effect of dust mite control. The study, however, is considerably underpowered if we assume a small effect size; the results from 785 participants would have been needed to have the recommended power of .8 with a significance level of .05. The study is also inadequate to detect a medium effect size (Muncer, 1999).

SUGGESTED PROCEDURE FOR ADOPTION OF POWER ANALYSIS IN META-ANALYSIS

The introduction of power analysis into the protocol of meta-analysis would counter the difficulties described above and make the results of meta-analysis more reli-

able and more likely to be consistent with large-scale randomized control trials. We, of course, accept that there are a variety of reasons why the results of meta-analysis and large-scale randomized control trials may differ. Our point is simply that the inclusion of low powered studies in a meta-analysis is likely to bias the overall effect size calculated to such a degree that the reliability of the statistic is in question and the meta-analysis is very likely to differ in its results from a large scale randomized control trial. We do not wish to state that the use of a minimum power criterion in meta-analysis would eliminate differences between meta-analysis and large-scale randomized control trials; however, it does remove a considerable source of error that contributes to such differences.

We suggest the following procedure:

1. Conduct a systematic review to ensure that all possible relevant articles have been collected.
2. Calculate the effect size for each study. At this stage a weighted mean effect size should be calculated and then the power of each study to test for the unbiased mean effect size should be calculated.
3. If the average power of the studies in the meta-analysis is low, an examination of the power of individual studies should be undertaken with a view to calculating an effect size for studies in which the power is acceptable. According to Cohen (1988), a single study should be considered adequately powered if it has a power of .8, but because meta-analysis involves combining studies, it may be argued that this criterion could be lowered. Muncer, Taylor, and Smith (1999) suggested using the .8 inclusion criterion for meta-analyses of health related studies and a lower criterion of .5 for studies in the social sciences in which errors are less likely to have life-threatening consequences. However, although the optimal power figure is yet to be firmly established, clearly the nearer it is to .8 and above, the better it would be.
4. Studies that meet a given power criterion should then be combined into a meta-analysis and the effect sizes recalculated.
5. Finally the power of the meta-analysis overall should be calculated and published with the meta-analysis.

Although this method has the potential to improve meta-analysis in the future by providing an a priori statistical basis for the inclusion of studies, it is also helpful in allowing us to evaluate existing meta-analyses. Note that the procedure in Steps 3 to 5 can be iterative. In cases in which the results of the first application of the power criterion to the individual studies in the meta-analysis still results in a meta-analysis whose overall power is too low, then the mean effect size of the included studies can be used to recalculate the power of all the studies. This is achieved by using the mean effect size calculated for the included studies as an estimate of the effect size in the populations being considered. The power criterion

can then be applied to the new power values for each study, and the procedure applied again so that a new set of effect sizes are calculated for those studies that pass the power criterion. This procedure iterates until there is no further change in the included studies, and therefore in the effect size and power calculated.

There will, however, be occasions when all studies in a meta-analysis are so lacking in statistical power that no further analysis should be undertaken. In these cases, it is advisable to calculate the mean weighted unbiased effect size for all studies and then calculate the mean power of the studies to support this effect size. These figures should be reported with the meta-analysis. In cases like this, it is important that future studies that are undertaken are sufficiently powered.

In the next section, we provide examples of these calculations using data from the meta-analysis from Kling, Hyde, Showers, and Buswell (1999), which examined sex differences in self-esteem, and Feingold's (1994) examination of sex differences in anxiety. Both articles were chosen because they are excellent examples of well-conducted meta-analyses and were published in a highly regarded journal. It is fairly easy to find poor examples of meta-analysis in the health and social sciences, but we have no intention of setting up straw men.

META-ANALYSIS EXAMPLES

Sex Differences in Self-Esteem

Kling et al. (1999) conducted a meta-analysis of 216 studies published between 1984 and 1997 that provided measures of self-esteem in males and females. The studies used participants from 6 age categories; 7 to 10 years old, 11 to 14 years old, 15 to 18 years old, 19 to 22 years old, 23 to 59 years old, and greater than 60 years old. Kling et al. provided a comprehensive set of tables to present their results and we used results presented in these tables to carry out our power analysis and all subsequent analysis.

The weighted mean effect size reported by Kling et al. (1999) was 0.21 with a confidence interval from 0.192 to 0.219, with an unweighted mean and median effect size of 0.2 and a modal effect size of 0.19. We calculated the weighted mean effect size from the information provided in the table using both a simple formula that weights each effect size directly by sample size (Cooper, 1998) and also using the weighting factor suggested by Hedges and Olkin (1985). Using the simple procedure the effect size is 0.207 and with the weighting factor to produce the weighted mean effect size 0.207 with a confidence interval ranging from 0.194 to 0.220. We recalculated the effect size so that we have a reasonable comparison with our future results and not as a check on Kling et al. We accept that their figures will be more accurate but have shown that our procedures using the results presented in the tables produce results that are accurate to within 0.002 of an effect size.

The next stage was to calculate the power of each of the studies to detect an effect size of 0.21, given the sample sizes that were used. This could either be done by hand, by the use of Cohen's tables to give approximate figures, or by a computer program called G*Power (Erdfelder et al., 1996). In this case we used G power to calculate the power of each study to detect the given effect with an independent samples *t* test. Given the findings of previous studies that examine power, it is not surprising that the mean power for these 216 studies was .503 with a standard deviation of .26 and a minimum of .126 and maximum of 1. The mean power of studies for the different age groups was, respectively, .414, .55, .548, .453, .424, and .599 for the over 60 age group. There was no significant difference in the power of studies from different age groups, $F(5, 207) = 2.21$, $p = .057$.

At this point we calculated the simple sample weighted and mean weighted effect size for studies with different levels of power. The first set of studies we examined all had power over the accepted level of .8. There were 42 studies that met this criterion. The simple effect size for these studies was 0.215 and the mean weighted effect size was 0.216 with a confidence interval ranging from 0.200 to 0.232. This is clearly very similar to the main result reported by Kling et al. (1999) and suggests that we can be fairly confident that the effect size is between about 0.2 and 0.23. We can, however, now be confident that our result is not being influenced by studies that should not be included.

The results are fairly similar for meta-analyses of the 57 studies with power over .7, which produces a weighted mean effect size of 0.218 with a confidence interval 0.203 and 0.233. Meta-analyses of the 69 studies with power over .6 produce a weighted mean effect size of 0.201 with a confidence interval of 0.196 and 0.225 and the 92 studies with power of over .5 produce an effect size of 0.209 with a confidence interval of 0.195 and 0.223. These results are all cumulative. Overall the results are either identical or slightly higher than that produced by a meta-analysis of all of the studies.

If we examine the studies that have very low power of less than .2, however, we find a different pattern. There are 21 studies that have a power of less than .2 and 6 of these have a negative effect size with females showing higher self-esteem. There are significantly more negative effect values in this group of studies, $\chi^2(1, N = 216) = 4.19$, $p = .041$, than in the remainder. The weighted mean effect size is also lower at 0.145 with a confidence interval of 0.016 to 0.274. Overall if we look at all studies that have a negative effect they have significantly less power with a mean of .403 as opposed to .519, $t(214) = 2.269$, $p = .024$. In this meta-analysis, which has a large number of studies, there is only a marginal effect on the estimate of effect size. In a smaller meta-analysis, however, studies with low power may be expected to exert a larger influence on the result as will be shown.

Kling et al. (1999) demonstrated that there are age differences with the late adolescent group showing a higher weighted mean effect of 0.33, it is not surprising, therefore, that the overall effect size is not homogeneous, $Q(215) = 629.77$, $p <$

.001. The Q statistic provides a test of heterogeneity that follows a χ^2 distribution The studies with power of .8 and above are also heterogeneous, $Q(41) = 360.48$, $p < .0001$, and indeed significantly more so, $Q(174) = 269.29$, $p < .001$. It is likely that this is also related to differences in effect size between the age groups. There are significant differences between the means of the effect sizes of the different age groups, $F(5, 33) = 5.17$, $p = .001$, with the late adolescent group having the largest mean and the over 60s group having the smallest. The weighted means for each age group for the studies with power over .8 and for all studies are presented in Table 1. Although the results for both sets of studies are similar, it is clear that the studies with power over .8 show more similarity for ages under 23 and greater difference between these under 23 studies and those with adults over 23. Both results suggest that at age 60 or over there is little difference in self-esteem ratings between the two sexes and that the largest difference is in the late adolescent period. The drop in the sex difference in self-esteem that accompanies adulthood is even clearer in the studies that have met the power criterion.

Although the reanalysis produces substantially similar results for self-esteem overall, it is important to note that the larger differences appear when the results are broken down into age categories. This is clearly because the impact of a study on the mean weighted effect size will be higher if there are fewer studies, as will also be demonstrated by the next example.

Sex Differences in Anxiety

Feingold (1994) carried out a number of meta-analyses to look at sex differences in self-esteem, locus of control, anxiety, and assertiveness. In this example we will focus on his replication of Hall's (1984) meta-analysis of anxiety, which examines 18 studies from 1986 to 1992. The mean effect size for the sex difference in anxiety was reported by Feingold as –0.15 (note in this part of the article Feingold reported unweighted mean effect sizes and we will use the same measure), which meant that females were slightly higher on anxiety. This would again be a small effect size as described by Cohen (1988).

TABLE 1
Weighted Mean Effect Sizes and Confidence Intervals for Each Age Group

All Studies	*Studies With Power of* ≥ .8
Age 7 to 10: 0.157 (0.101 to 0.213)	0.256 (0.161 to 0.351)
Age 11 to 14: 0.221 (0.197 to 0.245)	0.237 (0.209 to 0.265)
Age 15 to 18: 0.334 (0.307 to 0.361)	0.361 (0.328 to 0.394)
Age 19 to 22: 0.179 (0.148 to 0.21)	0.202 (0.14 to 0.264)
Age 23 to 59: 0.105 (0.064 to 0.146)	0.064 (0.014 to 0.114)
Age 60+: –0.025 (–0.074 to 0.024)	–0.027 (–0.078 to 0.024)

The mean power of the studies to detect an effect size of –0.15 was very low at .339 with a standard deviation of .21 and a range from .14 to .94. Only one study meets the .8 criterion and that had a larger effect size of –0.35. There is, therefore some reason to believe that a meta-analysis that takes power into consideration will produce different results. In this case, the number of studies and their relatively low power makes it sensible to initially adopt a power criterion of over .5 (Muncer, Taylor, & Smith, 1999).

There are three studies that meet this power criterion and their mean effect size is –0.36, which is substantially larger than the mean effect size for the total group of studies. Clearly, however, if this effect size is correct then we may be able to include more studies that would now meet the power criterion. In this case the mean power of the studies to detect an effect size of –0.36 is higher at .76. This makes it possible to use the preferred power criterion of .8 in future analyses. There are nine studies that now meet this power criterion (assuming a population effect size of –0.36) and these studies have an unweighted mean effect size –0.25.

Clearly if this effect size is correct then some of the nine studies previously included will no longer meet the power criterion of .8, as the new estimate of the population effect size (–0.25) is smaller than the previous estimate of –0.36. We, therefore, have to compute the power of studies to detect this effect size, –0.25. The mean power of studies to detect this effect size is .57 with a standard deviation of .24 and a range from .24 to 1. The unweighted mean effect size of these studies is –0.28. At this point, the iterative process ends as this new mean effect size is supported by all of the studies with a power of .8 and no new studies will be added. The weighted effect size for the studies that meet our .8 power criterion is 0.297 with a confidence interval ranging from 0.234 to 0.36. Interestingly this weighted mean effect size is similar to that found in Feingold's (1994) meta-analysis of studies cited in an earlier review by Maccoby and Jacklin (1974), in which the weighted mean effect size was –0.29. If we carry out a power-based meta-analysis on the studies cited in the Maccoby and Jacklin review we find that the weighted mean effect size is –0.3.

In our power-based analysis of the studies in the Hall (1984) meta-analysis, both the weighted and unweighted mean effect sizes are substantially larger than the 0.15 reported by Feingold (1994). This is not surprising as the Feingold analysis includes a number of low power studies with either low effects or effects in the opposite direction. There is indeed a trend for negative correlation between the power of a study and its effect size, $r(16) = -.449, p = .062$, for the studies in the Hall meta-analysis.

DISCUSSION

In both the examples we have given, the use of a power of .8 criterion has produced a larger effect size, although in the first case (Kling et al., 1999) the difference is

negligible. It should be noted, however, that this may not always be true; indeed, we might expect it to reduce the effect size more frequently. For example, if the procedure had been applied to the meta-analysis of magnesium trials (Teo et al., 1991) it would have been clear that the average power of the studies was low with a mean of .28; that none of the studies met a .8 power criterion and that the only study that met a power criterion of .5 and above, had a nonsignificant chi-square value indicating that magnesium did not reduce mortality. $\chi^2(1, N = 298) = 0.19$, $p = .66$, and an effect size of 0.04.

It should also be made clear that we are not suggesting that power analysis becomes the driving force in meta-analysis. There is still a need to look at the effect of moderator variables on the results and it may sometimes be true that these considerations outweigh the importance of power. There will also be occasions when none of the studies in an analysis meet a power criterion of .8 or even .5. Although we are not suggesting that such meta-analyses should not proceed, we believe that it is important that the small power of studies involved is explicitly stated. Such meta-analyses by their nature will be markedly affected by the results of a single study. For example, Bishop and Wahlsten (1997) in their analysis of the sex differences in the ratio of the splenium to the corpus callosum concluded that the effect size is extremely small at –0.11. It is worth noting that none of the studies in their sample meet the power criterion of .8 and that this result is enormously affected by a single study with an effect size of +1.496, involving 25 participants for whom the means and standard deviations are not reported.

When there are a relatively small number of studies in a meta-analysis the effects of single studies are going to have more effect, particularly if the other studies in the sample have low power. For example, the mean effect size in the Feingold (1994) meta-analysis was –0.15 and this would be lowered to –0.08 with the addition of three studies with fewer than 20 participants, showing effect sizes of 0.3. Although it would obviously take either more studies or studies with more power to affect the weighted mean effect size, it is clearly possible at present to make a difference with studies of low power. We are not arguing that the effect of low power in studies overrides the use of weighted means—this procedure obviously ameliorates the effect of low power somewhat. However, it does not eliminate the biasing effect of low power in the same way as applying a power criterion to the meta-analysis. We believe both procedures are required to ensure the reliability of meta-analyses. It has also been argued that at the moment because power is not a consideration in meta-analysis such "poor studies are encouraged or tolerated by a promise of eventual inclusion in meta-analyses" (Kraemer, Gardner, Brooks, & Yesavage, 1998, p. 24).

In this article we suggest some straightforward and simple procedures by which power analysis can be incorporated into meta-analysis. The importance to research of both power analysis and meta-analysis has been increasingly recognized (Rossi, 1998). Like others (Kraemer et al., 1998), we are suggesting that power analysis

should be explicitly combined with meta-analysis. In this article we suggest that all future meta-analyses should at least inform the reader of the average power of the studies that are being combined. Furthermore, we suggest that the goal of the researcher should be to produce a meta-analysis with acceptable power of .8 and have suggested ways in which that can be done. The goal should be to produce an effect size from studies that have sufficient power to support it. We are not suggesting that all other criteria should be abandoned, but we are suggesting that because meta-analysis is a statistical technique, statistical considerations like the power of a study are important.

It is important to note that we are not criticizing the meta-analyses that we have used in our examples. Both of them seem to be exemplary. However, we are suggesting and hope to have demonstrated that the methods used in them allow the effect size to be influenced by studies that would be better in the file drawer, or perhaps even the waste basket.

REFERENCES

Bangert-Drowns, R. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99,* 388–399.

Bishop, K. M., & Wahlsten, D. (1997). Sex differences in the human corpus callosum: Myth or reality. *Neuroscience and Biobehavioral Reviews, 21,* 581–601.

Brewer, J. K. (1972). On the power of statistical tests in the *American Educational Research Journal. American Educational Research Journal, 9,* 391–401.\

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399.

Chase, L., & Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology, 61,* 234–237.

Cooper, H. (1998). *Synthesizing research.* Thousand Oaks, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic.

Egger, M., & Davey-Smith, G. (1995). Misleading meta-analysis. Lessons from "an effective, safe, simple" intervention that wasn't. *British Medical Journal, 310,* 751–752.

Erdfelder, E., Faul, F., & Buchner, A. (1996). G*Power: A general power analysis program. *Behaviour Research Methods, Instruments & Computers, 28,* 1–11.

Eysenck, H. J. (1995). Problems with meta-analysis. In I. Chalmers & D. G. Altman (Eds.), *Systematic reviews* (pp. 64–74). London: BMJ.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9,* 83–96.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116,* 429–456.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology, 48,* 71–79.

Gøtzsche, P. C., Hammarquist, C., & Burr, M. (1998). House dust mite control measures in the management of asthma: Meta-analysis. *British Medical Journal, 317,* 1105–1110.

Gregoire, G., Derderian, F., & LeLorier, J. (1995). Selecting the language of the publications included in a meta-analysis: Is there a Tower of Babel bias? *Journal of Clinical Epidemiology, 48,* 159–163.

Hall, J. A. (1984). *Nonverbal sex differences: Communication accuracy and expressive style.* Baltimore: Johns Hopkins University Press.

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

Kazdin, A. E., & Bass, D. (1989). Power to detect differences between alternative treatments in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57*, 138–147.

Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin, 125*, 470–500.

Kraemer, H. C., Gardner, C., Brooks, J., & Yesavage, J. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3*(1), 23–31.

LeLorier, J., Gregoire, G., Benhaddad, A., LaPierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large scale randomized controlled trials. *The New England Journal of Medicine, 337*, 536–542.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

Muncer, S. J. (1999). Power dressing is important in meta-analysis. *British Medical Journal, 318*, 870–871.

Muncer, S. J., Taylor, S., & Smith, M. (1999). Power dressing and meta-analysis: New clothes for an emperor. *VI European Congress of Psychology*. Abstracts, 310–311.

Naylor, C. D. (1997). Meta-analysis and the meta-epidemiology of clinical research: Meta-analysis is an important contribution to research and practice but it's not a panacea. *British Medical Journal, 315*, 617–619.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Orme, J., & Combs-Orme, T. (1986, Fall). Statistical power and Type II errors in social work research. *Social Work Research*, 3–10.

Polit, D., & Sherman, R. (1990). Statistical power in nursing research. *Nursing Research, 39*, 365–369.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 102*, 159–163.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 683–641.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646–656.

Rossi, J. S. (1998). Meta-analysis, power analysis, and the null-hypothesis significance test procedure. *Behavioral and Brain Sciences, 21*, 216–217.

Teo, K., Yusuf, S., Collins, R., Held, P., & Peto, R. (1991). Effects of intravenous magnesium in suspected acute myocardial infarction: Overview of radomized trials. *British Medical Journal, 303*, 1499–1503.