Meta-analysis or best-evidence synthesis?

H.J. Eysenck PhD DSc
Institute of Psychiatry, University of London, London, UK

Correspondence
Prof. H.J. Eysenck
Institute of Psychiatry
University of London
Denmark Hill
London
UK

Keywords: curvilinearity, literature summary, meta-analysis, synergism

Accepted for publication:
8 June 1995

Abstract

This article examines the usefulness of meta-analysis, and articulates many of the criticisms that have been made of its workings. An attempt is made to outline the precautions that have to be taken before a scientifically useful and meaningful meta-analysis can be carried out. The problems encountered include heterogeneity of samples, conditions, interventions and end-points; narrow focus; curvilinearity of regression; lack of independence of determinants; synergistic interactions; contradictory experimental results. It is suggested that best-evidence synthesis, or theory-directed analysis, might be a safer option.

Types of analysis of scientific data

The purpose of scientific research is to support or disprove a theory, reach a conclusion on some important question, or settle a long-standing debate. In the hard sciences, this is usually achieved by carrying out some crucial experiment, as Newton said in his letter to Oldenburg (18 August 1676): 'For it is not number of Expts, but weight to be regarded; & where one will do what need many?' Of course, even in physics replication is necessary, and the notion of a 'crucial' experiment is no longer viable, but it remains an ideal. Contrasted with this ideal we have the medical and social sciences, where probabilistic solutions are the best that can often be achieved, and statistics is the ruler. To many scientists this is an abomination; as Rutherford used to say, 'if you need statistics to interpret your results you'd be advised to do a better experiment'!

However that may be, in the soft sciences experiments often give doubtful, ambiguous or even conflicting answers, and a way has to be found to come to some sort of conclusion. Mostly the experiments in question have several independent variables, rather than one, as is the ideal, and usually these variables are quantitatively differentiated. Thus in research on the relationship between class size and achievement in school (Glass & Smith 1979), the class sizes being compared will differ from study to study, as will the age and sex compositions of the children, the qualifications of the teachers, and the indices of achievement (by subject matter, by test, etc.). No single experiment can encompass all these variations, and hence there must be some method of integrating all these diverse studies. Always, when there are many studies devoted to a single topic, there are numerous attempts to find a reliable generalization, a worthwhile summary and a dependable law, and it becomes important to know how best to achieve this. There are several methods.

1 The traditional way of summarizing the literature on a given problem is for an experienced research worker in the field to consider the studies that have been carried out, give an opinion on their respective values, dismiss the worthless, and try to draw what conclusions it may be possible to extract from all this material. An investigator may or may not make a complete evaluation of all the studies published; he may exclude the worst studies, or include them only to criticize them for whatever faults they may exhibit.
2 Another way of summarizing data has been called by Slavin (1986) 'best evidence synthesis'. According to this method we should consider the 'best evidence'

in any field, from studies having the greatest internal and external validity, using well-specified, defined, explicit *a priori* inclusion and exclusion criteria, and favouring size-effect data over statistical significance alone. Such syntheses emphasize *numerical* findings, but their conclusions need not depend on a single estimation, nor on statistical significance.

3 A third, increasingly popular method is that of meta-analysis. As Huque (1988) states, '... the term "meta-analysis" refers to a statistical analysis which combines or integrates the results of several independent clinical trials considered by the analyst to be "combinable".' A distinctive characteristic of the strategy is a derivation of a single quantitative estimate of the effect of an interaction or a risk factor. The aims of a properly conducted meta-analysis are 'to increase statistical power; to deal with controversy when individual studies disagree; to improve estimates of size of effect, and to answer new questions not previously posed in component studies' (Hunter & Schmidt 1990).

The procedure advocated by meta-analysts is essentially one of (weighted) averaging (usually of size effects), and it consequently has all the advantages of averaging procedures. The main advantage, of course, is an increase in *statistical power* (Cohen 1988), as a result of the increase in the number of subjects over the number in any one constituent study. Studies in these areas often, or indeed usually, have too few subjects to allow an acceptable level of significance to be reached, given the small effect size (ES) expected, and averaging over all studies reduces Fisher Type 2 errors, which would normally be only too obvious. Small effects are difficult to detect in normal-sized studies; they become obvious in (combined) large-scale ones.

At the same time, meta-analysis largely absolves us from the illogical and mathematically inappropriate type of significance testing procedure normally used and advocated in statistical text-books. It is widely agreed that the typical procedures used for null-hypothesis significance testing (NHST) are illogical and error-fraught (Berkson 1938; Hogben 1957; Lykken 1968; Morrison & Henkel 1970; Sedlmeier & Gigerenzer 1989; Cohen 1994). Concern with ES rather than statistical significance is certainly an important advance, but can be found outside meta-analysis as well as inside; it is based on an argument that pre-dated the advent of meta-analysis, as the references cited above will testify.

The problem of heterogeneity

I will not deal with many of the problems that affect the details of meta-analytic procedures (Bahnson & Bliesener 1994; Cooper & Hedges 1994), but concentrate on weaknesses integral to the essential process of *averaging* that constitutes the central feature of meta-analysis. In doing so I shall refer to specific studies that are typical of the application of meta-analysis, and that have been widely praised as advancing the specific areas in which they have been carried out. The first and perhaps the most damaging criticism of any averaging procedure is that what is being averaged is *heterogeneous*, i.e. not suitable for averaging. In clinical studies we might say that treatment data can be averaged when: (1) samples are similar in sex and age composition, derivation, symptomatology, state of disease, etc.; (2) interventions are identical or sufficiently similar to make comparisons meaningful; (3) criteria for recovery, improvement, etc. are fundamentally similar. Given these degrees of homogeneity, meta-analysis is without doubt a useful procedure *unless results show a wide range of different values*, suggesting that conditions that have not been taken into account may be producing considerable heterogeneity. But, of course, if results are pretty similar meta-analysis is hardly needed; a simple survey of homogeneous findings is sufficient.

Let us now consider a typical meta-analytical study that has been widely praised and considered a 'classic' (Bahnson & Bliesener 1994). The study in question is a book by Smith *et al.* (1980), entitled *The Benefits of Psychotherapy* and concerned, as the title suggests, with the benefits of psychotherapy. Summarizing over 500 papers, these authors came to the conclusion (p. 183) that:

> psychotherapy is beneficial, consistently so and in many different ways. Its benefits are on a par with other expensive and ambitious interventions, such as schooling and medicine ... the evidence overwhelmingly supports the efficacy of psychotherapy.... Psychotherapy benefits people of all ages as

> reliably as schooling educates them, medicine cures them, or business turns a profit.

Many reviews have repeated these statements with approbation, relying on the *objectivity* of meta-analysis. I have expressed a contrary view (Eysenck 1978) related to an earlier publication by the same authors, categorizing it as 'an exercise in mega-silliness'. Why such an unparliamentary expression?

It is agreed that studies must be 'combinable' in order to be included. That means, in essence, that treatments, patients and end-points must be similar, or at least comparable. In the studies analysed by Smith *et al.* (1980), neither treatments, nor patients, nor end-points were remotely comparable. Patients could be severe neurotics, mild neurotics, students suffering from a specific phobic anxiety, or people suffering from some form of existentialist discomfort. Treatments were exceedingly varied; indeed, a table gives 18 different types of treatment! Endpoints were equally diverse, consisting of objective symptoms determined by examination, psychiatric opinion, questionnaire answers or some projective test such as the Rorschach Ink Blot Test. Let us now look at the table giving the effect size scores for the 18 different treatments. What we note right away is the title of treatment No. 18, which is 'placebo treatment'. The effects of a placebo treatment are very similar to those of psychodynamic therapy, Adlerian therapy, client-centred therapy, Gestalt therapy, rational-emotive therapy, transactional analysis or implosion treatment, and are significantly superior to those of undifferentiated counselling or reality therapy. In other words, what the study demonstrates, if anything, is that psychotherapy has no effect whatsoever beyond placebo treatment! This certainly is not the conclusion proffered by the authors, but it is the only one that makes any sense.

Smith *et al.* (1980) also conclude that 'different types of psychotherapy ... do not produce different types or degrees of benefit' (p. 184). A look at their table 5-1 shows that, of the 18 treatments surveyed, effect sizes vary from 0.14 to 2.38; this does not suggest that these different treatments do not produce different degrees of benefit. Cognitive-behavioural therapies are significantly superior to psychodynamic types of therapy, but this potentially important finding is suppressed. This 'closure' of meta-analysis thus combines very heterogeneous studies sharing neither treatments, nor patients, nor end-points, to arrive at meaningless effect sizes that are then misrepresented grossly in the summary! It might be argued that the differences in treatments, patients and end-points could be subjected to partial analyses, and, indeed, Smith *et al.* have attempted to do this. Their effort is not successful, however, for two reasons. As Beutler (1991) has shown, millions of studies would be needed to represent all possible combinations, and of course any such effort would only be acceptable if the different variables are uncorrelated, which clearly they are not.

We may compare this example of meta-analysis with an example of best-evidence synthesis, covering the same ground (Grawe *et al.* 1994). Here a much larger sample of studies is considered, the best are selected for detailed analysis, and each is assessed for good and bad points. The final book is a model for work of this kind. The major conclusion is the marked superiority of behaviour therapy over all other therapies, and there are many finer and more detailed conclusions concerning various types of therapy. No vacuous attempts are made at detailed and meaningless effect size computations, and no grandiose claims are made. Anyone interested in a comparison of the two methods should read and compare these two 'claims'; the obvious superiority of the Grawe *et al.* (1994) study will be immediately obvious.

Heterogeneity squared

One might have thought that absurdity could go no further, but that would seriously underestimate the capacity of academics to take to extremes a new method of analysis. Lipsey & Wilson (1993) have carried out a meta-analysis of meta-analyses, combining and extracting effect size estimates for a great variety of psychological, educational and behavioural treatments of a bewildering variety of disorders exemplifying an incredible degree of heterogeneity. As I have pointed out (Eysenck 1995):

> A method that averages apples, lice, and killer whales (here psychological, educational, and behavioural treatments) can hardly command scientific respect; there is little in common among psychotherapy for bulimia, cognitive behavioural

therapy with dysfunctional children, parent effectiveness training, diversion programmes for juvenile delinquents, effects of hypnosis or anxiety, group assertion training, career education programmes, social skills training, pre-operative preparation of children for surgery, biofeed-back for migraine, music therapy for pain reduction, adolescent pregnancy programmes, behavioural treatments for obesity, the Feingold diet for hyperactivity, computer-aided instruction, interactive video instruction, co-operative learning, positive reinforcement in the classroom, enrichment programmes, coaching for Scholastic Aptitude Tests, creativity training techniques, Frosting visual perception training, language intervention, science in-service training, career development courses, and mass media campaigns. To combine the outcomes of all these (and many more) meta-analyses seems to me a gigantic absurdity. To pretend that there is anything whatever in common among them seems difficult to justify and to have no ascertainable meaning. The mean effect size (ES) of 0.50 signifies what? It is an average of completely disparate methods, applied to completely disparate problems, with completely disparate controls. Would any physicist publish a meta-analysis of meta-analyses combining Bode's law, Boyle's law, $E = mc^2$, Kepler's law, the cosmological redshift interpretation, Heisenberg's indeterminacy principle, the laws of chromodynamics, quantum theory of fields, and hundreds more to prove that the sum total works out marginally better than nothing? I don't think so.

Narrowness of focus

If lack of homogeneity of method, sample and end-point is the major criticism of many applications of meta-analysis, excessive narrowness of focus is almost as bad a fault. Effect size is important, but looking at nothing but effect size can completely change the conclusion from negative to positive. It may be useful to consider an example of what I have in mind. The National Research Council (1986) published a meta-analysis of environmental tobacco smoke, trying to measure exposure and assessing health effects. In this study of the association between a non-smoker's exposure at home to environmental tobacco smoke (ETS) and the risk of lung cancer, the overall effect found by the NRC was a statistically significant 1.34 ($P < 0.001$), with a 95% confidence interval extending from 1.18 to 1.53. Fleiss & Gross (1991) have severely criticized the NRC study; as they conclude: 'The meta-analysis performed by the NRC must either be completely discounted or, as Stein (1988) conclude so succinctly in another context, considered a "mere" computational exercise.'

Fleiss & Gross 91991) have undertaken another meta-analysis of the American data, concluding that the overall effects are statistically non-significant, adding (p. 137):

> the fact that no significant association was found neither vindicates nor condemns the meta-analysis of these epidemiological studies. Given the biases that exist in each individual study, the safest conclusion from the present meta-analysis is a negative one: there is no convincing scientific evidence from the epidemiological literature of an association between exposure to ETS and the risk of lung cancer in the U.S.

Spitzer *et al.* (1990) have carried out a *best evidence* analysis, concluding that 'the weight of evidence is compatible with a positive association between residential exposure to environmental tobacco smoke (primarily from spousal smoking) and the risk of lung cancer.' There are some slight differences in the wording of the above conclusions, one being negative, the other not ruling out a positive effect, but on the whole they tend to agree with each other, as well as with a conclusion reached using a traditional method of viewing the data (Eysenck 1991).

The meta-analyses quoted not only contradict each other; by concentrating on ES estimation alone, they manage to forget and obscure the existence of a great deal of literature that is both relevant to the major issue (the effect on health of passive smoking) and destructive of the very research paradigm used in all the studies summarized. Consider that practically all of these studies are based on the verbal reports of respondents regarding their smoking habits; if these are incorrect and biased, results are meaningless.

Lee (1988, 1992), who has carried out a survey of the published literature on smoking habit misclassification, has pointed out that even a small proportion of smokers claiming to be non-smokers can cause a marked upward bias in estimates of the relative risk associated with marriage to a smoker. Because (as has been confirmed) smokers tend preferentially to marry smokers, subjects reporting being non-smokers married to smokers are more likely actually to be smokers than non-smokers married to non-smokers. Lee (who also noted that the reverse misclassification, of non-smokers as smokers, has only a minor biasing effect) concluded that bias resulting from misclassification of smokers as non-smokers could explain most, if not all, of the alleged effect of passive smoking on lung cancer (an allegation which is based in large part on evidence of a risk increase in relation to marriage to a smoker). Available data on over 100 studies (Lee 1988) confirm the prevalence of directionally biased reporting, and a recent study by Eysenck (1991) adds to this evidence.

If we add doubts about the reliability and validity of diagnoses (death certificates), detection bias and other sources of error (Eysenck 1991), we can see that omission of all this evidence which is relevant to the paradigm of research used makes meta-analysis meaningless, and that exclusive reliance on ES estimates regardless of how they are obtained makes any conclusions unacceptable. Consideration of the widely quoted National Research Council meta-analysis from this point of view clearly disqualifies it as a scientific summary of evidence on the effects of passive smoking; ES estimates simply serve to make it appear that something quantitative is being said when in reality no scientific meaning attaches to these figures.

Curvilinearity of regression

A third problem in the interpretation of meta-analytic data is the fact that regressions are often curvilinear, while ES estimation demands linear regressions. Again, an example may be useful.

Glass & Smith (1979) carried out meta-analysis research on class size and achievement and concluded that 'a clear and strong relationship between class size and achievement has emerged'. The study was carried out and analysed well; it might almost be cited as an example of what meta-analysis can do. Yet the conclusion is very misleading, as is the estimate of effect size it presents: 'between class-size of 40 pupils and one pupil lie more than 30 percentile ranks of achievement'. Such estimates imply a linear regression, yet the regression is extremely curvilinear, as one of the authors' figures shows; between class sizes of 20 and 40 there is absolutely no difference in achievement; it is only with unusually small classes that there seems to be an effect. For a teacher, the major result is that for 90% of all classes the number of pupils makes no difference at all to their achievement. The conclusions drawn by the authors from their meta-analysis are formally correct, but they are statistically meaningless and particularly misleading. No estimate of effect size is meaningful unless regressions are linear, yet such linearity is seldom investigated, or, if not present, taken seriously. A simple traditional review would not have made such an obvious error.

A related error is the implicit assumption in the Glass & Smith argument that the observed differences are caused by the variable investigated, i.e. class size. It is possible, or even probable, that small class sizes are more often found in private, or at least upper middle-class schools; there are few classes with single-figure numbers of pupils in inner-city schools! Good schools usually have outstanding teachers, while bad teachers drift towards inner-city schools, where the turnover is very great. Thus the observed differences may not be a result at all of size of class, as assumed by the authors, but of excellence of teaching, associated with size of class. Meta-analysis, by looking only at ES, fails to test alternative theories.

Synergistic interactions

This leads us to another problem with meta-analysis. The typical research paradigm assumes linear regression and orthogonality of relation with other variables. I have indicated how even in an apparently simple case these assumptions may not hold. But worse is to come: we may be dealing with synergistic interactions. Consider the effects of smoking on health; the usual analyses simply contrast smokers with non-smokers, and then use these figures to calculate risk ratios. But there is strong evidence to show that physical factors interact synergistically with smoking, not additively, and that psychosocial factors do the same (Eysenck 1994a). The effect is very strong, and the inevitable implication is that use of only one factor (smoking) for the meta-analysis gives entirely the wrong impression of ES.

Consider a study by Friedman *et al.* (1983), who ascertained smoking habits in patients suffering from myocardial infarction and in controls. They also used a personality questionnaire to sort both groups into four sections depend on proneness to coronary heart disease, from + + (very prone) through + and – to – – (not at all prone). The relative risks (odds ratios) were: + + = 4.4; + = 2.2; – = 1.1, and – – = 0.4. In other words, for the very prone cases and controls, the risk ratio for smoking was 4.4, while for the not at all prone it was 0.4! Thus, depending on the effect of the moderator variable (personality), the effect of smoking on myocardial infarction can be very positive or very negative. Clearly, any simple meta-analysis, with accompanying ES estimates, would be completely incapable of giving a proper idea of the underlying reality. Traditional methods of analysis, or best-evidence synthesis, would have no such problems (Eysenck 1991).

The problem of bad data

A final fault of meta-analysis is derived from one of its oft-proclaimed virtues. It is claimed that ordinary types of analysis leave out 'bad studies' because of alleged faults, but that doing so gives rise to subjectivity in making decisions; all available studies should become part of meta-analysis to avoid such subjectivity. But expert judgement is precisely what the reader should expect from the reviewer – otherwise a simple computer would do the job as well! Subjective evaluation is part and parcel of the special insight which the expert can bring to the discussion, and inclusion of bad studies may completely subvert the true outcome of a hypothetico-deductive analysis. Consider a small-scale example. Schmale & Iker (1971) tested the theory that hopelessness was a predictor of cervical cancer, using a directed interviewing technique and obtaining very positive results. They also administered the Minnesota Multiphasic Personality Inventory and the Rorschach inkblot test, with completely negative results. A minuscule meta-analysis of these three sets of data (and meta-analysts encourage separate analysis of different measures of the independent variable) would show a very small effect size of doubtful significance. Yet the interview was the only procedure relevant to the theory; the tests used are both all-purpose instruments of doubtful reliability and validity. A hypothetico-deductive approach would say the study strongly supported the hypothesis when measures directed at the hypothesis were used to test it; both sets of test results are irrelevant (and would have been even if they had been positive).

This argument will be persuasive to anyone familiar with the critical literature concerning the Minnesota Multiphasic Personality Inventory and the Rorschach test, yet how could a meta-analysis disregard these negative findings, other than by departing from its all-inclusiveness and using what might look like subjective considerations? Undoubtedly, many investigations use multipurpose instruments like these tests to investigate a specific hypothesis for which they are quite unsuited, and negative results so achieved are usually included in meta-analysis of data allegedly relevant to the original hypotheses.

Theory-directed approaches

I have suggested elsewhere (Eysenck 1984, 1992, 1994b) that in many cases a theory-directed approach might be scientifically more predictive than meta-analysis, traditional analysis, or even best-evidence synthesis. In any analysis of data relating to a given theory, one often encounters 'failure to replicate', or even downright contradictory results. Instead of attempting to obtain rough estimates of ES, it might be scientifically more useful to ask the reason for these contradictions. Consider an example (Eysenck 1981). I had put forward a theory according to which eye-blink conditioning should correlate positively with introversion, but not with neuroticism. Spence at Iowa had put forward a theory according to which eye-blink conditioning should correlate with neuroticism, but not with introversion. Both sides published convincing data that proved this theory to give the right prediction! To make matters worse, Amelang published data to show what both were wrong, and that neither introversion nor neuroticism correlated with eye-blink conditioning! (See Eysenck 1981 for references.) What would meta-analysis make of all this? Presumably it would show that both introversion and neuroticism increase eye-blink conditioning, but with very small ES. But such an estimate would disregard the essential disagreement between the studies. Actually the answer was quite simple. Spence had rigged up his test situation in such a way as to produce maximum anxiety in his subjects; hence differences in neuroticism became vitally important and so swamped any effects of introversion. Eysenck had arranged testing so as to minimize anxiety, thus reducing its importance and giving introversion a chance to show its effectiveness. Thus both theories were essentially correct, depending on the manipulation of the testing situation, in a theoretically predictable manner. And Amelang? Both Spence and Eysenck had spent many years building up a laboratory for eye-blink conditioning that was in the forefront of work in this area. Testing was carried out by colleagues with many years of experience. Amelang came into the field without any training or experience, used an inexpensive apparatus which probably did not test eye-blink conditioning at all, and had the testing carried out by inexperienced students. Failure to produce any results could have been predicted.

Of course, simple application of meta-analysis to these data could not have taken these facts into account. Expert knowledge, a determination to solve the problem of contradictory data, and a resolve to go behind the data as published are required. As Sohn (1995) has pointed out, meta-analysis uses printed accounts of experiments as primary data, in the same way as the typical investigator uses results of an experiment to produce these primary data; Lipsey & Wilson (1995) agree that this is so, and find nothing wrong in that. Yet there is a world of difference between these two types of data. We can justifiably average scores over subjects randomly allocated to conditions in our experiment because they are subjected to identical experimental arrangements. But this is not true in the case of meta-analysis of a number of different experiments; there is no random allocation of subjects, and no identity of conditions, as shown only too well in the example of eye-blink conditioning I have given. It is this difference between analysis of primary data and analysis of research reports that is at the basis of the problems many practising researchers have with meta-analysis. It sidesteps all the very real problems in analysis, and gives a pseudo-quantitative answer where conditions are often not yet ripe for such an answer to be meaningful.

Summary

I have criticized meta-analysis on several occasions (Eysenck 1984, 1992, 1994b) because where I have encountered it in the course of my own work in psychotherapy, smoking research and several other areas, I have found it wanting. I am not suggesting that the many criticisms I have made here and elsewhere discredit the method completely. The stress on ES estimation is a step in the right direction, but only if attention is paid to the conditions that must be fulfilled in order to make the estimates applicable. I have singled out homogeneity of samples, conditions and end-points; linear regressions; independence of determinants; absence of moderator variables, and lack of contradictory findings. I have also emphasized attention to criticisms of the experimental paradigm, neglect of which reduces the ES estimates to inconsequential guess-work.

This does not mean that meta-analysis may not be very useful in conditions where a simple drug is being tested for a specific illness, under double-blind conditions. Even in this case we may encounter problems of curvilinearity (dosage effects), and individual differences in reactivity, expectancy, etc., but Type 2 errors are less likely. Even in this case I feel that, where conditions for the use of meta-analysis are most propitious, the need is least pressing. When results of several studies are positive, but not significant because of insufficient statistical power, the position is clear enough not to require meta-analysis, and a statement of ES may be supererogatory, adding spurious assurance to an obvious conclusion. After all, as I have shown, different meta-analyses may lead to characteristically opposite conclusions, as in the case of passive smoking, and no precision in calculating ES estimates can disguise the fact. If this sounds unsympathetic, remember that negative finding are much less likely to be published than positive ones, for various obvious reasons, so that weak ES are likely to be swamped in a sea of unpublished null results (Rosenthal 1979). This point is often missed by enthusiasts for meta-analysis. And finally, if a large number of studies have to be added to give rise to a small ES (parturient montes, nascetur ridiculus mus), many people would feel that the effect might not be worth the expense and the possible danger. Are patients warned of the small effect size of the drug prescribed? It is certainly worth persevering with meta-analysis, but users should be wary of the many dangers in its uncritical use.

References

Bahnson A. & Bliesener T. (1994) Aktuelle Probleme und Strategien der Metaanalyse. *Psychologische Rundschan* **45**, 211–233.

Berkson J. (1938) Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* **33**, 526–542.

Beutler L. (1991) All have won and must all have prizes? Revisiting Luborsky et al.'s verdict. *Journal of Consulting and Clinical Psychology* **59**, 226–232.

Cohen J. (1988) *Statistical Power Analysis for the Behavior of Sciences*. Erlbaum, Hillsdale, N.J.

Cohen J. (1994) The earth is round ($p < 0.05$). *American Psychologist* **49**, 997–1003.

Cooper H.U. & Hedges L.V. (eds) (1994) *The Handbook of Research Synthesis*. Russell Sage Foundation, New York.

Eysenck H.J. (1978) An exercise in mega-silliness. *American Psychologist* **33**, 517.

Eysenck H.J. (ed.) (1981) *A Model for Personality*. Springer Verlag.

Eysenck H.J. (1984) Meta-analysis: an abuse of research integration. *Journal of Special Education* **18**, 41–59.

Eysenck H.J. (1991) *Smoking, Personality and Stress: Psychosocial Factors in the Prevention of Cancer and Coronary Heart Disease*. Springer, New York.

Eysenck H.J. (1992) Meta-analysis, sense or non-sense? *Pharmaceutical Medicine* **6**, 113–119.

Eysenck H.J. (1994a) Synergistic interaction between psychosocial and physical factors in the causation of lung cancer. In *The Psychoimmunology of Cancer* (eds C. Lewis, C. O'Sullivan & J. Barraclough), pp. 163–178. Oxford University Press, Oxford.

Eysenck H.J. (1994b) Meta-analysis and its problems. *British Medical Journal* **309**, 789–792.

Eysenck H.J. (1995) Meta-analysis squared – does it make sense? *American Psychologist* **50**, 110–111.

Fleiss J.L. & Gross A.J. (1991) Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: A critique. *Journal of Clinical Epidemiology* **44**, 439.

Friedman G., Bruce H., Fireman D., Petite D., Siegelant, A., Ury H. & Klatsky A. (1983) Psychological questionnaire score, cigarette smoking, and mycocardial infarction: A continuing enigma. *Preventive Medicine* **12**, 533–540.

Glass G. & Smith M.L. (1979) Meta-analysis of research in class size and achievement. *Educational Evaluation and Policy Analysis* **1**, 2–16.

Grawe K., Donati R. & Bernauer F. (1994) *Psychotherapie im Wandel*. Hogrefe.

Hogben L. (1957) *Statistical Theory*. Allen & Unwin, London.

Hunter J.E. & Schmidt F.L. (1990) Methods of Meta-Analysis. Newbury Park, Sage.

Huque M.E. (1988) Experience with meta-analysis in NDA submissions. *Proceedings of the Biopharmaceutical Section of the American Statistical Association* **2**, 28–33.

Lee P.N. (1988) *Misclassification of Smoking Habits and Passive Smoking*. Springer Verlag, New York.

Lee P.N. (1992) *Environmental Tobacco Smoke and Mortality*. Karger, London.

Lipsey M. & Wilson D. (1993) The efficacy of psychological educational, and behaviour treatments. Confirmation from meta-analysis. *American Psychologist* **48**, 1181–1209.

Lipsey M. & Wilson D. (1995) Reply to comments on Lipsey and Wilson. *American Psychologist* **50**, 113–115.

Lykken D.E. (1968) Statistical significance in psychological research. *Psychological Bulletin* **70**, 151–159.

Morrison D.E. & Henkel E.R. (eds) (1970) *The Significance Test Controversy*. Aldine, Chicago.

National Research Council (1986) *Environmental Tobacco Smoke: Measuring Exposure of Assessing Health Effects*. National Academy Press, Washington, D.C.

Rosenthal R. (1979) The 'file drawer problem' and tolerance for null results. *Psychological Bulletin* **83**, 638–641.

Schmale H.A. & Iker H. (1971) Hopelessness as a mediator of cervical cancer. *Social Science & Medicine* **5**, 55–100.

Sedlmeier P. & Gigerenzer G. (1989) Do studies of statistical power have an effect on the power of statistics? *Psychological Bulletin* **105**, 389–316.

Slavin R.E. (1986) Best evidence synthesis: an alternative to meta-analysis and traditional reviews. *Education Research* **15**, 9–11.

Smith M., Glass G. & Miller I. (1980) *The Benefits of Psychotherapy*. John Hopkins Press, Baltimore.

Sohn D. (1995) Meta-analysis as a means of discovery. *American Psychologist* **50**, 108–110.

Spitzer W.O., Lawrence V. & Dales R. (1990) Links between passive smoking and disease: A best evidence synthesis. *Clinical and Investigative Medicine* **13**, 17–42.

Stein R.A. (1988) Meta-analysis from one FOA reviewer's perspective. *Proceedings of the Biopharmaceutical Section of the American Statistical Association* **7**, 34–38.