Psychol. Belg.
1986, XXVI-2, 185-197.

IN SEARCH OF AN INTERPRETATION OF META-ANALYTIC FINDINGS

Iven VAN MECHELEN

Catholic University of Leuven

Meta-analysis was introduced as an effective means to integrate research findings. With regard to its definition and interpretation, however, it has remained rather controversial. If meta-analysis is defined as a series of review methods including statistical analysis and/or synthesis of the summary statistics of the studies considered, a statistical nucleus becomes evident. The relationship between summary statistics (e.g. significance levels and effect sizes) and their psychological interpretation, on the other hand, is unclear. It is therefore argued that in many instances it is impossible to link a psychological interpretation to the results of meta-analytic operations.

Nowadays researchers are facing the challenge of digesting an ever increasing and insurmountable stack of publications. In order to process this bulk of information, one should have recourse to serviceable expedients. As a response to this call for such expedients for the analysis and/or synthesis of aggregates of research data, from the middle of the 1970's onwards several authors (Rosenthal, 1978; Smith & Glass, 1977) have developed a set of techniques, known as meta-analysis. This endeavour, which first started off in a rather modest fashion, eventually grew out to become what has recently been referred to as the "meta-analytic revolution" (Fiske, 1983).

Meta-analysis has been applied to several major psychological research areas (for an overview see Green & Hall, 1984) as well as to other disciplines. Leading journals have paid attention to its study, surveys and handbooks of meta-analysis have appeared (Glass, McGaw & Smith, 1981; Hunter, Schmidt & Jackson, 1982; Rosenthal, 1980). The "meta-analytic revolution", however, has also provoked serious criticism (Abeles, 1981; Eysenck, 1978; Kazrin, Durac & Agteros, 1979; Presbey, 1978; Rachman & Wilson, 1980; Strube & Hartmann, 1982, 1983). To date, meta-analysis, to put it mildly, has remained a rather controversial issue.

This situation creates confusion amongst practitioners, who find no clear-cut answers to basic questions such as: What exactly is meta-

The author is research assistant of the National Fund for Scientific Research (Belgium). Thanks are due to L. Delbeke, E. Van Avermaet and S. Van den Broucke for their helpful comments on earlier drafts of this article.

analysis? How does it work? How should one interpret meta-analytic conclusions such as: "The average person who receives therapy is better off at the end of it than 80 percent of the persons who do not." (Smith, Glass & Miller, 1980, p. 87)? (This conclusion was based upon a meta-analysis of 475 controlled psychotherapy outcome studies.)

The present paper was conceived to provide some clarification regarding the three questions raised above. First, a number of alternative definitions of meta-analysis will be discussed. Next, its statistical framework will be presented. Together with the latter the psychological interpretation of the results of meta-analytical operations will be explored.

Definition

Upon their presentation, meta-analytic techniques have frequently been given different names, including "research integration", "quantitative judgment of research domains", and "quantitative review" (Green & Hall, 1984). In addition to this difference in labeling, several definitions for meta-analysis have been reported in literature. The lack of equivalence of these definitions has given rise to a vivid discussion among their respective proponents (Cooper & Arkin, 1981; Leviton & Cook, 1981). However, an unequivocal definition of meta-analysis is necessary as a starting point for a further explanation and discussion. Therefore, I will dwell for a while on the issue of how to define meta-analysis.

Meta-analysis is part of the research field focussing on the analysis and/or synthesis of aggregates of studies reported in the literature. To define meta-analysis, then, one first has to explore the process of reviewing an aggregate of studies (see also Smith, Glass & Miller, 1980). Seven consecutive steps can be distinguished here: (a) First a research question or hypothesis is chosen. (b) Next in the total field of literature the area to be included in the analysis/synthesis is selected. (c) A method is chosen to look up subject-related studies reported in literature. (d) A sample of studies is taken from the chosen research area. (e) From each study included in the sample a number of elements needing further specification are drawn. These elements can be related to research results as well as to methodological and substantive characteristics of the study (Glass, McGaw & Smith, 1981). (Note that in this context substantive means "specific to the content of the problem considered".) The collection and organisation of these elements require the construction of a classification system, a taxonomy (Fiske, 1983; Glass & Kliegl, 1983). (f) The collected elements are treated using the proper analysis and/or synthesis technique. (g) Finally, conclusions are drawn from the results of (f).

Each step allows for a number of options among which a choice is made more or less arbitrarily. For example, in order to trace studies (step c) several data bases may be used, each with a different reliability level (Glass, McGaw & Smith, 1981; Green & Hall, 1984). When collecting a study sample (step d) one can aim at achieving either completeness (i.e., try to map out the whole domain) or representativeness (e.g., random sampling). For the construction of a taxonomy (step e) several choices are available. With respect to methodological characteristics several authors (Strube & Hartmann, 1982, 1983; Shapiro & Shapiro, 1983) refer to the distinction between statistical conclusion, internal, construct, and external validity introduced by Cook and Campbell (1979); yet this framework can be filled up in many ways. The application of the analysis and synthesis techniques, as mentioned in (f), allows for several options also. Among them is the importance to be attached to methodological study characteristics in the aggregation of study results. One may, for example, exclude studies from analysis on account of methodological considerations — Smith, Glass, and Miller (1980) called this the "impeachment procedure" — or use methodological elements to construct a weighting procedure. One may also simply carry out the analysis or synthesis of study results first and relate the findings to methodological characteristics afterwards. Generally speaking, it should be pointed out that all options taken by a reviewer can be more or less implicit, public and replicable, and that a number of them may imply more or less quantification. Whereas options *are* necessary, it is important to realize that they define the particular reviewing method that will be used.

While acknowledging the process of reviewing studies reported in literature and its implications, let us now consider a number of formulations put forward to define meta-analysis. Smith et al. (1980) state that:

> [Meta-analysis] was designed to satisfy three basic requirements: (1) studies should not be excluded from consideration on arbitrary and a priori grounds; boundaries must be drawn around fields, but it is better to draw them wide than narrow; (2) study findings should be transformed to commensurable expressions of magnitude of experimental effect or correlational relationship; (3) features of studies that might mediate their findings should be defined, measured, and their covariation with findings should be studied. (...) [One] must begin with the results of data analyses (means,

> variances, t-tests, and the like) that are typically reported by researchers. An integrative analysis is then carried out on the results of the primary data analyses. (p. 39).

Cooper and Arkin (1981) give a much simpler definition, suggesting that meta-analysis is a kind of literature review characterized primarily by its emphasis on quantification. Christensen and Reick (1983) refer to "the use of data analytic procedures, practices, or techniques where multiple studies or sets of data are used" (p. 33).

Those definitions do not only differ with respect to comprehensiveness and clarity, but also represent different options for the consecutive steps of the review process as outlined above. Since ultimately one is free to fill out these options, discussions about such definitions of meta-analysis (Cooper & Arkin, 1982; Leviton & Cook, 1981) are somewhat academic. Nevertheless it can be argued that a combined definition (like that of Smith et al.), containing a number of options which are mutually independent — i.e., one needs not take them together — is not a convenient starting-point for a settlement of the controversy about meta-analysis. Therefore, we suggest that, in line with Cook and Leviton (Cook & Leviton, 1980; Leviton & Cook, 1981), only two central conditions should be retained to define meta-analysis: (1) (step [e] of the review process) Summary statistics (e.g., significance levels or effect sizes) of the studies are collected or derived from the data mentioned in the study reports. (2) (step [f] of the review process) Statistical operations of analysis and/or synthesis are applied to the elements mentioned in (1).

It is obvious from the preceding statements that statistical procedures are of major importance in meta-analysis. In the remainder of this paper this statistical basis will be discussed further, together with the psychological interpretation of the results of the statistical operations.

BUILDING BLOCKS OF META-ANALYSIS

According to the definition presented above meta-analysis starts from summary statistics, which in fact constitute the building blocks of the technique. We shall first focus on the two summary statistics that are most frequently used in meta-analysis (i.e., significance level and effect size), and then discuss their aggregation (analysis/synthesis) in the next paragraph.

Building block one: Significance

Significance levels (p) indicate the risk of wrongly rejecting a correct null hypothesis (H_0) (given a particular statistical model). They are also used (often incorrectly) to compare results (e.g., "result A is more significant than result B"). A presentation of as many "tabular asterisks" as possible is popular among authors as means to prove the strength of their findings (Meehl, 1978).

The exaggerated importance attached to significance levels has been denounced by many commentators. In nature exact null relations occur only rarely (Cooper, 1981). Authors nearly always start from false null hypotheses (Meehl, 1978) and they are evidently prejudiced against H_0 (Greenwald, 1975). Accepting or rejecting H_0 with a specific significance level is relatively uninformative (Cooper, 1981), as is exemplified by the distinction between statistical and psychological significance. Moreover, since a level of significance is dependant upon sample sizes it is not a convenient parameter to measure the magnitude of an effect.

Therefore, several researchers looked for alternative solutions that could replace significance level as an index of effect strength. Meehl (1978) has argued strongly in favor of a confirmatory approach; Kazdin (1978) has proposed so-called social validation techniques to determine "clinical" significance, such as comparison with norms of "nondeviant peers".

It is evident that one must further inspect measures of statistical and psychological effect size in addition to *p*-levels. Significance level and effect size, then, do not include each other, but provide complementary information (Cooper, 1981).

Building block two: Effect size

As I have pointed out above, statistical significance is not a very suitable measure to determine the strength of relations. Therefore, applied statisticians have developed measures of effect size (ES). Many (kinds of) measures have been proposed, and the task of making a good choice amongst this jungle of possibilities is certainly not an easy one. In order to draw up a pragmatic classification of the proposed measures one may investigate them starting from three different angles: (a) the type of scale that is used to express the measure (Sechrest & Yeaton, 1982); (b) the underlying statistical model (Glass & Hakstian, 1969; Hedges, 1981); and (c) the psychological interpretation.

Types of scales. Four types of scales of ES can be distinguished, but they are not all equally suitable for every design. For instance, the first two types of scales will be more easily applicable for measuring the strength of relationship between a dichotomous independent variable X (e.g., experimental vs. control condition) and a continuous dependent variable Y, than in the case of a polytomous independent variable. The strength of relationship can be expressed by: (1) a standardized mean difference (in the situation mentioned above $\delta = (\mu_E - \mu_C)/\sigma_C$ where μ_E, μ_C are the population means of the experimental and control groups respectively, and σ_C is the population standard deviation of the control group) — amongst other measures, Cohen's *d* (Cohen, 1977) and Glass' *d* (Smith et al., 1980) fall in this category —; (2) (in the situation mentioned above) a percentage of nonoverlap of the density distributions of the two conditions of X (Cohen, 1977); (3) the percentage of variance in Y accounted for by X — Hays' η^2, ω^2 and Kelley's ε^2 belong to this category (Glass & Hakstian, 1969) —; (4) a correlational quantity — this category contains, amongst others, Friedman's r_m, C, φ, r_{pb} (Cohen, 1965; Sechrest & Yeaton, 1982) —.

It is possible to move from one scale to another. Several authors have drawn up transformation tables for this purpose (Cohen, 1977; Friedman, 1968; Glass, McGaw & Smith, 1981).

Underlying statistical models. When investigating ES measures with regard to an underlying statistical model it is important to keep in mind the distinction between sample and population. In general, one is interested in the estimation of population ES's departing from sample ES's. In the situation with an experimental and control condition (as outlined in the preceding paragraph), for instance Glass' sample ES *d* is an estimator of the population ES δ (with $d = (M_E - M_C)/S_C$, where M_E and M_C are the sample means for the experimental and control groups, S_C the sample standard deviation of the control group). To estimate the same population ES different sample ES measures may be used (e.g., Cohen's *d* is also an estimator of the same δ). Being an estimator of the same population ES, then, is one possible connection between measures of ES. Moreover each such estimator may have a (different) bias. Provided the traditional assumption that experimental and control groups are independently and normally distributed with common variance, neither Glass' *d* nor Cohen's *d* are unbiased estimators of δ (Hedges, 1981). On the other hand, seemingly related sample ES measures may estimate essentially different population parameters (Glass & Hakstian, 1969; Sechrest & Yeaton, 1982).

Psychological interpretation. One can attempt to make an intuitive interpretation of ES. The problem is that different ES measures create different intuitive impressions of effect: 10% of variance accounted for corresponds to a correlation of .32 and to the kind of relationship between dichotomous independent and dependent variables as represented in table 1.

Tab. 1 — Hypothetical 2 × 2 frequency table with association coefficient φ = .32

	dependent variable		
independent variable	0	1	Σ
0	66	34	100
1	34	66	100
Σ	100	100	200

Some authors have therefore proposed to express ES in a uniform manner, such as the percentage of variance accounted for (Sohn, 1980). Cooper (1981), on the other hand, argued that one would be better off acknowledging the (nonintuitive) equivalence of different ES measures.

It appears one needs a better basis for ES interpretation than mere intuition. To this end one can look for norms. Cohen (1977) suggested to settle on a convention to classify ES's in the behavioral sciences into "small", "medium", and "large", using critical Pearson r-values of .10, .30, and .50 respectively. This classification is arbitrary. Cooper (1981) judged it to be too broad a categorisation. Moreover, a firm ground for its application is lacking (Sechrest & Yeaton, 1982). An alternative (or support) for intuitive norms is to make an appeal to results of meta-analyses (Yeaton & Sechrest, 1981).

The aforementioned norms, however, are just an abstract point of reference for interpretation. They do not suffice for the determination of the practical importance of an effect. To catch this "practical importance" Rosenthal and Rubin (1979b, 1982) and Rosenthal (1983) propose to convert ES's into a frequency table like the one presented in Table 1, with, in the example of psychotherapy outcome research, therapy/no therapy as the independent variable and alive/dead as the dependent variable. (Such a table is called a "Binomial Effect Size Display".) However, this proposal disguises the problem of interpretation by fixing an independent and dependent variable, for the psychological meaning of an effect is often determined by the very nature of the independent and dependent variables. With respect to the independent variable this may be, amongst others, the "strength" of the manipulation carried out (see e.g., Yeaton & Sechrest, 1981; O'Grady, 1982). Consider as an example controlled psychotherapy outcome research with the comparison between a therapy and control group. Assume that for two such comparisons the statistical ES's are identical, but that for one case the duration of therapy amounts to 10 hours and for the other case to 40 hours. In these two cases the same statistical ES will probably get a different psychological interpretation.

As to the dependent variable there are also complications. Take again as an example two comparisons between therapy and control groups. Again, assume that the ES's are identical. Furthermore, assume that for each of the studies the dependent variable is dichotomous, but that in one case the classes are severe depressive/moderate depressive and in the other case severe depressive/"normal" mood. Again the same statistical ES will probably get different psychological interpretations.

In summary, we may conclude that interpretation of ES's is a complex matter, for which an instant solution is not available at present. It is evident, however, that the relationship between a statistical ES and its psychological interpretation is often not one to one: a particular statistical ES can have more than one psychological meaning and vice versa.

Analysis/Synthesis

Now that the building blocks of meta-analysis have been identified and commented upon we can move to the second step: the analysis and/or synthesis of the results of the primary analyses. The situation which we are dealing with can roughly be represented as follows: a number of studies have been carried out, each involving one or more tests; from each test a basic value, usually a significance level (p-value) or an ES, is retrieved; the aggregation of these basic values should lead to a conclusion.

As was the case with the building blocks of meta-analysis there are also different techniques of aggregation. Rosenthal (1983) classified them on the basis of the analytic process for which they are intended, that is a *combination* or a (directed or nondirected) *comparison* of study results. A *combination* of p-values allows for an estimated risk to reject wrongly the common null hypothesis of the studies involved (Rosenthal, 1978). Further, a *comparison* of p-values allows to examine so-called differences in statistical significance (Rosenthal & Rubin, 1979b). A *combination* of ES's may be used to estimate the population ES, which is assumed to be identical for all studies (c.q. tests) (Hedges, 1981). Finally, a *comparison* of ES's may serve as a method to examine the heterogeneity of the population ES's of studies (or groups of studies with assumed common population ES within each group) (Hedges, 1982a, 1982b). In addition to these rather simple aggregation techniques more complex ones are also available, such as multivariate procedures. One could, for instance, determine the relationship between ES's and several independent variables by means of multiple regression (Hedges, 1982c), provided that the variance of ES's across studies is not entirely due to artifacts such as sampling error and measurement error (Hunter, Schmidt & Jackson, 1982). Eventually each aggregation procedure comes down to either a hypothesis testing or a parameter estimation procedure. The use of these statistical notions implies that combinations and comparisons in meta-analysis must always be based upon a statistical model.

In most of the meta-analytic models reported thus far, three basic assumptions prevail. They concern: (a) the sample sizes of the studies involved; (b) the independence of the basic values; and (c) commensurability.

Sample sizes. Meta-analytic models generally require "large" sample sizes (Hedges, 1981, 1982a, 1982b, 1982c; Rosenthal, 1978; Rosenthal & Rubin, 1979a, 1982). Yet Hedges (1982a, 1982b, 1982c) referred to computer simulations to support the accuracy of his computations for sample sizes up to n = 10.

Independence of basic values. Every meta-analytic model is based upon the assumption that the basic values are derived from independent observations. In practice, however, this assumption of independence is often violated. The meta-analysis of controlled psychotherapy outcome studies reported by Smith, Glass, and Miller (1980), for example, contained several ES's per study (475 studies yielded 1776 ES's); other dependencies also contaminated this review (see Landman & Dawes, 1982). Although Glass, McGaw and Smith (1981) and Strube (1985) have provided partial solutions for the problem of nonindependance, a completely satisfactory statistical strategy is still lacking.

Commensurability. This is perhaps the most essential, but at the same time one of the most controversial assumptions of meta-analysis. The core of this assumption is that the aggregation of the basic values must be possible and "meaningful". For example, a combination of *p*-values does not make any sense unless the null hypotheses associated with these values are identical.

Smith et al. have frequently been criticized for not respecting the commensurability condition and thus for mixing "apples and oranges"

(Rachman & Wilson, 1980). In reply Smith and her associates (1980) stated that if one is interested in studying "fruit" as such, a mixing of apples and oranges can indeed be meaningful. Evidently what matters here is the choice of a level of abstraction (see also Glass & Kliegl, 1983).

Rather than getting carried away by this fruit discussion I will discuss the issue of commensurability starting from a statistical point of view. Therefore we need an explicit underlying statistical model. Suppose (Hedges, 1982a; Hedges & Olkin, 1983) our data consist of a series of k independent studies, each of which involves a comparison of an experimental group (E) with a control group (C). Let Y_i^E and Y_i^C be the distributions of the experimental and control groups, respectively, where the subscript i refers to the i-th experiment. For fixed i, we assume that Y_i^E and Y_i^C are independently and normally distributed with means μ_i^E and μ_i^C and common variance σ_i^2; that is $Y_i^E \sim (\mu_i^E, \sigma_i^2)$ and $Y_i^C \sim (\mu_i^C, \sigma_i^2)$, $i = 1,\ldots k$. The (population) ES of the i-th experiment can be defined as $\delta_i = (\mu_i^E - \mu_i^C)/\sigma_i$. An additional possible assumption for the model is that all the population ES's are equal, that is $\delta_1 = \delta_2 = \ldots = \delta_k = \delta$. This equation is a condition for statistical commensurability. A combination of the k (sample) ES's will aim at estimating the common population ES δ. (As the common δ assumption is not trivial, control is necessary. Hedges (1982a, 1982b) has introduced testing procedures to check the heterogeneity of ES's.)

For a psychological interpretation, however, statistical commensurability does not suffice. As has been demonstrated in the preceding section, a one to one correspondence between statistical and psychological ES can probably be ruled out. The psychological significance of the k (statistical population) ES's may therefore differ (e.g., on account of considerations on the "strength" of the independent variables), even if the common δ assumption holds. In such cases no meaningful psychological interpretation links up with the statistical estimation of the common population ES.

Instead of the common δ assumption, a random effects model may also be introduced (Hedges, 1983). This would imply the assumption that the k population ES's are not fixed, but instead are sample realizations retrieved from a distribution (Δ) of possible population ES's. The mean of this distribution (i.e. $\bar{\Delta}$) can then be estimated. However, it is easy to show with an argument similar to the one outlined above, that a psychological interpretation of such an estimate will be even far more difficult.

CONCLUSION

Meta-analysis, as part of a movement that started in the 1970's, encompasses more than a cluster of techniques. Among other things, it also concerns a thorough analysis of the reviewing process and a critical evaluation of several more or less traditional methods to integrate research findings. In addition, proposals are made to overcome deficiencies of techniques used in the past.

In this contribution meta-analysis was defined rather narrowly as a technique based upon two statistical options. This definition allows for a limited yet more thorough appraisal.

As principal building blocks of meta-analysis, significance values and effect sizes bear relevance to both a statistical and a psychological context. Their use requires a rigorous statistical model as well as an adequate psychological judgment. Yet, a coherent statistical/psychological interpretative system is not available at present. The fact that the relationship between statistical and psychological effect size is frequently not one to one often prevents the ascription of a psychological interpretation to meta-analytic findings.

Eventually, everything comes down to the lack of an accurate definition of what is psychologically significant and by which variables this is determined. With such a definition at hand a meaningful meta-analysis could become possible, either by restoring the one to one relationship by means of studying groups that are homogeneous with respect to the determining variables, or by creating new statistical ES's that take those variables into account.

REFERENCES

Abeles, N. (1981). Psychotherapy works! *Contemporary Psychology, 26,* 821-823.

Bandura, A. (1978). On paradigms and recycled ideologies. *Cognitive Therapy and Research, 2,* 79-103.

Christensen, E.R., & Reick, D.T. (1983). Meta-analysis in mental health treatment outcome evaluation. In M.J. Lambert, E.R. Christensen & S.S. De Julio (Eds.), *The assessment of psychotherapy outcome* (pp. 33-55). New York: Wiley.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-Experimentation. Design and analysis issues for field settings.* Chicago: Rand McNally.

Cook, T.D., & Leviton, L.C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality, 48,* 449-472.

Cooper, H.M. (1981). On the significance of effects and the effects of significance. *Journal of Personality and Social Psychology, 41,* 1013-1018.

Cooper, H.M., & Arkin, R.M. (1981). On quantitative reviewing. *Journal of Personality, 49,* 225-230.

Eysenck, H.J. (1978). An exercise in mega-silliness. *American Psychologist, 39,* 517.

Fiske, D.W. (1983). The meta-analytic revolution in outcome research. *Journal of Consulting and Clinical Psychology, 51,* 65-70.

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin, 70,* 245-251.

Garfield, S. (1981). Psychotherapy: A 40-year appraisal. *American Psychologist, 36,* 174-183.

Glass, G.V., & Hakstian, A.R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal, 6,* 403-414.

Glass, G.V., & Kliegl, R.M. (1983). An apology for research integration in the study of psychotherapy. *Journal of Consulting and Clinical Psychology, 51,* 28-41.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Green, B.F. & Hall, J.A. (1984). Quantative methods for literature reviews. *Annual Review of Psychology, 35,* 37-53.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1-20.

Hedges, L.V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107-128.

Hedges, L.V. (1982a). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490-499.

Hedges, L.V. (1982b). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics, 7,* 119-137.

Hedges, L.V. (1982c). Fitting continuous models to effect size data. *Journal of Educational Statistics, 7,* 245-270.

Hedges, L.V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93,* 388-395.

Hedges, L.V., & Olkin, I. (1983). Clustering estimates of effect magnitude from independent studies. *Psychological Bulletin, 93,* 563-573.

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

Kazdin, A.E. (1978). Methodological and interpretative problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology, 46,* 629-642.

Kazrin, A., Durac, J., & Agteros, T. (1979). Meta-meta analysis: A new method for evaluating therapy outcome. *Behaviour Research and Therapy, 17,* 397-399.

Landman, J.T., & Dawes, R.M. (1982). Psychotherapy outcome: Smith & Glass's conclusions stand up under scrutiny. *American Psychologist, 37,* 504-516.

Leviton, L.C., & Cook, T.D. (1981). What differentiates meta-analysis from other forms of review? *Journal of Personality, 49,* 231-236.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806-834.

O'Grady, K.E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin, 92,* 766-777.

Presbey, S. (1978). Overly broad categories obscure important differences between therapies. *American Psychologist, 33,* 514-515.

Rachman, S.J., & Wilson, G.T. (1980). *The effects of psychological therapy* (2nd enlarged ed.). New York: Pergamon Press.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185-193.

Rosenthal, R. (Ed.). (1980). *New directions for methodology of social and behavioral sciences: Quantitative assessment of research domains.* San Francisco: Jossey-Bass.

Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. *Journal of Consulting and Clinical Psychology, 51,* 4-13.

Rosenthal, R., & Rubin, D.B. (1979a). Comparing significance levels of independent studies. *Psychological Bulletin, 86,* 1165-1168.

Rosenthal, R., & Rubin, D.B. (1979b). A note on percentage of variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology, 9,* 395-396.

Rosenthal, R., & Rubin, D.B. (1982). A simple general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166-169.

Sechrest, L., & Yeaton, W.H. (1982). Magnitudes of experimental effects in social science research. *Evaluation Review, 6,* 579-600.

Shapiro, D.A., & Shapiro, D. (1983). Comparative psychotherapy outcome research: Methodological implications of meta-analysis. *Journal of Consulting and Clinical Psychology, 51,* 42-53.

Smith, M.L., & Glass, G.V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752-760.

Smith, M.L., Glass, G.V., & Miller, T.I. (1980). *The benefits of psychotherapy.* Baltimore: The Johns Hopkins University Press.

Sohn, D. (1980). Critique of Cooper's meta-analytic assessment of the findings on sex differences in conformity behavior. *Journal of Personality and Social Psychology, 39,* 1215-1221.

Strube, M.J. (1985). Combining and comparing significance levels from nonindependent hypothesis tests. *Psychological Bulletin, 97,* 334-341.

Strube, M.J., & Hartmann, D.P. (1982). A critical appraisal of meta-analysis. *British Journal of Clinical Psychology, 21,* 129-139.

Strube, M.J., & Hartmann, D.P. (1983). Meta-analysis: techniques, applications and functions. *Journal of Consulting and Clinical Psychology, 51,* 14-27.

Yeaton, W.H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity and effectiveness. *Journal of Consulting and Clinical Psychology, 49,* 156-167.

Department of Psychology
Tiensestraat 102
3000 Leuven

Received May 1986