British Education Research Journal, Vol 10, Issue 2, Pages 135-144, 1984

Meta-analysis: an explication

C.T. FITZ-GIBBON, University of Newcastle-upon-Tyne
(now *University of Durham*)

ABSTRACT *Meta-analysis, as developed over the last few years by Glass and others, is a quantitative method for synthesising research results. Its use is illustrated here by means of examples ranging from irreproachable to dubious. Being simple to use and easily understood, meta-analysis will undoubtedly become popular and this increasing use may well bring about some notable changes. The well-controlled, small-scale experiment is likely to become more important and better use will be made of many existing research reports, reports which can now be dusted off and incorporated into meta-analyses. Because meta-analysis focuses on how much difference something makes (the magnitude of an effect) and not on whether or not the difference was statistically significant at a pre-specified level, its use encourages a more scientific approach to the interpretation of quantitative results. It also offers some hope that we might eventually have a clearer idea of the conditions under which research findings can be generalised. Progress in this direction will require mutual support between quantitative and qualitative research methods.*

Explication

Educational researchers may have produced another best seller in the social sciences, a book which, like the Campbell & Stanley (1966) classic, will be widely cited in journals of many disciplines as well as having long-lasting effects on the methods of educational research. The publication of *Meta-analysis in Social Research* (Glass, McGaw & Smith, 1981) follows some years of work during which the concepts of meta-analysis have been developed and applied to a rapidly growing number of research topics. To give some examples from the field of education, meta-analysis has been used to synthesise the empirical evidence regarding the effects on achievement of:

four methods of individualizing Mathematics instruction (Hartley, 1977);
racial desegregation (Krol, 1979);
class size (Glass & Smith, 1979);
modern versus traditional mathematics instruction (Athappilly, 1980);
cognitive levels of teachers' questions (Redfield & Rousseau, 1981);
peer tutoring (Cohen, Kulik & Kulik, 1982);
ability grouping of secondary school students (Kulik & Kulik, 1982).

The use of meta-analysis has not, however, been without controversy. "An exercise in mega-silliness" Eysenck (1978) called it. It will be suggested below that existing meta-analyses might be assessed along a continuum, ranging from totally acceptable through dubious to unacceptable. The explanation of how such assessments might be made will provide an introduction to meta-analysis for anyone not familiar with this development. This introduction assumes no more than a passing acquaintance with statistical technique since it is hoped that those who do not usually work in quantitative research will find this article of interest. The implications of meta-analysis (MA), the topic of the second half of this article, concern the entire research community. Moreover, one of the great virtues of meta-analysis is that it rests essentially on a measurement, 'Effect Size', which can be understood in the first week of a statistics course or by anyone who can remember what a *z*-score is.

Meta-analysis is essentially the application of statistical techniques to the task of synthesising research findings. Several approaches to the task of synthesis are possible within the broad analytical strategy (Cooper, 1982), but the term meta-analysis has become associated with the approach which was pioneered primarily by Glass and which utilises the concept of an 'Effect Size'.

Take a very simple example of the need to synthesise results: Fig. 1(a) shows the effect on measured achievement of two methods of teaching, say method A and method B. Pupils were randomly assigned to either method, hence their equivalence at pre-test. At post-test three weeks later, method A pupils scored higher than method B pupils, on average, and about three months later method A pupils still had scores above their pre-test scores whereas method B pupils had more or less returned to ground zero.

Traditionally researchers have been trained to ask immediately "But were the results statistically significant?" Significance at the 0.05 level has been a hurdle assigned prime importance in which Carver (1978) has designated "A corrupt form of the scientific method". Regardless of the statistical significance or insignificance of the difference between the two post-test means in Fig. 1(a) this difference would still represent the best available evidence of the effect of the treatment in that situation and would provide, therefore, an estimate of the magnitude of the effect. It is the *magnitude of the effect* and the interpretation of how *educationally significant* this was which are the most important pieces of information to arise from an experiment. The educational significance of an effect requires a judgment made in terms of the outcome measure, a procedure which is important enough generally, and to this discussion in particular, to need a special term. Let it be called *interpreting the metric*. The meaning can be made clear by an example: suppose that the difference between the two post-test scores represented just one more item correct after three weeks of considerable effort. One would not be particularly impressed by the result, whatever the statistical significance – unless there were some other factor, such as each item being incredibly difficult for the group of pupils concerned. On the other hand, a gain which brought performance of the bottom of the class up to the level usually achieved by the top of the class, might be judged educationally significant even though, because of small numbers it might not be statistically significant. In short, to interpret the educational significance of a difference between two groups one must *interpret the difference in terms of the metric* in which the outcomes were measured. The size of the obtained difference must then be evaluated in relation to the nature of the treatment (its costs in terms of time and effort, its feasibility, acceptability and so on). Often it is necessary to refer to prior expectations in order to produce a judgment as to how good the results were. To examine the magnitude of the effect in this way is far more important than discussing the statistical significance, which depends too heavily on sample size.

[Insert Figure 1]

But would not such an approach leave us prey to interpreting one-off, unstable results which would not replicate? Yes indeed, but the solution to this problem lies not in striving for a single instance of statistical significance at the magic 0.05 level but in conducting actual replications. Only replication will establish replicability, generalisability. The experiment in which method A was contrasted with method B was in fact replicated with three other classrooms. The results are shown in Fig. 1(b, c, d). In each case the outcome was in the same direction and, regardless of the statistical significance of any one result, the collection of results begins to seem persuasive.

How should the four results be interpreted? Table I presents the *t*-test for each difference between means. One approach, which Light and Smith have called 'vote counting' (Light & Smith, 1971) is to note that two out of the four experiments yielded results statistically significant at the 0.05 level. Should we, then, regard the treatment as having an equal chance of succeeding or failing, since two results were statistically significant and two were not? Such a conclusion ignores both the dependence of statistical significance on sample size and the overall consistency in the direction of the results. Moreover, the *coup de grace* for vote counting research reviews has been administered by Hedges & Olkin (1980) who showed that type II errors actually become *more* likely the larger the sample of studies analysed. (A Type II error is the error of failing to recognise a true difference.)

TABLE 1: Results of four replications

Experiment	*t*	df	One tail probability	$-\ln P$
Class 1	1.84	18	0.04	3.219
Class 2	1.47	14	0.08	2.526
Class 3	4.02	13	0.003	5.809
Class 4	1.13	18	0.14	1.966
Sum				13.520

As explained in Winer (1971, p. 49)

$X^2 = 2\ \Sigma\ (-\ln P_i)$ and

df = 1 (*n*).

Thus $X^2 = 2(13.520) = 27.04$, df = 2(4) = 8, $p<0.005$

A better approach might be to ask 'How likely is it that these four results would have been obtained were there really no difference between method A and method B?' Jones & Fiske (1953) described Pearson's chi square procedure for assessing this probability, and this has been applied in Table I yielding, for the four studies, a probability of less than 0.01. (It is unfortunate that this simple procedure, which has been available for so long, is not included in elementary statistical texts. Its omission is symptomatic of the lack of emphasis on *replication*. Replication should be presented as an essential feature of hypothesis testing and, indeed, of any line of research.) Rosenthal (1978) summarised other statistical methods of combining probabilities from repeated experiments. However, whilst useful to some extent, approaches which deal with aggregating probability levels, lose the most important information: the magnitude of the effect.

The approach to such a set of data in a meta-analysis is to compute the *Effect Size* for each experiment. For comparing the mean of an experimental group (E-group) with the mean of a control group (C-group) the Effect Size has the appealingly simple value:

$$\frac{[(\text{E-group mean}) - (\text{C-group mean})]}{[\text{C-group SD}]}$$

Essentially the effect size is indicated by locating the E-group mean as a *z*-score in the control group distribution (see Fig. 2). The Effect Size is the best estimate of the likely magnitude of the effect in a series of replications. With such information there is some chance that the magnitude of this effect can be interpreted for its educational significance.

[Insert Figure 2]

Unhappily the intuitively reasonable estimation for population values is biased (Hedges, 1981). The smaller the sample the more the ratio given above over-estimates the Effect Size. However, to correct for this the above values needs only to be multiplied by a correction factor and a table of these correction factors is supplied in the Glass *et al.* book (1979, p. 113, table 5.4). All Effect Sizes should be corrected before further analyses are carried out.

Another important complication is the question of any possible restriction in range in the control group, that is, the question of the size of the control group standard deviation (SD) (McGaw, 1982; McGaw & Glass, 1980). The control group might have a small SD, i.e. a restricted range of scores on the outcome measure, because, for example, they were a selected remedial group not showing the normal spread of scores on a cognitive test. Since the control group SD is the denominator for each Effect Size this alone would make the Effect Size appear large in comparison with Effect Sizes based on control groups with the normal spread of scores. For various experiments to be compared on the basis of control group distributions these distributions need to represent unrestricted sample or else a correction must be applied for restrictions in range (McGaw, 1982; McGaw & Glass, 1980).

Corrected for bias and also for restrictions in range, the Effect Sizes for the four replications in Table I were 0.58, 0.78, 1.87 and 0.28. This gave an average of 0.88. Considering a normal distribution, the average E-group participant scored higher than approximately 81% of the C-group participants. (Readers may recall that 34% of cases in a normal distribution lie between the mean and one standard deviation above the mean. Thus a *z*-score of 0.88 represents the 81st percentile.)

This small-scale example illustrates a type of meta-analysis – the synthesis of results via effect sizes – which would seem to be irreproachable. In this situation there were replications of the same treatment, using the same measure of outcome. Aggregation in some form was justifiable. The effect sizes were commensurate.

Following the computation of effect size one might well speculate on what it was about each experiment which caused the results to differ. That is, one starts to consider 'context variables'. A context variable is a variable which is constant for a given experiment and, therefore, its effects can only be determined from a consideration of many experiments across which the contexts change. For example, Hartley (1977) coded each study of individualised instruction in Mathematics according to whether or not the person running the project was employed by the school authority or was an outside researcher. This variable, affiliation of the researcher, was a constant for any one study, so its influence on any one project could only be surmised or studied qualitatively. Examining Effect Sizes across all studies Hartley found larger Effect Sizes associated with the researcher's being an employee of the school system rather than an external researcher. In general, Effect Sizes from the many studies of meta-analysis can be related to the many context variables to see if any of them appear to 'explain' some of the variations in Effect Sizes. Thus meta-analysis provides some empirical evidence as to whether or not a context variable (CV) is likely to be important, i.e. some evidence as to whether or not a CV is one of the conditions which must be taken into account in generalising results.

Since generalisation, the question of the external validity of results, has always been a thorny problem, the capacity of meta-analysis to provide tentative guidelines for generalisation is one of its most valuable features.

Given only a few experiments, as in Table I, speculations about variables related to Effect Sizes would not need the aid of statistics. But in examining effect sizes from large numbers of experiments one would need the aid of statistical summaries, such as those provided by correlations or ANOVAs. It must be emphasised however that the procedure of relating context variables to Effect Sizes represents hypothesis *generating* activity, not hypothesis testing. Hartley's finding of the relationship of Effect Size to affiliation of the researcher, for example, does not establish any causal relationship or explain the relationship. The procedure of relating Effect Sizes to CVs belongs to the kind of research activity variously called exploratory data analysis (Tukey, 1977), the search for grounded theory (Glaser & Strauss, 1967) correlational studies (Cronbach, 1957), passive observational studies (Cook & Campbell, 1979) or the stage of romance (Alfred North Whitehead as quoted in Dockrell, 1980). These terms may strike the reader as referring to very disparate activities but they all share one essential feature: the examination of data for relationships which suggest hypotheses. They are distinguished from the activity of *testing* hypotheses for which one needs, ideally, controlled experiments. Thus to test hypotheses about the context effects one would need controlled experiments in which context variables were deliberately manipulated.

Moving, perhaps, a little further from the irreproachable end of the continuum, we might consider Cohen *et al.* (1982) who aggregated various measures not just on one achievement dimension but on two – mathematics and reading. Using some 65 independent evaluations of school tutoring programmes, Cohen *et al.* computed Effect Sizes and related these to such study characteristics as adequacy of the research design, whether or not the tutoring task was highly structured, the duration of the treatment, cross-age or same-age, and topic being taught (mathematics or reading). Here the Effect Sizes begin to be less obviously commensurate. We could consider basic cognitive skills but *are* they justifiably taken as representing a single dimension? Cohen *et al.* found effect sizes, for tutor achievement, of 0.62 for mathematics and 0.21 for reading. For tutees the Effect Sizes were also substantially different: 0.60 for mathematics but only 0.29 for reading. Since the Effect Sizes were of quite different magnitudes, many other relationships with context variables would depend heavily on the proportion of the studies which were in the area of reading as opposed to mathematics. In addition, it can be argued that the guidance to further practice is lost if the two dependent variables are not kept distinct: factors associated with success in mathematics projects might not be important in reading projects.

The amalgamation of the mathematics and reading outcome measures may militate against the chance to evaluate projects fairly. For example, it may be much more difficult to change scores on a reading test than on a mathematics achievement test, particularly with older pupils for whom reading tests are essentially measures of comprehension, a skill which might not be sensitive to instructional interventions. The important general point for the technique of meta-analysis is that, despite the attractions of examining large bodies of data, researchers should stop short of commensurations which lose the meaning of the metric. *The ultimate aim of statistical analysis is interpretation and this requires an interpretable metric.*

Let us consider now, a type of meta-analysis which moves even further along the continuum towards dubious, i.e. which is open to more criticism as to whether or not it is appropriate. Schlessinger, Mumford & Glass (1978) coded 12 studies, producing 22 Effect Sizes indicating the effects of various kinds of treatments on asthma. In these studies, the dependent variables (DVs) were a variety of indicators of improvement, such as, for example, reduced wheezing or reduced hospitalisation. It could be argued that these two metrics, wheezing and hospitalisation, are *too* different to be considered at all commensurate. An improvement in wheezing score might be fine, but does not seem to match in importance or financial implications an improvement in hospitalisation rates. Similarly, Gallo (1978) commenting on a meta-analysis by Smith & Glass (1980) on the outcomes of psychotherapy wrote:

> Smith and Glass have aggregated dozens of different dependent measures together, ranging from elaborate clinical judgments to scores on paper-and-pencil tests. Any attempt to extricate meaningful information from such a hodgepodge is impossible. (Gallo, 1978, p. 516)

In summary, when scales for the outcome measures lose their meaningfulness upon aggregation, the meta-analysis might lie in the dubious to unacceptable end of the continuum. Different researchers will differ in their opinions as to how far from strict replication a study must be before it cannot be placed into a meta-analysis and in the extent to which different measures can be aggregated without loss to their interpretability. Some resolution of these differences of opinion must probably await more experience with the techniques of meta-analysis.

Before summarising this brief introduction to meta-analysis, mention might be made of a particular genre of criticism represented by two articles: Eysenck (1978) and Horan (1982). These both make good reading. Essentially their point is that certain fields of research are in too poorly defined a state to be worth meta-analysing. They seem to be arguing that in psychotherapy and counselling there is little agreement as to what constitutes adequate operationalisation for a treatment of a particular type (e.g. what is and what is not 'reality therapy'?) and little agreement as to what constitutes an acceptable measure of outcomes. These authors work in very difficult areas dealing with emotional and cultural variables (phobias, depressions, criminal behaviour) and one certainly would not wish to criticise them for their piercing glances at their own murky fields. But are they correct in their assumption that the mists of confusion are as dense elsewhere? Certainly, in the development of a research area, the state of meta-analysis can only *follow* the stage in which there are a fair number of well conceptualised and internally valid experiments. Eysenck's and Horan's criticisms do not constitute grounds for rejecting meta-analysis in fields of inquiry which have reached such a level of development.

Summary: the steps in meta-analysis

Meta-analysis has been presented here as having one main essential feature: the use of ESs (Effect Sizes) as a means of synthesising research findings. In the following paragraphs the four steps in a meta-analysis are enumerated and at the same time some more of the objections which have been made are described, since they help to draw attention to key features of the procedure. A meta-analysis proceeds by the following steps:

(1) *Finding studies: dissertations, theses, published work, and previous reviews are searched to draw up an exhaustive list of studies dealing with the effect of interest.* It has disturbed some researchers (e.g. Eysenck, 1978) that all studies are included for which Effect Sizes can be computed. Glass *et al.* specifically reject the procedure of selecting only 'well-designed' studies for analysis. Too often, they fear, this permits "the ad hoc impeaching on

methodological grounds of the studies of one's enemies" (p. 220). Nor will they accept that only published studies should be analysed. Those wishing to argue for such arbitrary ways of reducing the number of studies to be considered must present evidence which justifies such "rampant a priorism" (p. 226).

(2) *Coding study characteristics: Each study is read and its characteristics (e.g. date, publication status, details of design, ratings of design quality, status of the researcher, and any other variables which might influence the effects) are recorded on a coding sheet, with checks for inter-coder reliability.* Design quality, in various aspects, is always among the context variables or study characteristics which are coded. The fact that all studies are included does not mean an indifference to design quality but, rather, a willingness to examine empirically whether well-controlled studies do in fact yield different Effect Sizes from less well-controlled studies, information arising from the next two steps.

(3) *Measuring effect sizes: essentially the effect size is indicated by locating the E-group mean as a z-score in the control group distribution.* In order to compare outcomes from a variety of experiments some common scale is required. Glass *et al.* adopted the effect size as an appealingly simple measure for comparing the mean of an experimental group (E-group) with the mean of a control group (C-group). Since the differences arising are scaled by the control group deviations, these SDs must all be from either unrestricted or comparably-restricted samples. Thus corrections may be needed for restrictions in range. Other corrections are needed for small sample sizes.

One other problem in the step of estimating Effect Sizes is that those from the same study may not be independent, a problem of 'lumpy data'. A simple solution is to average all ESs from one study and use the study as the unit of analysis. No general rule can be given as the procedures adopted must reflect the extent to which the studies themselves contained independent replications. As so often, the statistics will be easier the better the initial designs.

(4) *Correlating effect sizes with context variables: the data are explored to see if there are any relationships between the Effect Sizes and the contexts in which the effects were found.* For example, it is at this step that the question of how ESs relate to design quality can be taken up. Meta-analyses so far have shown that in some research areas the Effect Sizes for well-controlled versus poorly controlled studies are very similar, while in other areas there must be systematic biases in the poorly controlled studies since they yield a different average Effect Size. In research on the effects of class size on achievement, for example, surveys (poorly controlled studies) often show lower achievement associated with smaller class sizes, probably because of school policies for keeping difficult classes small. The ESs from such studies are different from ESs derived from studies in which class size was a manipulated variable, i.e. from well-designed experiments. Given this situation, confidence would only be placed in the Effect Sizes computed from the well-controlled studies.

More generally, from meta-analyses of 12 research areas Glass *et al.* (1981, p. 226) found "seldom much more than .1 standard deviation difference between average effects for high validity and low validity experiments".

A major virtue of synthesising in terms of an ES is that this facilitates the interpretation of the findings, in particular it facilitates judgment of the *substantive*as opposed to *statistical* significance of the findings. Meta-analysis may move us into an area of parameter estimation rather than significance testing (cf. Simon, 1974). The essential point is that for all kinds of outcome variables, affective, cognitive, behavioural, etc. we must know the magnitude of the effect on an interpretable metric.

Correspondence: C.T. Fitz-Gibbon, Curriculum, Evaluation and Management Centre, Mountjoy Research Centre 4, University of Durham, Durham DH1 3UZ, England.

NOTE

[1] This is the first of two articles on meta-analysis. Meta-analysis is introduced and described here while the implications hinted at in the abstract will be considered in the next article.

REFERENCES

Athappilly, K.K. (1980) *A meta-analysis of the effects of modern mathematics in comparison with traditional mathematics in the American educational system* (Ann Arbor, Michigan, University Microfilms International).

Campbell, D.T. & Stanley, J.C. (1966) *Experimental and quasi-experimental designs for research* (Chicago, Rand McNally).

Carver, R.P. (1978) The case against statistical significance testing, *Harvard Educational Review*, 48(3), pp. 378-399.

Cohen, P.A., Kulik, J.A. & Kulik, C.C. (1982) Educational outcomes of tutoring: a meta-analysis of findings, *American Educational Research Journal*, 19(2), pp. 237-248.

Cook, T.D. & Campbell, D.T. (1979) *Quasi-experimentation: designs for field research* (Chicago, Rand McNally).

Cooper, H.M. (1982) Scientific guidelines for conducting integrative research reviews, *Review of Educational Research*, 52(2), pp. 291-302.

Cronbach, Lee J. (1957) Beyond the two disciplines of scientific Psychology, *American Psychologist*, 12, pp. 671-684.

Dockrell, W.B. (1980) The contribution of research to knowledge and practice: truth – what is that? in: Dockrell, W.B. & Hamilton, D. (Eds) *Re-thinking Educational Research*, pp. 11-22 (London, Hodder & Stoughton).

Eysenck, H. (1978) An exercise in mega-silliness, *American Psychologist*, p. 157.

Fitz-Gibbon, C.T. & Morris, L.M. (1975) Theory based evaluation, *Evaluation Comment*, 5(1), pp. 1-4.

Gallo, P. (1978) Meta-analysis – a mixed meta-phor?, *American Psychologist*, 33 (May), pp. 515-517.

Glaser, B.G. & Strauss, A.L. (1967) *The Discovery of Grounded Theory: strategies for qualitative research* (New York, Aldine).

Glass, G.V., McGaw, B. & Smith, M.L. (1981) *Meta-analysis in Social Research* (London, Sage).

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

Hargreaves, D.H. (1981) Schooling for delinquency, in: Barton, L. & Walker, S. (Eds) *Schools, Teachers and Teaching* (London, Falmer).

Hartley, S.S. (1977) Meta-analysis of the effects of individually paced instruction in mathematics, *doctoral dissertation*, University of Colorado.

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

Hedges, L.V. & Olkin, I. (1980) Vote-counting methods in research synthesis, *Psychological Bulletin*, 88(2), pp. 359-369.

Horan, J.J. (1982) Experimentation in counselling and psychotherapy. part I: new myths about old realities, *Educational Researcher*, 9(11), pp. 5-10.

Jones, L.V. & Fiske, D.W. (1953) Models for testing the significance of combined results, *Psychological Bulletin*, 50, pp. 375-382.

Krol, R.A. (1979) *A meta analysis of comparative research on the effects of desegregation on academic achievement* (Ann Arbor, Michigan, University Microfilm International).

Kulik, C.C. & Kulik, J.A. (1982) Effects of ability grouping on secondary school students: a meta-analysis of evaluation findings, *American Educational Research Journal*, 19(3), pp. 415-428.

Light, R.J. & Smith, P.V. (1981) Accumulating evidence and procedures for resolving contradictions among different research studies, *Harvard Educational Review*, 41, pp. 429-471.

McGaw, B. (1982) Ensuring a common metric in meta-analysis, *paper presented at annual meeting of American Educational Research Association*, New York.

McGaw, B. & Glass, G.V. (1980) Choice of metric for effect size, *American Educational Research Journal*, 17(3), pp. 325-338.

Redfield, D.L. & Rousseau, E.W. (1981) A meta-analysis of experimental research on teacher questioning behaviour, *Review of Educational Research*, 51(2), pp. 237-245.

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

Schlessinger, H.J., Mumford, E. & Glass, G.V. (1978) *A critical review and indexed bibliography of the literature up to 1978 on the effects of psychotherapy on medical utilization* (Denver, Department of Psychiatry, University of Colorado Medical Centre).

Simon, H.A. (1974) How big is a chunk?, *Science*, 183, pp. 482-487.

Smith, M.L. & Glass, G.V. (1980) Meta-analysis of research on class size and its relationship to attitudes and instruction, *American Educational Research Journal*, 17(4), pp. 419-434.

Tukey, J.W. (1977) *Exploratory Data Analysis* (London, Addison Wesley).

Winer, B.J. (1971) *Statistical Principles in Experimental Design*, 2nd Edn (New York, McGraw-Hill).