



The Relevance of Systematic Reviews to Educational Policy and Practice

PHILIP DAVIES

ABSTRACT *This paper argues that educational policy and practice has much to gain from systematic reviews and other methods of research synthesis. Different types of reviews are considered, including narrative reviews, vote-counting reviews, meta-analyses, best evidence synthesis, and meta-ethnography. It is argued that systematic reviews allow researchers, and users of research, to go beyond the limitations of single studies and to discover the consistencies and variability in seemingly similar studies. This, in turn, allows for some degree of cumulative knowledge of educational research that is often missing in the absence of systematic reviews. Some limitations of systematic reviews and research synthesis for educational policy and practice are also discussed. The work of the Campbell Collaboration as an international organisation that promotes the use of systematic reviews in educational policy and practice is outlined.*

INTRODUCTION

Recent criticisms of educational research and policy have suggested that there is a gap between the knowledge and research needs of those who provide education and those who undertake research on educational policy and practice (Hargreaves, 1996, 1997; Hillage *et al.*, 1998; Tooley & Darby, 1998). Educational research has been criticised for serving the interests of researchers, rather than those of policy-makers, providers and users of educational services. Consequently, the relevance, applicability and quality of educational research has come under the close scrutiny of those who question the value of educational research. Yet another criticism has been that educational research fails the policy-making and broader educational community by the non-cumulative nature of its findings.

These criticisms are not new. Hunter and Schmidt (1990) provide the following statement from Senator Walter Mondale's invited address to the American Psychological Association Convention in 1970:

What I have not learned is what we should do about these problems. I had hoped to find research to support or to conclusively oppose my belief that quality integrated education is the most promising approach. But I have found very little conclusive evidence. For every study, statistical or theoretical, that contains a proposed solution or recommendation, there is always another, equally well documented, challenging the assumption or conclusions of the first. No one seems to agree with anyone else's approach. But more distressing: no one seem to know what works. As a result I must confess, I stand with

my colleagues confused and often disheartened. (quoted in Hunter and Schmidt, 1990, p. 35)

No doubt some academic researchers will take the view that such policy issues, let alone their resolution, is none of their business. Others may argue that the task of academic researchers is not to seek common agreement on such matters, but is primarily one of providing competing accounts and evidence of whatever it is they are studying. Post-modernist researchers will undoubtedly argue that there is no such thing as 'conclusive evidence' about anything, and that confusion and uncertainty is the very stuff of good academic research.

For those who think that it is a legitimate task of academic research to address issues such as the effectiveness of educational policy and practice, and who seek the best cumulative evidence of 'what works' and what does not (and, indeed, what *constitutes* 'working' and 'not working' in education), research synthesis provides a valuable means to this end.

Research synthesis, of which systematic reviews are one type, is built upon the observation that single studies 'are limited in the generalisability of the knowledge they produce about concepts, populations, settings and times' and 'frequently illuminate only one part of a larger explanatory puzzle' (Cook *et al.*, 1992, p. 3). Single studies, even if they are randomised controlled trials or other types of experimental inquiry, have limitations of time-, sample- and context-specificity which can undermine their applicability, relevance and usefulness in other contexts. Research synthesis, or research integration, 'involves the attempt to discover the consistencies and account for the variability in similar-appearing studies' (Cooper & Hedges, 1994, p. 4). In turn, this implies that 'seeking generalisations also involves seeking the limits and modifiers of generalisations' and, thereby, identifying the contextual-specificity of available research and evidence.

There is currently much interest in the U.K. and other countries in research synthesis and systematic reviews of available evidence. In the U.K. the Department for Education and Employment (DfEE) has established a Centre for Evidence-Informed Policy and Practice in Education at the Social Science Research Unit of the Institute of Education at London University. A broader based Co-ordinating Centre for Evidence-Based Policy and Practice is also being established by the UK Economic and Social Research Council (ESRC). More generally within the UK Government, research synthesis and evidence-based practice have become guiding principles of policy development and implementation (Cabinet Office, 1999). The Centre for Management and Policy Studies (CMPS) within the Cabinet Office has been established to provide Ministers and civil servants with evidence of best practice based on systematic and critical appraisal of research evidence from the social and political sciences. Organisations such as the National Health Service Centre for Reviews and Dissemination (CRD) at the University of York, and the National Institute of Clinical Excellence (NICE), complement the work of the Cochrane Collaboration. The latter is an international organisation that prepares, maintains and disseminates systematic reviews of the effects of interventions in health care, to provide a high quality database of best evidence for policy and practice in health care. The Campbell Collaboration has recently been established as a sibling organisation to the Cochrane Collaboration to help people make well-informed decisions about education, criminal justice, and social work and welfare by putting the best available evidence from systematic reviews of research at the heart of policy and practice in these areas of public service. Given this

degree of public interest in research synthesis and systematic reviews, it is important to establish what these terms cover, and their relevance to educational policy and practice.

TYPES OF RESEARCH SYNTHESIS

Narrative Reviews

Research synthesis is the collective term for the family of methods for summarising, integrating and, where possible, cumulating the findings of different studies on a topic or research question. The simplest form of research synthesis is the traditional qualitative literature review, often referred to as the *narrative review*. In its simplest form, the narrative review attempts to identify what has been written on a subject or topic, using which methodologies, on what samples or populations, and with what findings. Often, there is no attempt to seek generalisation or cumulative knowledge from what is reviewed. Rather, the task is to identify the range and diversity of the available literature, much of which will be inconclusive, and to find a gap which new research might attempt to fill. Such literature reviews are almost always *selective*, if not haphazard, in that they do not involve a systematic, rigorous and exhaustive search of *all* the relevant literature, using electronic and print media as well as hand searching and ways of finding the 'grey' literature. Instead, narrative literature reviews are often *opportunistic* in that they review only that literature and evidence that is readily available to the researcher (the file drawer phenomenon). Such narrative reviews may also involve discarding studies that use methodologies in which the researcher has little or no interest.

Vote Counting Reviews

A more systematic and sophisticated type of research synthesis is the *voting method*. This attempts to accumulate the results of a collection of relevant studies by counting 'how many results are statistically significant in one direction, how many are neutral (i.e. 'no effect'), and how many are statistically significant in the other direction' (Cook *et al.*, 1992, p. 4). The category that has the most counts, or votes, is taken to represent the modal or typical finding, thereby indicating the most effective means of intervention. As Cook *et al.* point out, however, this approach fails to indicate 'the possibility that a treatment might have different consequences under different conditions'. This is a point that was repeatedly acknowledged by Donald T. Campbell (after whom the Campbell Collaboration is named). Commenting on Campbell's work (Campbell, 1957; Campbell & Stanley, 1966), and the work of Cronbach (1982), Cook *et al.* point out that for these authors:

person and setting factors are especially likely to moderate causal relationships and help explain why a treatment has the effects it does. (Campbell adds time to this list). Both authors assume that social affairs are multiply determined in complex ways and that diversity typically found among people, settings, and historical climates creates unique context for each study. (Cook *et al.* 1992, p. 22)

Another problem with the voting method is that it takes no notice of the fact that some studies are superior to others methodologically and, consequently, are more valuable and deserve special weighting (see Cook *et al.*, 1992) The method of research synthesis

that does acknowledge the differential quality and value of seemingly similar studies is *meta-analysis*.

Meta-Analysis

The term 'meta-analysis' is usually attributed to Gene Glass (1976) who used the term to refer to 'the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings'. Meta-analysis 'combines the individual study treatment effects into a 'pooled' treatment effect for all studies combined, and/or for specific subgroups of studies or patients, and makes statistical inferences' (Morton, 1999). In the two decades or more since Glass's original meta-analytic work on psychotherapy (Smith *et al.*, 1980) and class size (Smith & Glass, 1980; Glass *et al.*, 1982), meta-analysis has developed considerably in terms of the range and sophistication of data-pooling and statistical analysis of independent studies (see Kulik & Kulik, 1989, Cook *et al.*, 1992, and Cooper & Hedges, 1994, for more detailed accounts of these developments).

Meta-analysis, then, allows researchers to go beyond the limitations of single studies and establish what are the consistent findings about intervention effects across studies, based on large, aggregate samples. Meta-analysis, however, has its own limitations. Cooper and Hedges (1994) point out that where meta-analyses include primary studies that were not based on the random assignment of study participants to experimental and control groups, 'causal inferences are not permissible by either the primary researcher or the synthesiser', and one is left with only correlational evidence (which may, of course, have its own merits and insights for researchers and users of research).

Meta-analysis, in keeping with other forms of research synthesis, is also limited by the specificity of the question it is addressing, and the comprehensiveness of the literature and data searches upon which it is based. Good meta-analysis requires good and specific questions about the intervention being undertaken, the population (or sample) that is the focus of the study, and the outcomes that are being assessed. Given that each of these variables may be different across studies, this presents a challenge for the meta-analyst to ensure that there is real consistency between primary studies on all three dimensions. The heterogeneity of samples, research questions asked, and outcomes measured raises the 'apples and oranges' question of comparability of different primary studies (Slavin, 1984) and, as Preiss (1988) has pointed out, 'questions regarding judgement calls made during meta-analysis'. Random-effects models are used where meta-analysts believe that the primary studies are 'different from one another in ways too complex to capture by a few simple study characteristics' (Cooper & Hedges, 1994, p. 526). Where there is greater homogeneity between primary studies, fixed-effects models of meta-analysis may be used.

The limitations of meta-analysis are also dependent on how systematic and comprehensive the meta-analyst is in searching for relevant and appropriate primary studies. This requires extensive, if not exhaustive, searches of databases, textbooks, journals, conference proceedings, dissertation abstracts, and research-in-progress, using electronic and hand searching methods. The need to search unpublished sources (including research-in-progress) is crucial given the problems of publication bias which may favour positive outcome studies in some journals and negative outcome studies in others.

The independence, data quality and adequacy of statistical reporting in primary studies are also potentially limiting factors of meta-analysis. The independence of primary studies refers to the possibility that one study may be reported more than once

in the research literature, and that different subgroups from a single study may be reported on in different papers. This can lead to the results of individual studies being included more than once in a meta-analysis, and a problem of double counting effect size. Good meta-analysts will ensure that this is not done, but the reader should be aware of this potential problem.

Missing data also confound meta-analyses. The failure to report follow-up data in a way that is consistent with baseline data is one problem. Another problem occurs where studies do not report, or account for, participants lost-to-follow-up at different data collection points. The problem may be very significant where there is a high lost-to-follow-up rate, and reporting of this is minimal, of those participants who drop out of the primary study. These may be the very people in whom the study (or meta-analysis) is interested, such as non-attenders in school truancy studies, problem drinkers in studies of alcoholism treatment, or recidivists in criminological intervention studies. Missing data on moderator and mediating variables also present problems and limitations for meta-analysis (Cooper & Hedges, 1994).

The adequacy of statistical reporting in primary studies is also variable. Different studies use different descriptive and inferential statistics. Some use only means and standard deviations; others use chi-squares, odds ratio and confidence intervals. Measures of dispersion are sometimes not included in research reports with measures of central tendency, which limits the statistical and practical usefulness of the findings. The use of different types of statistical manipulation, such as logarithmic transformations of data, is also very variable and may not be reported in any great detail. The validity and reliability of tests and outcome measures may also be variable and not reported. As Cook *et al.* (1992) point out, there are several ways in which problems of inadequate statistical reporting can be handled by meta-analysts. These include the use of external sources to establish the validity and reliability of instruments used in primary studies, contacting the primary investigator(s) to obtain additional data or clarification of procedures used, and reporting deficiencies of primary data in the meta-analysis, thereby distinguishing between good and poor data.

Best Evidence Synthesis

Slavin (1984, 1986) has criticised meta-analysis for not always being selective enough in terms of the methodological quality of studies that are included in reviews. Slavin characterises meta-analysis as using 'exhaustive inclusion followed by statistical tests' and of 'including all studies that meet broad standards in terms of independent and dependent variables, avoiding any judgement of quality' (Slavin, 1986, p. 6). This is done, says Slavin, in order to avoid the reviewer's own subjective biases entering decisions about which studies are 'good' and which are 'bad'. In contrast to this Slavin suggests 'best evidence synthesis', whereby 'reviewers apply consistent, well justified, and clearly stated *a priori* inclusion criteria' of studies to be reviewed. Slavin suggests some guiding principles for choosing *a priori* criteria, including that primary studies should be germane to the issue at hand, should be based on a study design that minimises bias, and should have external validity. By germane, Slavin means that:

a meta-analysis focusing on school achievement as a dependent measure must explicitly describe what is meant by school achievement and must only include studies that measured what is commonly understood by school achievement on individual assessments, not swimming, tennis, block stacking, time-on-

task, task completion rate, group productivity, attitudes, or other measures perhaps related to but not identical with student academic achievement. (Slavin, 1986, p. 6)

By study design that minimises bias Slavin acknowledges that where ‘the independent variable is strongly correlated with academic ability, motivation, and many other factors that go into a decision to, for example, promote or retain a student ... random assignment to experimental or control groups is essential’ (Slavin, 1986, p. 7) In other contexts, however, where the independent variable is less correlated with dependent variables, says Slavin, ‘then random assignment, though still desirable, may be less essential’.

By external validity, Slavin calls for outcome variables that have some ‘real life’ educational significance rather than ‘extremely brief laboratory studies or other highly artificial experiments’ (1986, p. 7). This underlines Slavin’s concern about the use of diverse measures of educational activities and outcomes, many of which are only remotely related to what is commonly understood by school achievement. This has considerable implications for the relevance of systematic reviews for educational policy and practice, which will be discussed later in this paper.

Slavin’s ‘best evidence syntheses’ have been criticised by meta-analysts such as Kulik and Kulik (1989) on the grounds that they ‘usually cover relatively few studies’, and that they involve ‘analyst biases’ (Kulik and Kulik, 1989, p. 255). This tension between the statistical benefits of exhaustive inclusion and a large number of primary studies on the one hand, and high quality reviews of fewer studies using more selective methodological criteria of inclusion and exclusion, is a recurrent theme in systematic review methodology. In the decade and a half since Slavin’s proposal for best evidence synthesis, many of the quality criteria that he called for—explicit *a priori* criteria, exhaustive literature searches of published and unpublished studies, listings of included and excluded studies with the reasons for doing so, and the study characteristics, transparency of reviewers’ procedures and conclusions—have become common practice in systematic reviews. They are central requirements of the systematic reviews that appear in the Cochrane Library of health care interventions, and will undoubtedly play a similar role in the developing Campbell Library of interventions in education, criminal justice and social work.

More recently Slavin and Fashola (1998) have presented a review of ‘proven and promising programs for America’s schools’, which uses a rather pragmatic notion of best evidence synthesis. They noted that:

Ideally, programs emphasised in this book would be those that present rigorous evaluation evidence in comparison and control groups showing significant and lasting impacts on the achievement of students placed at risk, have active dissemination programs that have implemented the program in many schools serving at-risk students, and have evidence of effectiveness in dissemination sites, ideally from studies conducted by third parties. To require all of these conditions would limit this review to very few programs. To include a much broader range of programs, we had to compromise on one or more criteria. (Slavin & Fashola, 1998, p. 10)

Some studies are included in this review even though Slavin and Fashola had reservations about some aspects of the primary studies in question. They note, for instance, that the comparison groups used in Mehan *et al.*’s (1996) AVID project may be susceptible to bias, yet they conclude that ‘the college enrollment rates for AVID are

impressive, and the program has a good track record in serving students throughout the United States, and for these reasons is worthy of consideration by other schools serving many students placed at risk' (Slavin & Fashola, 1998, p. 87). The inclusion of an excellent study such as the AVID project in a major (though not necessarily systematic) review of programmes for American Schools may not meet with the approval of some meta-analysts. This study, however, provides valuable evidence not only of what seems to work in terms of promoting the college enrollment of low achieving students with good academic potential, but also good qualitative evidence from case studies, interviews with students and teachers, and ethnographic research, of *why* and *how* the AVID programme succeeds, and has limitations. The synthesis of good qualitative research is less developed than that of controlled experiments, but is attracting considerable attention from researchers interested in evidence-based policy and practice in education and other areas of public services.

Meta-ethnography

Meta-ethnography attempts to summarise and synthesise the findings of qualitative studies, especially ethnographies and interpretive studies. It is embedded in the interpretive paradigm of social scientific research and claims to 'be interpretive rather than aggregative' (Noblit & Hare, 1988, p. 11). Noblit and Hare define the interpretive paradigm as:

research that is termed ethnographic, interactive, qualitative, naturalistic, hermeneutic, or phenomenological. All these types of research are interpretive in that they seek an explanation for social or cultural events based upon the perspectives and experiences of the people being studied. In this way, all interpretive research is 'grounded' in the everyday lives of people. (1988, p. 12)

Like meta-analysis, meta-ethnography 'seeks to go beyond single accounts' (Noblit & Hare, 1988, p. 13), but instead of doing so by aggregating samples and identifying consistencies and variability between different studies, it does this by 'constructing interpretations, not analyses' and by revealing 'the analogies between the accounts' (*ibid*). Meta-ethnography, say Noblit and Hare, 'reduces the accounts while preserving the sense of the account through the selection of key metaphors and organisers' (*ibid*). In an attempt to clarify this, Noblit and Hare suggest that 'when we talk about the key metaphors of a study, we are referring to what others may call themes, perspectives, organisers, and/or concepts revealed by qualitative studies' (1988, p. 14). To this extent, meta-ethnography would appear to have more in common with narrative reviews than with vote counting systematic reviews, meta-analyses or best evidence synthesis.

Meta-ethnography has some of the same problems as meta-analysis and other types of research synthesis, such as establishing criteria for which studies to include and exclude in a meta-ethnographic review. This is possibly even more difficult with qualitative research in that there seems to be even greater diversity than with quantitative studies in terms of the questions being asked and the theoretical perspectives from which these questions are generated. In other words, the heterogeneity of primary studies may be greater with qualitative research, and the potential for homogeneity may be less than is the case with quantitative primary studies.

From an interpretive perspective, meta-ethnography also has a problem of balancing

summary statements of qualitative studies with their contextual specificity. Notwithstanding the point made by Noblit and Hare that meta-ethnography is 'interpretive rather than aggregative', there have been attempts by qualitative researchers and meta-ethnographers to 'venture towards achieving more general conclusions from the ethnographic specifics of the separate cases' (Wax, 1979, p. 1). These attempts at aggregation, however, have 'avoided a full exploration of context and did not enable an explanatory synthesis' (Noblit & Hare, 1988, p. 21). Instead, the authors of these attempts at aggregation of qualitative studies complain that they ignore the 'meaning in context' and the 'ethnographic uniqueness' that is so central to ethnographic and qualitative inquiry.

From the more positivistic perspective of meta-analysis, meta-ethnography is seen as being limited by its inability to provide statistical accumulation of findings, its inability to allow prediction or to specify any degree of confidence about qualitative findings, and by its inability to allow for the statistical control of bias. The apparent lack of any systematic way for meta-ethnography to test for, and control, the heterogeneity/homogeneity of different studies, also concerns meta-analysts and those more disposed to quantitative approaches to research synthesis. These concerns and limitations, however, are somewhat cross-paradigmatic and seem to miss the point of what ethnographies and other qualitative studies are trying to achieve (Davies, 2000).

SYSTEMATIC REVIEWS AND EDUCATION

Meta-analysis and other types of systematic review have a distinguished record in medicine and health care (Smith *et al.*, 1980; Crowley *et al.*, 1990; Antman *et al.*, 1992), criminal justice (Farrington, 1983; Dennis, 1988; Weisburd, 1993; Petrosino, 1997), and social work (Didden *et al.*, 1997; Hoag & Burlingame, 1997; Kavale *et al.*, 1997; Reeker *et al.*, 1997; Gorey *et al.*, 1998; Guterman, 1999). It is in educational research, however, that systematic reviews had their origins (Smith & Glass, 1980; Glass *et al.*, 1982) and have been developed extensively over the past two decades (Kulik & Kulik, 1989; Lipsey & Wilson, 1993). Kulik & Kulik reviewed some 150 meta-analyses of educational interventions, and Lipsey & Wilson reviewed 302 meta-analyses of psychological, educational and behavioural interventions, two-thirds of which were in education. These systematic reviews covered a wide range of substantive topics, subject areas, methods of teaching and learning, types of school and class organisation, different students and learners, stages of education, and measures of educational outcome and achievement (Davies *et al.*, 2000). It is perhaps no exaggeration to suggest that a whole industry of systematic reviews and meta-analyses has developed in educational research, particularly in the USA. The question to be asked is: what has been learned from all this industry of systematic reviews and meta-analyses in education?

Meta-analyses and systematic reviews typically measure the *effect size* of interventions in terms of the difference between the mean change in some outcomes measures in the experimental group and the mean change in the control group, divided by the standard deviation of either the control group or the combined standard deviations of the experimental and control groups. Using this summary statistic, meta-analysts such as Kulik & Kulik and Lipsey & Wilson have noted that most educational interventions have moderate effects, and that both large effects and negative effects occur infrequently. Lipsey & Wilson, for instance, found only six meta-analyses (out of a total of 302) that identified negative mean effect sizes. These were in the areas of open

classroom learning (N = 3 meta-analyses), between and within ability groupings of secondary students (N = 1), self-paced modularised individualised mathematics teaching (N = 1), and special education classroom placement of 'exceptional children' (N = 1).

So strong is the evidence from systematic reviews of a positive effect of most educational interventions that one must ask whether this is an artifact of the methodology of systematic reviews, or of the types of research designs used in studies of educational interventions. Lipsey & Wilson (1993) have addressed this issue by considering the possible inflation of effect size that derives from the type of research design (randomised versus non-randomised allocation of experimental and control groups), the methodological quality of primary studies, small sample bias, publication bias, and a generalised placebo effect. After reviewing the evidence on each of these factors Lipsey & Wilson concluded that two of them—one-group pre- and post-designs without a control or comparison group, and publication bias—inflate the mean effect size of educational interventions. Lipsey & Wilson then excluded all the meta-analyses in their corpus that were based on one-group pre- and post-designs, as well as those that did not include unpublished studies, and found that there was a very similar distribution of mean effect sizes as before, albeit on a smaller number of studies (N = 156). That is, most meta-analyses report moderate positive effect sizes (0.20 standard deviations higher than the mean outcome measures in the control groups) with only a few studies showing strong effect sizes and very few showing effect sizes in the zero or negative range.

If one accepts these findings and interpretations of Lipsey & Wilson's review, the question still remains as to the practical meaning and relevance of meta-analytic findings for teachers, learners and other members of the educational community. What, for instance, does it mean to users of research evidence to know that 'the average treatment [*sic*] group scores 0.47 standard deviations higher on the average outcome measure than did the average control group' (Lipsey & Wilson, 1993, p. 1198). The generality of such findings may mean very little to teachers, for instance, who have different groups of students, pursuing various types of educational activities, with various outcomes, all of which are anything but 'average'. Similarly, it may be hard for teachers and educational managers to grasp what it means to have students scoring '0.47 standard deviations higher on the average outcome measure than did the average control group'. It is often difficult, if not impossible, to convert such findings to the currency in which teachers, learners, parents and educational managers trade, such as (in the UK) SATs, GCSE Grades A–C, GCE 'A' Levels, degree classifications, and positions in league tables of educational performance.

A closely related issue is the relationship between *statistical* significance and *educational* significance. A class or school may indeed score x or y standard deviations higher than some control group norm, and this may be statistically significant. If, however, the criterion of educational achievement in the primary studies is the accuracy of throwing a tennis ball against a wall (cited by Kulik and Kulik of one of the studies in Glass and Smith's (1980) meta-analysis of class size), then this may have little or no relevance, or educational significance, to teachers, learners or educational managers charged with the demands of the National Curriculum of England and Wales (and similar educational initiatives in other countries).

Class size provides an example of how difficult it can be to grasp the practical relevance of systematic reviews and meta-analyses. There have been a number of systematic reviews of the effects of class size on students' achievement (Glass & Smith,

1979; Hedges & Stock, 1983), and on students' and teachers' attitudes and feelings (Smith & Glass, 1980). Glass and Smith found that in terms of academic achievement there is an average effect size 0.21 standard deviations higher in classes of less than 30 students than in classes with more than 30 students. Amongst less-controlled primary studies the average effect size dropped to 0.09 standard deviations. This suggests that class size plays a marginal role (one tenth to one fifth of a standard deviation) in determining student achievement. In terms of students' attitudes to learning, the effect size of smaller classes, compared to larger classes, was larger (0.53 standard deviations), suggesting that on this dimension class size may be more important. Hedges and Stock (1983) reanalysed Glass and Smith's data using different statistical methods, and found very little difference in these effect sizes.

These findings, however, give a *generalised* overview (that is the point of them) of the effects of class size on measures of educational achievement, some of which (such as swimming, tennis, block stacking and group productivity) have dubious external and/or ecological validity. The threshold criterion of small classes versus large classes in the Glass and Smith study was 30 students, leaving a wide range of 'small class size' from 1–29 (including one-to-one tutorials), and a wide range of 'large class size' from 30–40, or even 30–60. Moreover, Glass and Smith suggest that the effect sizes of class size are algorithmic, and that they have greater effect in classes of fewer than 20 students, or even as low as 10 students, than in classes of 20–40 students. This raises questions about the practical relevance of such findings in educational climates where classes of 20–40 students are normal and class sizes of fewer than 20 (and certainly fewer than 10) are exceptional. There are also questions about the relevance of the outcome measures of academic achievement used in some of the studies included in Glass and Smith's meta-analysis (e.g. throwing a tennis ball against a wall) and those that are used in other educational contexts (such as the SATs in the National Curriculum of England and Wales).

Teachers, learners, parents and educational managers tend to have *particular* and *context-specific* concerns about education, such as whether class size of more than 20 students in *their* primary school reception class has an effect on *their* children's/students' reading and cognitive abilities *at this point in time*, on their SAT scores at age seven, on their social and interactional development, on their abilities for independent as well as collaborative learning, and so forth. Systematic reviews and meta-analysis can only rarely give clear and unambiguous evidence about the best way to achieve these quite reasonable goals. This tension between the generalised and particular demands upon evidence may lessen the apparent relevance of systematic reviews and meta-analyses. This, however, does not render systematic reviews, or evidence-based education, redundant. Just as evidence-based health care means 'integrating individual clinical expertise with the best available external evidence from systematic research' (Sackett *et al.*, 1996), so evidence-based education means integrating individual teaching and learning expertise with the best available evidence from systematic research on educational interventions and practice. Similarly, just as evidence-based medicine is perhaps best expressed as context-sensitive medicine (Greenhalgh & Worrall, 1997), so evidence-based education should be conceived as context-sensitive education. This requires systematic reviews, and single study findings, of the best available evidence from ethnographic studies, sociological observational studies, and the whole range of qualitative studies in education and social scientific research (Davies, 2000). When taken together with the findings of meta-analyses, users of educational research will have a broad basis of sound empirical evidence with which to exercise their professional

judgement, and make context-sensitive decisions that are more fully informed and robust.

CONCLUSIONS

This paper has attempted to summarise different types of systematic reviews and to discuss their relevance to educational policy and practice. A number of problems of doing, and using, systematic reviews have been identified. Notwithstanding the limitations posed by these problems, systematic reviews have much to offer the educational community in terms of providing unbiased evidence from a wide range of studies of educational policy and practice. The recently formed Campbell Collaboration provides an international mechanism for preparing, maintaining and disseminating systematic reviews of the effects of educational interventions, policy and practice. The Education Group of the Campbell Collaboration is establishing review groups in educational methods, mathematics, science, early years learning, work-related learning and transferable skills, information technology in teaching and learning, leadership and management, medical education, and dissemination. Other topics for review groups are emerging and suggestions from readers who may wish to become involved in the Campbell Collaboration are welcome [1].

The quality of systematic reviews and meta-analyses depends upon the quality of the primary studies on which they are based, and on the rigour, transparency and reporting of the inclusion and exclusion criteria used by reviewers when doing a systematic review. This is essential if systematic reviews and meta-analyses are to deliver what they promise—high quality reviews of existing and emerging evidence in which bias is minimised. If the methodological quality standards reviewed in this paper are not adhered to, the words of caution of Iyengar (1991) are highly relevant:

Meta-analysis is like any other item in the scientist's toolkit. When it is applied thoughtfully and carefully to a body of data it can provide valuable insights. Otherwise, it merely adds to the confusion that it was intended to clear up.

NOTE

- [1] The Campbell Collaboration is co-ordinated in the UK by Philip Davies (philip.davies@conted.ox.ac.uk).

REFERENCES

- ANTMAN, E.M., LAU, J., KUPELNICK, B. *et al.* (1992) A comparison of results of meta-analysis of randomised controls trials and recommendations of clinical experts' treatments for myocardial infarction, *Journal of the American Medical Association*, 269, pp. 240–248.
- CABINET OFFICE (1999) *Modernising Government*, White Paper (London, Stationery Office) Cmnd. 4310.
- CAMPBELL, D.T. (1957) Factors relevant to the validity of experiments in social settings: a challenge to conventional interpretations, *Psychological Bulletin*, 54, pp. 297–312.
- CAMPBELL, D.T. & STANLEY, J.C. (1966) *Experimental and Quasi-experimental Designs for Research* (Chicago, Rand McNally).

- COOK, T.D., COOPER, H., CORDRAY, D.S., HARTMANN, H., LIGHT, R.J., LOUIS, T.A. & MOSTELLER, F. (1992) *Meta-Analysis for Explanation*, (New York, Russell Sage Foundation).
- COOPER, H. & HEDGES, L.V. (eds) (1994) *The Handbook of Research Synthesis* (New York, Russell Sage Foundation).
- CRONBACH, L.J. (1982) *Designing Evaluations of Educational and Social Programs* (San Francisco, Jossey-Bass).
- CROWLEY, P., CHALMERS, I.G. & KEIRSE, M.J.N.C. (1990) The effects of corticosteroid administration before preterm delivery: an overview of the evidence from controlled trials, *British Journal of Obstetrics and Gynaecology*, 97, pp. 11–25.
- DAVIES, P.T. (2000) Qualitative Research Methods in Evidence-Based Policy and Practice, in: H.T.O. DAVIES, S.M. NUTLEY & P.C. SMITH (eds), *What Works? Evidence and Public Policy* (Bristol, The Policy Press).
- DAVIES, P.T., HOLMES, E. & WOLF, F. (2000) An organisational framework for preparing and maintaining systematic reviews in education, Paper presented at the Inaugural Meeting of the Campbell Collaboration, Philadelphia, Pennsylvania, February 25–26, 2000.
- DENNIS, M.L. (1988) *Implementing Randomised Field Experiments: An Analysis of Criminal and Civil Justice Research*, PhD Dissertation, University Microforms, Northwestern University, Ann Arbor, Michigan.
- DIDDEN, R., DUKER, P.C. & KORZILIUS, H. (1997) Meta-analytic study on treatment effectiveness for problem behaviours with individuals who have mental retardation, *American Journal of Mental Retardation*, 101, 4, pp. 387–399.
- FARRINGTON, D. (1983) Randomised experiments on crime and justice, in: M. TONRY & N. MORRIS (eds), *Crime and Justice: An Annual Reviews of Research, Volume IV* (Chicago, University of Chicago Press).
- GLASS, G.V. (1976) Primary, secondary and meta-analysis of research, *Educational Researcher*, 5, pp. 3–8.
- GLASS, G.V. & SMITH, M.L. (1979) Meta-analysis of research on class size and achievement, *Educational Evaluation and Policy Analysis*, 1, 2–16.
- GLASS, G.V., CAHEN, L.S., SMITH, M.L. & FILBY, N.N. (1982) *School Class Size: Research and Policy* (Beverly Hills, Sage Publications).
- GOREY, K.M., THYER B.A. & PAWLICK, D.E. (1998) Differential effectiveness of prevalent social work practice models: a meta-analysis, *Social Work*, 43, 3, pp. 269–278.
- GREENHALGH, T. & WORRALL, J.G. (1997) From EBM to CSM: the evolution of context-sensitive medicine, *Journal of Evaluation in Clinical Practice*, 3, 2, pp. 105–108.
- GUTERMAN, N.B. (1999) Enrollment strategies in early home visitation to prevent physical child abuse and neglect and the ‘universal versus targeted’ debate: a meta-analysis of population-based and screening-based programs, *Child Abuse and Neglect*, 23, 9, pp. 863–90.
- HARGREAVES, D.H. (1996) *Teaching as a Research-Based Profession: Possibilities and Prospects* (Cambridge, Teacher Training Agency Annual Lecture).
- HARGREAVES, D.H. (1997) In defence of research for evidence-based teaching: a rejoinder to Martyn Hammersley, *British Educational Research Journal*, 23, 4, pp. 405–419.
- HEDGES, L.V. & STOCK, W. (1983) The effects of class size: an examination of rival hypotheses, *American Educational Research Journal*, 20, pp. 63–85.

- HILLAGE, J., PEARSON, R., ANDERSON, A. & TAMKIN, P. (1998) *Excellence in Research on Schools: Research Report RR74* (Sudbury, DfEE Publications).
- HOAG M.J. & BURLINGAME G.M. (1997) Evaluating the effectiveness of child and adolescent group treatment: a meta-analytic review, *Journal of Clinical Child Psychology*, 26, 3, pp 234–246.
- HUNTER, J.E. & SCHMIDT, F.L. (1990) *Methods of Meta-Analysis* (Newbury Park, Sage Publications).
- IYENGAR, S. (1991) Much ado about meta-analysis, *Chance*, 4, 1, pp 33–40.
- KAVALE, K.A., MATHUR, S.R., FORNESS, S.R., RUTHERFORD, R.B. JR, QUINN, M.M. (1997) Effectiveness of social skills training for students with behaviour disorders: a meta-analysis, *Advances in Learning and Behavioural Disabilities*, 11, pp. 1–26.
- KULIK, J.A. & KULIK, C.-L.C. (1989) Meta-analysis in education, *International Journal of Educational Research*, 13, pp. 221–340.
- LIPSEY, M.W. & WILSON, D.B. (1993) The efficacy of psychological, educational and behavioural treatment: Confirmation from meta-analysis, *American Psychologist*, 48, 12, pp 1181–1209.
- MEHAN, H., VILLANUEVA, T., HUBBARD, L. & LINTZ, A. (1996) *Constructing School Success: The Consequences of Untracking Low-Achieving Students* (Cambridge, Cambridge University Press).
- MORTON, S. (1999) Systematic Reviews and Meta-Analysis, Workshop materials on Evidence-Based Health Care (University of California, San Diego, La Jolla, California, Extended Studies and Public Programs).
- NOBLIT, G.W. & HARE, R.D. (1988) *Meta-Ethnography: Synthesizing Qualitative Studies* (Newbury Park, Sage Publications).
- PETROSINO, A.J. (1997) *What Works? Revisited Again: a Meta-Analysis of Randomized Experiments in Delinquency Prevention, Offender Rehabilitation and Deterrence* (Ann Arbor, Michigan, University of Michigan).
- PREISS, R.W. (1988) *Meta-Analysis: A Bibliography of Conceptual Issues and Statistical Methods* (Annandale, Virginia, Speech Communication Association).
- REEKER, J., ENSING, D., ELLIOTT, R. (1997) A meta-analytic investigation of group treatment outcomes for sexually abused children. *Child Abuse and Neglect*, 21, 7, pp. 669–680.
- SACKETT, D.L., ROSENBERG, W., GRAY, J.A.M., HAYNES, R.B. & RICHARDSON, W. (1996) Evidence-based medicine: what it is and what it isn't, *British Medical Journal*, 312, pp. 71–72.
- SLAVIN, R.E. (1984) Meta-analysis in education: How has it been used?, *Educational Researcher*, 13, pp. 6–15.
- SLAVIN, R.E. (1986) Best evidence synthesis: An alternative to meta-analysis and traditional reviews, *Educational Researcher*, 15, pp. 5–11.
- SLAVIN, R.E. & FASHOLA, O.S. (1998) *Show Me the Evidence!: Proven and Promising Programs for American Schools* (Thousand Oaks, California, Corwin Press).
- 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, pp. 419–433.
- SMITH, M.L., GLASS, G.V. & MILLER, T.I. (1980) *The Benefits of Psychotherapy* (Baltimore, Johns Hopkins University Press).
- TOOLEY, J., & DARBY, D. (1998) *Educational Research: An Ofsted Critique* (London, OFSTED).

- WAX, M. (1979) *Desegregated Schools: An Intimate Portrait Based on Five Ethnographic Studies* (Washington D.C., National Council of Education).
- WEISBURD, D. (1993) Design sensitivity in criminal justice experiments: Reassessing the relationship between sample size and statistical power, in: M. TONRY & N. MORRIS (eds) *Crime and Justice: An Annual Reviews of Research, Volume IV* (Chicago, University of Chicago Press).

Correspondence: Dr Philip Davies, Department for Continuing Education, University of Oxford, 1 Wellington Square, Oxford OX1 2JA, UK.