Trying to Do More Good than Harm in Policy and Practice: The Role of Rigorous, Transparent, Up-to-Date Evaluations

By
IAIN CHALMERS

Because professionals sometimes do more harm than good when they intervene in the lives of other people, their policies and practices should be informed by rigorous, transparent, up-to-date evaluations. Surveys often reveal wide variations in the type and frequency of practice and policy interventions, and this evidence of collective uncertainty should prompt the humility that is a precondition for rigorous evaluation. Evaluation should begin with systematic assessment of as high a proportion as possible of existing relevant, reliable research, and then, if appropriate, additional research. Systematic, up-to-date reviews of research—such as those that the Cochrane and Campbell Collaborations endeavor to prepare and maintain—are designed to minimize the likelihood that the effects of interventions will be confused with the effects of biases and chance. Policy makers and practitioners can choose whether, and if so how, they wish their policies and practices to be informed by research. They should be clear, however, that the lives of other people will often be affected by the validity of their judgments.

Keywords: evaluation; research synthesis; research methodology; ethics

Why Do We Need Rigorous, Transparent, Up-to-Date Evaluations of Policy and Practice?

It is the business of policy makers and practitioners to intervene in other people's lives. Although they usually act with the best of intentions, however, their policies and practices sometimes have unintended, unwanted effects, and they occasionally do more harm than good.

Iain Chalmers qualified in medicine in the mid-1960s and practiced as a clinician for seven years in the United Kingdom and the Gaza Strip. In the mid-1970s, after further training at the London School of Hygiene and Tropical Medicine and the London School of Economics and Political Science, he became a full-time health services researcher with a particular interest in assessing the effects of health care. He directed the National Perinatal Epidemiology Unit between 1978 and 1992 and the U.K. Cochrane Centre between 1992 and 2002.

DOI: 10.1177/0002716203254762

This reality should be their main motivation for ensuring that their prescriptions and proscriptions for others are informed by reliable research evidence.

In her address at the opening of the Nordic Campbell Center, Merete Konnerup, the director, gave three examples showing how the road to hell can be paved with the best of intentions (Konnerup 2002). An analysis of more than fifty studies suggests that effective reading instruction requires phonics and that promotion of the whole-language approach by educational theorists during the 1970s and 1980s seems likely to have compromised children's learning (National Institute of Child Health and Human Development 2000). A review of controlled assessments of driver education programs in schools suggests that these programs may increase road deaths involving teenagers: they prompt young people to start driving at an earlier age but provide no evidence that they affect crash rates (Achara et al. 2001). A review of controlled studies of "scared straight" programs for teenage delinquents shows that, far from reducing offending, they actually increase it (Petrosino, Turpin-Petrosino, and Finchenauer 2000; Petrosino, Turpin-Petrosino, and Buehler 2003).

One example of several that I could use to illustrate how my good intentions as a medical practitioner turned out to be lethal is the advice I promulgated after reading Benjamin Spock's (1966) record-breaking bestseller, *Baby and Child Care*. I bought the book when I was a recent medical graduate in the mid-1960s and marked the following passage:

> There are two disadvantages to a baby's sleeping on his back. If he vomits, he's more likely to choke on the vomitus. Also he tends to keep his head turned towards the same side. . . . this may flatten the side of his head. . . . I think it is preferable to accustom a baby to sleeping on his stomach from the start. (Pp. 163-64)

No doubt like millions of Spock's other readers, I passed on this apparently rational and authoritative advice. We now know that it led to thousands, if not tens of thousands, of avoidable sudden infant deaths (Chalmers 2001).

Uncertainty and Humility: Preconditions for Unbiased Evaluations

Individual policy makers and practitioners are often certain about things that are a matter of opinion. But surveys of practice reveal that these individual certain-

---

He is the editor of The James Lind Library, *and his main current interest is the history of the development of methods to test the effects of medical treatments (see www.jameslindlibrary.org).*

NOTE: This article is based on a presentation on 6 June 2002 at a symposium on "Randomised Controlled Trials in the Social Sciences," Nuffield College, Oxford, United Kingdom; and on preparations for the Jerry Lee Lecture at the third annual Campbell Colloquium, 27 February 2003, Stockholm, Sweden. I am grateful to Phil Alderson, Mike Clarke, Diana Elbourne, David Farrington, Judith Gueron, Tim Newburn, and Jan Vandenbroucke for comments on an earlier draft of this article and to Ann Oakley for providing unpublished information.

ties are often manifested in a very wide range of practices, not infrequently providing indirect evidence of mutually incompatible opinions. This evidence of collective uncertainty about the effects of policies and practices should prompt professionals and the public to find out which opinions are likely to be correct. A lack of empirical evidence supporting opinions does not mean that all the opinions are wrong or that, for the time being, policy and practice should not be based on people's best guesses. On matters of public importance, however, it should prompt efforts to obtain relevant evidence through evaluative research to help adjudicate among conflicting opinions.

Because professionals sometimes do more harm than good when they intervene in the lives of other people, their policies and practices should be informed by rigorous, transparent, up-to-date evaluations.

If advice as apparently innocuous and "theoretically sound" as recommending a baby's sleeping position can be lethal, there is clearly no room for complacency among professionals about their potential for harming those whom they purport to help. Evidence of collective uncertainty about the effects of their policies and practices should prompt the humility that is a precondition for rigorous evaluation. In a moving account, Judith Gueron (2002, 27-28) has reported how professionals delivering an education and training program for high school dropouts agreed to a randomized trial to assess its effects, in spite of their concern that this might fail to find any beneficial effects of their work. (In fact, the results of the trial were positive and led to a fifteen-site expansion serving hundreds of disadvantaged youth.)

A recent example from medical research illustrates the importance of remaining uncertain about the effects of an intervention until reliable evidence is available showing that it has at least some beneficial effects that outweigh negative effects (Freed et al. 2001). There have been reasons to hope that transplantation of fetal tissue into the brains of people with Parkinson's disease can improve the symptoms of that distressing condition. Accordingly, a randomized trial comparing fetal implants with placebo surgery was done to assess whether these hopes were borne out by experience. Not only did the study fail to detect any beneficial effects of the implants; it eventually showed that they seemed to cause a serious deterioration in symptoms in some patients.

Those patients who had been randomly assigned to placebo surgery were initially protected from these unanticipated adverse effects. But because they had been desperate to receive this new treatment, the clinical investigators had promised at the time they were randomized to placebo that they, too, would receive fetal implants after one year of follow-up. Unfortunately, the full extent of the adverse effects had not become clear within this time period, so the controls, too, were exposed to an intervention for which only adverse effects, and no benefits, have so far been shown in controlled trials (Freed et al. 2001).

One of the factors preventing a wider appreciation of the need for professional uncertainty and humility about the effects of interventions is that disappointing results tend to get hidden and forgotten. Studies that have yielded "disappointing" or "negative" results are less likely to be presented at scientific meetings; less likely to be reported in print; less likely to be published promptly, in full, in journals that are widely read, in English, and more than once; and less likely to be cited in reports of subsequent studies (Sterne, Egger, and Davey Smith 2001). An analysis of "successful case studies" in situational crime prevention (Clarke 1997), for example, is likely to be less informative than a systematic review of all relevant case studies—successful and unsuccessful.

How Should Uncertainties about the Effects of Policy and Practice Be Addressed?

Systematic reviews of existing research evidence: A scientific and ethical imperative

Whatever the study designs considered appropriate for reliable detection of the effects of policies and practices, individual studies should not be considered in isolation but interpreted in the context of systematic reviews incorporating any other, similar studies. Application of this principle in practice is no more or less than an acknowledgement that science is a cumulative activity. Yet the principle is widely ignored within academia, not only in "stand alone" reviews but also in the Discussion sections of reports of new studies (Clarke, Alderson, and Chalmers 2002).

The science of research synthesis—as in any other scientific research—implies that those who practice it will take steps to avoid misleading themselves and others by ignoring biases and the effects of chance. A systematic review thus has the same basic components as any other scientific investigation, and so involves

- stating the objectives of the research;
- defining eligibility criteria for studies to be included;
- identifying (all) potentially eligible studies;
- applying eligibility criteria;
- assembling the most complete data set feasible;
- analyzing this data set, using statistical synthesis and sensitivity analyses, if appropriate and possible; and
- preparing a structured report of the research.

It is not easy to conceptualize any justification for ignoring these principles, regardless of the sphere of scientific activity or the study designs and type of data available. Social scientists in the United States and their statistician colleagues have played a key role in the evolution of research synthesis, particularly since the late 1970s (Chalmers, Hedges, and Cooper 2002). Significant interest emerged in the medical field only in the late 1980s.

Some academics question the very notion of a systematic review (see, for example, Learmonth and Watson 1999; Webb 2001). The British educational researcher Martyn Hammersley (2001), for example, criticized the concept because the positivist model is committed to "procedural objectivity"; and he rejects the notion that bias "can and must be minimised," because this is "assumed to maximise the chances of producing valid conclusions" (p. 545). Hammersley's unfamiliarity with the field of research synthesis is revealed most clearly in the following:

> To be even more provocative, we could ask whether some of these forms of synthesis actually constitute reviewing the literature at all. A few seem to be closer to actually *doing* research, rather than reviewing it. (Hammersley 2002, 4)

This is a remarkably tardy insight, coming as it does two decades after a fellow educational researcher published a seminal paper pointing out that "integrative research reviews" are research projects in their own right (Cooper 1982).

Ignorance about the field of research synthesis and cavalier lack of concern about bias in reviews may simply reflect views about the purposes of research. Towards the end of his critique, Hammersley (2001) suggested,

> It is not proven that providing solutions to practical problems, or evaluating them, is the most important contribution which research can make to policy making and practice. (P. 550)

Views such as this may have prompted the U.K. secretary of state for education and employment, David Blunkett, to question the relevance of social science to government. The decision by Blunkett's department to establish a Centre for Evidence-Informed Policy and Practice at London's Institute of Education appears to have been driven partly by concerns about "ideology parading as intellectual inquiry and about the relevance and timeliness of research and the intelligibility of its results" (Boruch and Mosteller 2002, 2).

What kind of studies should be included in research syntheses to reduce biases in estimating intervention effects?

Study designs must be "fit for purpose"

Failure to distinguish research designs intended to lead to reliable causal inferences about the effects of interventions from other research designs, appropriate

for other purposes, is not uncommon (see, for example, Webb 2001). Researchers need to draw on a variety of research designs (Oakley 1999, 2000; Macintyre and Petticrew 2000), for example, to develop defining criteria for attention-deficit hyperactivity disorder, to survey the frequency of mental illness in prison populations, to investigate the validity of methods used to assess school performance, and to explore and record the subjective experiences of asylum seekers.

Surveys often reveal wide variations in the type and frequency of practice and policy interventions, and this evidence of collective uncertainty should prompt the humility that is a precondition for rigorous evaluation.

Indeed, a variety of study designs are required to assess the effects of specific factors on some health or social characteristic, life course, or putative "outcome." As the British sociologist John Goldthorpe (2001) has noted, a fundamental issue is whether the researchers can manipulate the factors concerned. Often this will not be possible, for example, in efforts to understand the effects on child development of genetic characteristics or of divorce. Studies of the relationship between child development and these factors may help to develop theory about the nature of the relationship and lead to ideas about how to intervene in an effort to protect or improve child development.

It is at this point—when interventions have been conceptualized on the basis of theory derived from observed associations—that it is important to ensure rigorous evaluation of the effects of these interventions, for example, gene therapy, marriage guidance, or child counseling. All such interventions can, in principle, be manipulated, and empirical evaluation in controlled experiments can assess whether they have the effects predicted by theory.

Sometimes the results of controlled experiments will be consistent with theory and can inform the development of policy and practices. On other occasions, controlled experiments will not yield evidence of the intervention effects predicted by theory. This does not necessarily mean that the theory is wrong; but it does mean that the possible reasons for the discrepancy between the predicted and observed effects should be explored, possibly leading to a refinement or rejection of the theory; and it should certainly be a warning that deploying the intervention in practice may do more harm than good.

Estimates of intervention effects vary with study design

Reliable studies of the effects of interventions are those in which the effects of policies or practices are unlikely to be confused with the effects of biases or chance. Rarely, estimates of the effects of interventions are so large that they are very unlikely to reflect the effects of insufficiently controlled biases or chance. Returning to an earlier example, once the adverse effects of placing babies to sleep on their tummies had been recognized, the effect of promulgating the opposite advice to the public in "Back to Sleep" campaigns was dramatic—a reduction in death rates to between a half and a quarter of their previous levels—and unlikely to be explained by biases or regression to the mean (Gilbert 1994; Wennergren et al. 1997).

Usually, however, plausible effects of policies are modest but worth knowing about. In these circumstances, research syntheses must be designed in ways that minimize the effects of biases and chance. For example, we would probably still not have learned that very low doses of aspirin offer the potential for an important reduction in the risk of suffering cardiovascular morbidity and mortality had investigators not prepared scientifically robust syntheses of scientifically robust studies (Antiplatelet Trialists' Collaboration 1988).

Reliable detection of moderate but important real intervention effects requires adequate control of the biases that may distort estimates of effects and of the effects of chance. The effects of biases and chance can mislead people into believing that useless or harmful interventions are worthwhile (as has been the case with long-standing claims that postmenopausal hormone therapy reduces the risk of cardiovascular disease) or that interventions are useless when, in truth, they have beneficial effects (see explanation of Cochrane Collaboration logo, below).

People considering which studies should be included in systematic reviews of research assessing the effects of interventions must take into account that studies with different research designs tend to yield different estimates of the effects of interventions (Kunz and Oxman 1998; Britton et al. 1998; MacLehose et al. 2000; Kunz, Vist, and Oxman 2003). For example, in a comparison of the results of studies to assess the effects of crime reduction strategies, Weisburd, Lum, and Petrosino (2001) found that estimates of effect sizes were larger in studies in which there had been fewer precautions to minimize biases. Even in studies that purport to have used random allocation or alternation to create comparison groups, those in which the allocation schedule has not been concealed from the people making decisions about the eligibility and assignment of participants yield larger estimates of treatment effects (Juni, Altman, and Egger 2001; Kunz and Oxman 1998; Kunz, Vist, and Oxman 2003).

There is no easy escape from the dilemma posed by these differences. Although observational data yield estimates of effects that are larger, *on average*, than those using data from randomized trials, in any particular instance it is not possible to predict whether different estimates will emerge using the two different approaches. One cannot even predict with confidence the direction of any differences that are found (Kunz, Vist, and Oxman 2003).

Random allocation is the only defining characteristic of randomized trials

Just as social scientists in the United States have pioneered research synthesis, so also have they pioneered the use of randomized trials to assess the effects of social and educational interventions (Boruch 1997; Petrosino et al. 2000). Some commentators reject the use of randomized trials to test social and educational interventions (see, for example, Dobash and Dobash 2000; Prideaux 2002; Kippax 2003). These comments sometimes reveal a failure to understand that the *one and only* defining feature of randomized trials is random allocation to comparison groups to abolish selection bias and, thus, to ensure that unmeasured as well as measured factors of prognostic importance in the comparison groups differ only by chance (Kleijnen et al. 1997).

A professor of education writing in the *British Medical Journal*, for example, stated,

> Randomisation relies on the maintenance of blind allocation. Maintaining blinding is rarely possible in research on educational interventions. (Prideaux 2002)

And a reviewer consulted by the Economic and Social Research Council about a proposal to prepare systematic reviews of randomized trials and studies with other designs stated (Ann Oakley, personal communication 2002),

> With double bind [*sic*] and other safeguards generally impossible in social science research, and typically with biases due to differential attrition, it is not evident that randomised control trials are invariably preferable.

A comment from another, anonymous, reviewer of the proposal is illustrative of the genre of vague statements, unsupported by any reference to empirical evidence, that often characterize comments about randomized trials:

> The straightforward extrapolation of judgements about rigour and generalisability from medical to behavioural evaluation by randomised comparison can, of course, be subjected to a quite serious empirical and theoretical critique. Such a critique would argue that randomised comparisons can yield biased assessments of true effects of interventions.

Sometimes comments on randomized trials are little more than polemic and the erection of straw men:

> Randomized designs have, like all designs, important limitations. (Dobash and Dobash 2000, 257)
>
> It is not the case, even in abstract terms, that some research designs have all the advantages and others have none. (Hammersley 2001, 547).
>
> The orthodoxy of experimental manipulation and RCTs is dangerous when applied unthinkingly to health promotion. (Kippax 2003, 30)

Those who reject randomization are implying they are sufficiently knowledgeable about the complexities of influences in the social world that they know how to take account of all potentially confounding factors of prognostic importance, including those they have not measured, when comparing groups to estimate intervention effects.

Double standards on the ethics of experimentation

Additional misconceptions result from unacknowledged double standards on the ethics of evaluative studies. As Donald Campbell (1969) noted many years ago, selectively designating some interventions as "experiments"—a term loaded with negative associations—ignores the reality that policy makers and practitioners are experimenting on other people most of the time. The problem is that their experiments are usually poorly controlled. Dr Spock's ill-founded advice would

Evaluation should begin with systematic assessment of as high a proportion as possible of existing relevant, reliable research, and then, if appropriate, additional research.

probably not be conceptualized by many people as a poorly controlled experiment, yet that is just what it was. Had he proposed testing the effect of his advice on infant mortality in a well-controlled evaluation, however, many people would have had no hesitation in characterizing that as "an experiment," invoking all the "guinea pig" images conjured up by that term in the public's mind.

As noted in a *Lancet* editorial published more than a decade ago, "The clinician who is convinced that a certain treatment works will almost never find an ethicist in his path, whereas his colleague who wonders and doubts and wants to learn will stumble over piles of them" (Medical ethics 1990, 846). Or, as put more bluntly by the pediatrician Richard Smithells (1975, 41), "I need permission to give a drug to half of my patients, but not to give it to them all."

This double standard (Chalmers and Lindley 2000) results in some bizarre ethical analyses (see, for example, Graebsch 2000). Professionals who are uncertain about whether a particular intervention (a policy or practice) will do more good than harm, and so wish to offer it only within the context of a controlled trial so that they protect people in the face of current uncertainty and learn about its effects, are expected to observe elaborate informed consent rituals. If exactly the same

intervention is offered by other professionals—because it was recommended during their professional training three decades previously, or because there is a plausible theory that suggests it will be helpful, or because it is an accepted routine, or because they or the institutions for which they work have a vested financial or political interest in promulgating it (Oxman, Chalmers, and Sackett 2001)—the standard of consent is relaxed.

People not infrequently raise questions about the ethics of well-controlled, randomized experiments designed to address uncertainties about the effects of inadequately evaluated policies and practices. They would do well to consider the ethics of acquiescing in professional promulgation of the same policies and practices among recipients who have not been made aware either of the lack of reliable evidence of their effects or of the real reasons that they are being recommended to accept these interventions.

What can be done to reduce the effects of chance?

As with the methods to reduce biases in systematic reviews, social scientists and statisticians in the United States were prominent among those developing methods to reduce the effects of chance using quantitative synthesis of the results of separate but similar studies (Chalmers, Hedges, and Cooper 2002). Indeed, it was an American social scientist who coined the term "meta-analysis" to describe this process (Glass 1967).

Sometimes meta-analysis is impossible with the data available, and even when it is possible it may not be appropriate. When it is both possible and judged appropriate, however, meta-analysis can reveal "reconcilable differences" among studies. The Cochrane Collaboration logo (Figure 1), for example, is based on a meta-analysis of data from seven randomized trials. Each horizontal line represents the results of one trial (the shorter the line, the more certain the result), and the diamond represents their combined results. The vertical line indicates the position around which the horizontal lines would cluster if the two treatments compared in the trials had similar effects; if a horizontal line touches the vertical line, it means that that particular trial found no statistically significant difference between the treatments. The position of the diamond to the left of the vertical line indicates that the treatment studied is beneficial.

This diagram shows the results of a systematic review of randomized trials of a short, inexpensive course of a corticosteroid given to women expected to give birth prematurely. The first of these randomized trials was reported in 1972. The diagram summarizes the evidence that would have been revealed had the available randomized trials been reviewed systematically a decade later: it indicates strongly that corticosteroids reduce the risk of babies dying from the complications of immaturity. By 1991, seven more trials had been reported, and the picture in the logo had become still stronger. This treatment reduces the odds of the babies of these women dying from the complications of immaturity by 30 to 50 percent. Because no systematic review of these trials had been published until 1989, however, most obstetricians had not realized that the treatment was so effective. As a

FIGURE 1
THE COCHRANE COLLABORATION®

result, tens of thousands of premature babies have probably suffered and died unnecessarily (and cost the health services more than was necessary). This is just one of many examples of the human costs resulting from failure to perform systematic, up-to-date reviews of randomized trials of health care.

One of the reasons that the Cochrane logo conveys the message it does is that estimates of the effects of the treatment have been shown as 95 percent confidence intervals. Emphasis on point estimates of effects and reliance on p values derived from statistical tests can result in failure to detect possible effects of interventions that may be important. This danger is illustrated in a paper by two British criminologists titled "The Controlled Trial in Institutional Research—Paradigm or Pitfall for Penal Evaluators?" (Clarke and Cornish 1972). This drew on the authors' experience of a randomized trial of a therapeutic community for young offenders. Because similar numbers of boys in the experimental and control groups went on to reoffend, the authors concluded that therapeutic communities were ineffective and that randomized trials are inappropriate for assessing the effects of institutional interventions.

Had they taken account of the confidence interval surrounding the point estimate of the difference between experimental and control groups, as well as the results of other, similar studies, they might have come to a more cautious conclusion (Table 1). An overall estimate of the effects of therapeutic communities based on a systematic review of eight randomized trials suggests that this category of intervention may halve the odds of adverse outcomes, an effect of great public

TABLE 1
SYSTEMATIC REVIEW OF EIGHT RANDOMIZED CONTROLLED TRIALS ASSESSING THE EFFECTS OF THERAPEUTIC COMMUNITIES ON ADVERSE OUTCOMES

	Odds Ratio	95 Percent Confidence Interval
All (N = 8)	0.46	0.39 to 0.54
Secure democratic		
Cornish and Clarke (1975)	1.04	0.76 to 2.79
Auerbach (1978)	0.52	0.28 to 0.98

SOURCE: NHS Centre (1999).

importance if true. An analysis restricted to the two trials of the "secure democratic" model studied by Clarke and Cornish (1972) suggests that although the beneficial effect may be somewhat less in these, the evidence is still suggestive of a potentially very important benefit. As a consequence of a failure to take proper account of the effects of chance, a useful methodology and a useful intervention may both have been jettisoned prematurely.

Systematic Reviews Need to Be Rigorous, Transparent, and Up to Date

Whatever decisions are made about which studies are eligible for inclusion in systematic reviews, and whether or not meta-analysis is used to analyze them, reviews should be published in sufficient detail to enable readers to judge their reliability. The advent of electronic publishing has transformed the potential for providing the detail required and allows systematic reviews to be updated when additional data become available and improved in other ways when ways of doing this are identified, for example, to incorporate relevant qualitative data (see, for example, Burns et al. 2001). Electronic publication also facilitates prompt publication of comments and criticisms.

The advantages of electronic publication are particularly welcome when the matter at issue is very contentious. A very extensive systematic review of the effects of water fluoridation (www.york.ac.uk/inst/crd/fluorid.htm) shows how electronic media enable research synthesis to be done transparently, accountably, and democratically. For many years, there have been two opposing lobbies on this issue in the United Kingdom. Following a debate in the House of Lords, the government commissioned the NHS Centre for Reviews and Dissemination in York to review the relevant evidence. An advisory group, on which the two main warring parties were both represented, was established to agree a protocol for the review before the data collection started. The list of studies to be assessed for eligibility was posted on a public Web site, and people were invited to suggest additional studies for consider-

ation. As the review progressed, the Web site showed the results of applying the agreed inclusion and exclusion criteria and displayed the data abstracted from eligible studies and eventually the draft data tables. As it happens, the investigators were unable to identify any randomized experiments of water fluoridation, and they were disappointed with the quality of most of the observational data (McDonagh et al. 2000). (These suggested a modest reduction in caries and an increase in disfiguring dental fluorosis.)

This transparent process is relevant to a point made by the president of the Royal Statistical Society in 1996. After referring approvingly to the Cochrane Collaboration—which prepares, maintains, and disseminates systematic reviews of the effects of healthcare interventions (Chalmers 1993)—he wrote,

> But what's so special about medicine? We are, through the media, as ordinary citizens, confronted daily with controversy and debate across a whole spectrum of public policy issues. But typically, we have no access to any form of systematic "evidence base"—and therefore no means of participating in the debate in a mature and informed manner. Obvious topical examples include education—what does work in the classroom?—and penal policy—what is effective in preventing reoffending? (Smith 1996)

It was after reading this presidential address and Robert Boruch's excellent book, *Randomized Experiments for Planning and Evaluation* (1997), that I decided to beat a path to the latter's door in October 1998. I wanted to try to persuade him to take up the challenge of leading an effort to establish an analogue to the Cochrane Collaboration to prepare systematic reviews of social and educational interventions. For reasons that should now be clear, although I felt it was essential that such collaboration should be international, I believed that it would fail without the leadership and active involvement of social scientists in the United States, and I suggested that it might be named after one of them—Donald Campbell.

The Cochrane Collaboration and the Campbell Collaboration are both exploiting the advantages of electronic media. Electronic publication means that protocols (containing the introduction to and materials and methods planned for each review) as well as complete reports of systematic reviews can be made publicly available in considerably more detail and promptly after submission than is usually possible with print journals, and that they can be modified in the light of new data or comments.

As far as I am aware, these two collaborations currently provide the only international infrastructure for preparing *and maintaining* systematic reviews in the fields of health and social care and education. Estimates suggest that more than ten thousand people are now contributing to the Cochrane Collaboration (which was inaugurated in 1993), most of them through one or more of fifty Collaborative Review Groups (all international), which have collectively published nearly two thousand systematic reviews in *The Cochrane Database of Systematic Reviews*. Members of these groups are supported by ten Cochrane Methods Groups (all international) and twelve Cochrane Centres, which are geographically based, and share collective responsibility for global coverage (www.cochrane.org).

The Campbell Collaboration (which was inaugurated in 2000) currently consists of Coordinating Groups in Crime and Justice, Education, and Social Welfare, with more than fifty registered titles of reviews in preparation. These groups preparing reviews are supported by a Methods Group, as well as by the Collaboration's Secretariat and staff at the Nordic Campbell Centre (www.sfi.dk/sw1270.asp). There is growing recognition of the need for international collaboration in preparing systematic reviews. For example, a review of British educational research conducted in 2002 by the Organization of Economic Cooperation and Development (2002) commended the work on systematic reviews being coordinated by the Evidence for Policy and Practice Information and Co-ordinating Centre (the EPPI-Centre) but noted,

> Making the EPPI Centre activities broadly international in scope (perhaps by increasing collaboration with the Campbell Collaborative [*sic*]) could further increase the gain policy makers and the research community may expect from the EPPI Centre. If similar centres could be created in other countries and similar reviews conducted, the gain in knowledge would be greater and some economies of scale could be expected in terms of methodology. (Para. 66)

There are encouraging examples of Campbell and Cochrane groups working collaboratively, especially to tackle methodological challenges, for example, to explore how best to incorporate qualitative and economic data within systematic reviews assessing the effects of policies and practices.

The Role of Systematic Reviews of Research Evidence in the Development of Policy and Practice

Conclusions about the effects of policies and practices will always remain a matter of judgement. As Xenophanes put it in the sixth century B.C., "Through seeking we may learn and know things better. But as for certain truth, no man hath known it, for all is but a woven web of guesses." Or, in Mervyn Susser's words twenty-five centuries later,

> Our many errors show that the practice of causal inference . . . remains an art. Although to assist us, we have acquired analytic techniques, statistical methods and conventions and logical criteria, ultimately the conclusions we reach are a matter of judgement. (1984, 846)

Our judgments can affect other people's lives, however. After comparing the results of systematic reviews with the recommendations of experts writing textbooks and narrative review articles, Antman and his colleagues (1992) concluded that because reviewers have not used scientific methods, advice on some life-saving therapies has been delayed for more than a decade, while other treatments have been recommended long after controlled research has shown them to be harmful.

One needs to bear in mind Xenophanes' words and empirical evidence of this kind when assessing nonspecific questions about the validity of systematic reviews. Hammersley (2001, 547) is not the only person to have asked the question, "Where is the evidence that systematic reviews produce more valid conclusions than narrative reviews?"

Not only do those who pose such questions ignore the existing evidence, they almost never confront the reality that different methods of reviewing tend to lead to different conclusions or explore the reasons for and consequences of this. For policy makers, practitioners, and others wishing to use research evidence to inform

---

[Policy makers and practitioners] should be clear, however, that the lives of other people will often be affected by the validity of their judgments.

---

their choices about interventions, these discrepancies obviously have practical implications: which reviews should they believe? As I have made clear elsewhere, evidence of the discrepant conclusions of systematic and narrative reviews in the health field leave me in no doubt about which type of review I wish to be taken into account when I am a patient (Chalmers 1995, 2000, 2001). And if I had a delinquent teenage son, I would be in no doubt that I would not wish him to be exposed to a "Scared Straight" program, however many uncontrolled before-and-after observational studies suggested that this would divert him from a criminal career (Petrosino, Turpin-Petrosino, and Buehler 2003). Put bluntly, it is time that those academics who offer general—often polemic—criticisms of efforts to reduce the effects of bias and chance in reviews begin to face up to the reality that different materials and methods used for reviews usually result in different conclusions, and show which conclusions they would prefer, and why.

None of the foregoing is meant to suggest that systematic reviews of research evidence speak for themselves. They do not, as has been stated repeatedly by those involved in this work. But up-to-date, reliable, systematic reviews of research evidence, or a demonstration that no relevant research exists, should be regarded as desirable and often essential for informing policy and practice. Judgments will always be needed about how to use the evidence derived from evaluative research. As well as the research evidence, these judgments need to take account of needs, resources, priorities and preferences, and other factors.

I can illustrate how evidence does not speak for itself by drawing on a personal experience. Many years ago, I worked for two years as a United Nations medical officer in a Palestinian refugee camp in the Gaza Strip. Some of my child patients who developed measles, who were often malnourished, died from the complications of the disease. During my medical training, it had been impressed on me that I should not prescribe antibiotics routinely to children with measles. Had I had access then to the Cochrane review of controlled trials of prophylactic antibiotics in measles (Shann, D'Souza, and D'Souza 2002), the authority of the research evidence would have trumped the authority of my teachers at medical school, and I would have used antibiotic prophylaxis. I believe that this would have prevented some of my child patients from suffering and dying. Although that would have been my response to the research evidence, however, this is not to suggest that everyone would have responded similarly (Chalmers 2002). Factors that might weigh more heavily in the judgments of others in different circumstances include the moderate quality of the available relevant studies, the likely magnitude of the beneficial effects, costs, and concerns about the development of antibiotic resistance.

This brings me back to the rationale for evaluations of policy and practice that are rigorous, transparent, and up to date, namely, that policy makers and practitioners who intervene in other people's lives should acknowledge that although they act with best of intentions, they may sometimes do more harm than good. That possibility should be sufficient motivation for them to ensure that their prescriptions and proscriptions are informed—even if not dictated—by reliable research evidence.

Concluding Observations

I have tried to make clear and to justify in this article how I conceptualize reliable research evidence. This entails the preparation of systematic reviews designed to minimize bias, drawing on research studies designed to minimize bias. I have deliberately concentrated on bias because the other important issue, taking account of the effects of chance, is a more straightforward matter (by using meta-analysis and doing larger studies). I believe that the principle of minimizing bias applies across all of science, and certainly in applied fields like the health and social sciences, because of the impact research may have on policies and practices.

In conclusion, my interest in research to assess the effects of interventions arises from a long-standing concern that, acting with the best of intentions, policy makers and practitioners have sometimes done more harm than good when interfering in the lives of others. I believe that the empirical evidence showing associations between study design and study results—whether among reviews or among individual studies—is likely to be explained by differential success in controlling biases. If only as a patient, therefore, I want decisions about my care to take account of the results of systematic reviews and studies that have taken measures to reduce the effects of biases and chance. As a citizen, too, I want these principles to be respected more generally—by policy makers, practitioners, and the public—

than they are currently. However, to return to my starting point, uncertainty and humility among policy makers, practitioners, and researchers are the preconditions for wider endorsement of the approaches I have outlined. Sadly, these qualities are too often in short supply.

References

Achara, S., B. Adeyemi, E. Dosekun, S. Kelleher, M. Lansley, I. Male, N. Muhialdin, L. Reynolds, I. Roberts, M. Smailbegovic, and N. van der Spek. 2001. Evidence based road safety: The Driving Standards Agency's schools programme. *Lancet* 358:230-32.

Antiplatelet Trialists' Collaboration. 1988. Secondary prevention of vascular disease by prolonged antiplatelet treatment. *BMJ* 296:320-31.

Antman, E. M., J. Lau, B. Kupelnick, F. Mosteller, and T. C. Chalmers. 1992. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *Journal of the American Medical Association* 268:240-48.

Auerbach, A. W. 1978. The role of the therapeutic community "Street Prison" in the rehabilitation of youthful offenders. University Microfilms no. 78-01086. Doctoral diss., George Washington University, Washington, DC.

Boruch, R. 1997. *Randomized experiments for planning and evaluation: A practical guide.* Thousand Oaks, CA: Sage.

Boruch, R., and F. Mosteller. 2002. Overview and new directions. In *Evidence matters: Randomised trials in education research*, edited by F. Mosteller and R. Boruch, 1-14. Washington, DC: Brookings Institution.

Britton, A., M. McKee, N. Black, K. McPherson, C. Sanderson, and C. Bain. 1998. Choosing between randomised and non-randomised studies: A systematic review. *Health Technology Assessment* 2 (13): 1-124.

Burns, T., M. Knapp, J. Catty, A. Healey, J. Henderson, H. Watt, and C. Wright. 2001. Home treatment for mental health problems: A systematic review. *Health Technology Assessment* 5 (15): 1-139.

Campbell, D. T. 1969. Reforms as experiments. *American Psychologist* 24:409-29.

Chalmers, I. 1993. The Cochrane Collaboration: Preparing, maintaining and disseminating systematic reviews of the effects of health care. In "Doing more good than harm: The evaluation of health care interventions," edited by K. S. Warren and F. Mosteller. *Annals of the New York Academy of Sciences* 703 (special iss.): 156-63.

———. 1995. What do I want from health research and researchers when I am a patient? *BMJ* 310:1315-18.

———. 2000. A patient's attitude to the use of research evidence to guide individual choices and decisions in health care. *Clinical Risk* 6:227-30.

———. 2001. Invalid health information is potentially lethal. *BMJ* 322:998.

———. 2002. Why we need to know whether prophylactic antibiotics can reduce measles-related morbidity. *Pediatrics* 109:312-15.

Chalmers, I., L. V. Hedges, and H. Cooper. 2002. A brief history of research synthesis. *Evaluation and the Health Professions* 25:12-37.

Chalmers, I., and R. Lindley. 2000. Double standards on informed consent to treatment. In *Informed consent in medical research*, edited by L. Doyal and J. S. Tobias, 266-75. London: BMJ Publications.

Clarke, M., P. Alderson, and I. Chalmers. 2002. Discussion sections in reports of controlled trials published in general medical journals. *Journal of the American Medical Association* 287:2799-801.

Clarke, R. V. 1997. *Situational crime prevention: successful case studies.* 2d ed. New York: Harrow and Heston.

Clarke, R. V. G., and D. B. Cornish. 1972. *The controlled trial in institutional research—Paradigm or pitfall for penal evaluators?* Home Office Research Study no. 15. London: Her Majesty's Stationery Office.

Cooper, H. M. 1982. Scientific principles for conducting integrative research reviews. *Review of Educational Research* 52:291-302.

Cornish, D. B., and R. V. G. Clarke. 1975. *Residential treatment and its effects on deliquency.* Home Office Research Study no. 32. London: Her Majesty's Stationery Office.

Dobash, R. E., and R. P. Dobash. 2000. Evaluating criminal justice interventions for domestic violence. *Crime & Delinquency* 46:252-70.

Freed, C. R., P. E. Greene, R. E. Breeze, W. Y. Tsai, W. DuMouchel, R. Kao, S. Dillon, H. Winfield, S. Culver, J. Q. Trojanowski, D. Eidelberg, and S. Fahn. 2001. Transplantation of embryonic dopamine neurons for severe Parkinson's disease. *New England Journal of Medicine* 344:710-19.

Gilbert, R. 1994. The changing epidemiology of SIDS. *Archives of Disease in Childhood* 70:445-49.

Glass, G. V. 1967. Primary, secondary and meta-analysis of research. *Educational Researcher* 10:3-8.

Goldthorpe, J. H. 2001. Causation, statistics, and sociology. *European Sociological Review* 17:1-20.

Graebsch, C. 2000. Legal issues of randomized experiments on sanctioning. *Crime & Delinquency* 46:271-82.

Gueron, J. 2002. The politics of random assignment: Implementing studies and affecting policy. In *Evidence matters: Randomised trials in education research*, edited by F. Mosteller and R. Boruch, 15-49. Washington, DC: Brookings Institution.

Hammersley, M. 2001. On "systematic" reviews of research literatures: A "narrative" response to Evans & Benefield. *British Educational Research Journal* 27:543-54.

———. 2002. Systematic or unsystematic, is that the question? Some reflections on the science, art, and politics of reviewing research evidence. Text of a talk given to the Public Health Evidence Steering Group of the Health Development Agency, October, in London.

Juni, P., D. G. Altman, and M. Egger. 2001. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 323:42-46.

Kippax, S. 2003. Sexual health interventions are unsuitable for experimental evaluation. In *Effective sexual health interventions: Issues in experimental evaluation*, edited by J. Stephenson, J. Imrie, and C. Bonell, 17-34. Oxford: Oxford University Press.

Kleijnen, J., P. Gøtzsche, R. H. Kunz, A. D. Oxman, and I. Chalmers. 1997. So what's so special about randomisation? In *Non-random reflections on health services research: On the 25 anniversary of Archie Cochrane's* Effectiveness and efficiency, edited by A. Maynard and I. Chalmers, 93-106. London: BMJ Books.

Konnerup, M. 2002. The three main pillars of the Campbell Collaboration. Presented at Nordic Campbell Center Inauguratiou Seminar, 12 November, in Copenhagen, Denmark. Retrieved 18 January 2003 from from http://www.nordic-campbell.dk/MereteKonnerupsforedrag4/index.htm.

Kunz, R., and A. D. Oxman. 1998. The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 317:1185-90.

Kunz, R., G. Vist, and A. D. Oxman. 2003. Randomisation to protect against selection bias in healthcare trials (Cochrane methodology review). In *The Cochrane Library*, iss. 1. Oxford: Update Software.

Learmonth, A. M., and N. J. Watson. 1999. Construing evidence-based health promotion: perspectives from the field. *Critical Public Health* 9:317-33.

Macintyre, S., and M. Petticrew. 2000. Good intentions and received wisdom are not enough. *Journal of Epidemiology and Community Health* 54:802-3.

MacLehose, R. R., B. C. Reeves, I. M. Harvey, T. A. Sheldon, I. T. Russell, and A. M. Black. 2000. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment* 4 (34): 1-154.

McDonagh, M. S., P. F. Whiting, P. M. Wilson, A. J. Sutton, I. Chestnutt, J. Cooper, K. Misso, M. Bradley, E. Treasure, and J. Kleijnen. 2000. Systematic review of water fluoridation. *BMJ* 321:855-59.

Medical ethics—Should medicine turn the other cheek? 1990. *Lancet* 336:846-47.

National Institute of Child Health and Human Development. 2000. *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. NIH Publication no. 00-4769. Washington, DC: Government Printing Office.

NHS Centre for Reviews and Dissemination, York and School of Sociology & Social Policy, Nottingham. 1999. *Therapeutic community effectiveness*. CRD Report 17. Retrieved 18 January 2003 from www.york.ac.uk/inst/crd/.

Oakley, A. 1999. Paradigm wars. *International Journal of Social Research Methodology* 2:247-54.

———. 2000. *Experiments in knowing*. Oxford: Polity Press.

Organization of Economic Cooperation and Development. 2002. *Educational research and development in England. Examiners' report.* CERI/CD10, September, p. 21.

Oxman, A. D., I. Chalmers, and D. L. Sackett. 2001. A practical guide to informed consent to treatment. *BMJ* 323:1464-66.

Petrosino, A., R. F. Boruch, C. Rounding, S. McDonald, and I. Chalmers. 2000. *The Campbell Collaboration Social, Psychological, Educational & Criminological Trials Register (C2-SPECTR). Evaluation and Research in Education* 14:206-19.

Petrosino, A., C. Turpin-Petrosino, and J. Buehler. 2003. "Scared Straight" and other juvenile awareness programs for preventing juvenile delinquency (Cochrane Review). In *The Cochrane Library*, iss. 1. Oxford: Update Software.

Petrosino, A., C. Turpin-Petrosino, and J. O. Finchenauer. 2000. Well-meaning programs can have harmful effects! Lessons from experiments such as Scared Straight. *Crime & Delinquency* 46:354-79.

Prideaux, D. 2002. Researching the outcomes of educational interventions: A matter of design *BMJ* 324:126-27.

Shann, F., R. M. D'Souza, and R. D'Souza. 2002. Antibiotics for preventing pneumonia in children with measles (Cochrane Review). In *The Cochrane Library*, iss. 2. Oxford: Update Software.

Smith, A. 1996. Mad cows and ecstasy. *Journal of the Royal Statistical Society* 159:367-83.

Smithells, R. W. 1975. Iatrogenic hazards and their effects. *Postgraduate Medical Journal* 15:39-52.

Spock, B. 1966. *Baby and child care.* 165th printing. New York: Pocket Books.

Sterne, J. A. C., M. Egger, and G. Davey Smith. 2001. Investigating and dealing with publication and other biases. In *Systematic reviews in health care: Meta-analysis in context*, 2d ed. of Systematic Reviews, edited by M. Egger, G. Davey Smith, and D. Altman, 189-208. London: BMJ Books.

Susser, M. 1984. Causal thinking in practice: Strengths and weaknesses of the clinical vantage point. *Pediatrics* 74:842-49.

Webb, S. A. 2001. Some considerations of the validity of evidence-based practice in social work. *British Journal of Social Work* 31:57-79.

Weisburd, D., C. M. Lum, and A. Petrosino. 2001. Does research design affect study outcomes in criminal justice? *Annals of the American Academy of Political and Social Science* 578:50-70.

Wennergren, G., B. Alm, N. Oyen, K. Helweg-Larsen, J. Milerad, R. Skjaerven, S. G. Norvenius, H. Lagercrantz, M. Wennborg, A. K. Daltveit, T. Markestad, and L. M. Irgens. 1997. The decline in the incidence of SIDS in Scandinavia and its relation to risk-intervention campaigns. Nordic Epidemiological SIDS Study. *Acta Paediatr* 86:963-68.