Designing Randomised Trials in Health, Education and the Social Sciences

An Introduction

David J. Torgerson and
Carole J. Torgerson

Designing Randomised Trials in Health, Education and the Social Sciences

Also by Carole J. Torgerson
SYSTEMATIC REVIEWS

Designing Randomised Trials in Health, Education and the Social Sciences

An Introduction

David J. Torgerson and Carole J. Torgerson
University of York

palgrave
macmillan

Softcover reprint of the hardcover 1st edition 2008 978- 0-230-53735-4

First published 2008 by
PALGRAVE MACMILLAN
Houndmills, Basingstoke, Hampshire RG21 6XS and
175 Fifth Avenue, New York N.Y. 10010
Companies and representatives throughout the world

PALGRAVE MACMILLAN is the global academic imprint of the Palgrave Macmillan division of St. Martin's Press, LLC and of Palgrave Macmillan Ltd. Macmillan® is a registered trademark in the United States, United Kingdom and other countries. Palgrave is a registered trademark in the European Union and other countries.

ISBN 978-0-230-53736-1 ISBN 978-0-230-58399-3 (eBook)
DOI 10.1057/9780230583993

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Torgerson, David J. (David John), 1960–
Designing randomised trials in health, education and the social sciences / David J. Torgerson and Carole J. Torgerson.
p. ; cm.
Includes bibliographical references and index.

ISBN 978-0-230-53736-1 (paperack : alk. paper) 1. Clinical trials. 2. Social sciences—Research—Methodology. I. Torgerson, Carole. II. Title.
[DNLM: 1. Health Services Research. 2. Randomized Controlled Trials as Topic—methods. 3. Education. 4. Research Design. 5. Social Sciences. W 84.3 T682d 2008]
R853.C55T67 2008
610.72'4—dc22 2007052309

10 9 8 7 6 5 4 3 2 1
17 16 15 14 13 12 11 10 09 08

Transferred to Digital Printing 2012

Contents

List of Figures, Boxes and Tables

Figures

Boxes

Tables

Preface

The randomised controlled trial (RCT) has long been considered the 'gold-standard' method for establishing effectiveness in health care research. Many hundreds of thousands of health care RCTs have been published. The fiftieth anniversary of the 1948 RCT of streptomycin was widely celebrated by health care researchers in 1998. However, one wonders how many educational or social science researchers are aware of the larger 1931 and 1932 randomised trials of an educational intervention conducted by Walters? And if they had known would the seventy-fifth anniversary, in 2007, of those trials have been a cause for celebration? Relatively few RCTs have been undertaken in the wider social sciences. Many methodological advances in the design of trials have been undertaken in health care, which are directly applicable to other areas. Whilst many social science research methods texts have been published, little attention or detail is given to the design and conduct of the RCT in such texts. In health sciences research several excellent texts describe the RCT, usually from a statistical standpoint, which make them less accessible to the non-statistician or general research methodologist. This book is an attempt to remedy this deficit. We avoid, as far as possible, detailed statistical arguments or formulae. Instead we focus on the importance of trial design.

Since the early descriptions of the RCT there has been a tremendous amount of methodological work, mainly in health care trials. As health trialists have widened their remit away from the placebo controlled drug trial, methodological innovations have been developed to deal with the threat of post-randomisation biases. In this book we detail these threats and describe different trial designs that can as easily be applied to the wider social sciences as they can be to health care trials. Many published trials make elementary mistakes that undermine their validity. By way of example we discuss how to avoid these problems through proper design and thereby, hopefully, develop a design that will produce reliable results. We think that there will be a revived interest in the RCT across the social sciences as politicians and policy-makers begin to crave evidence for 'what works?' The most reliable guide to providing evidence on what works is the RCT. Other approaches, long over-used by researchers and practitioners, are nearly always subject to inherent flaws, which can render their results

uninterpretable. In this book we describe the main justification for the RCT, and details of how we should randomise, and information about potential bias. We also look in detail at different trial designs, at how to appraise trial quality and we outline the importance of economic analysis alongside RCTs.

Acknowledgements

We thank Joy Adamson, Martin Bland, Iain Chalmers and Jo Dumville for their helpful comments on earlier draft chapters and Alison Robinson for proof reading the manuscript. We also thank Jo North for her careful copy editing of the book. Any remaining errors are the authors' responsibility. We also thank William Torgerson. We acknowledge support from the ESRC Researcher Development Initiative: Training in pragmatic social interventions.

Glossary of Terms

Active or on treatment analysis – An analytical method whereby only those who comply with their assigned intervention are included in the analysis. Non-compliers may be analysed in the control condition. Method violates randomisation and can introduce bias.

Allocation concealment – This prevents foreknowledge of allocation of an individual by the researcher, participant or practitioner. This is important because random allocation can be undermined if participants are chosen to be in a desired group.

Alternation – A non-random method of forming comparator groups, whereby trial participants are alternately assigned to treatment or to act as controls.

Attrition – Some participants are lost during the study and cannot be included in the analysis. This is termed attrition.

Before and after (pre- and post-test) – The weakest form of quantitative evaluation. Participants are measured at a point in time, given an intervention and then re-measured. Any change is attributed to the intervention. Bias is a strong possibility due to temporal and regression to the mean effects.

Bias – A term denoting that a known or unknown variable is or may be responsible for an observed effect other than the intervention.

Blinding – This denotes that the researcher is masked or 'blinded' to the identity of the group allocation of the participants when undertaking post-tests. This prevents biased assessment. Sometimes participants are also blinded to the true nature of the experiment.

Blocked randomisation – This method of randomisation prevents groups becoming either numerically unbalanced or suffering from chance bias. It does this by randomising in blocks (e.g., block of four). Thus, a block of four can be: ABAB, AABB, BBAA, BABA, ABBA, BAAB. This means that the study will be balanced, although the block size must be kept secret to conceal the allocation sequence.

Case control study – A study where participants are identified with a specific outcome (cases) and then compared with a control group of participants without the outcome.

Campbell Collaboration – Inspired by the Cochrane Collaboration (see below) but aims to synthesise controlled studies in education, crime and justice and social welfare (www.campbellcollaboration.org/).

Cochrane Collaboration – A world-wide collaboration, the aim of which is to collect and review all of the controlled trials in the health care field, to inform clinicians and policy-makers (www.cochrane.org/).

Comprehensive cohort design – A study design whereby participants who do not consent to be randomised, or cannot be randomised, are followed up alongside the randomised groups.

Confidence intervals – A method of expressing sample uncertainty around the estimate of treatment effect. They are usually 95 per cent intervals: if an identical trial is conducted many times then 95 per cent of the trials will have confidence intervals which contain the true estimate of effect.

Confounders – A variable associated with cause and outcome; can mask a true relationship between another variable and outcome.

CONSORT – Consolidated Standards for Reporting Trials is a descriptive method adopted by many medical journals for publication of RCTs.

Cost-benefit analysis – An economic technique that measures both costs and benefits in monetary terms. If costs are lower than benefits then the intervention should be adopted.

Cost-effectiveness analysis – An economic method that measures costs in monetary terms but measures benefits in 'natural' units. When comparing two mutually exclusive alternatives, the intervention with the lowest cost-effectiveness ratio should be adopted.

Cost-utility analysis – A form of cost-effectiveness analysis where the outcomes are measured in units of utility. An intervention should be adopted if the cost-utility ratio is lower than a decision-maker's willingness to pay threshold.

Effect size – This is the difference between two groups described in standard deviation units (i.e., difference divided by the standard deviation), which is termed the effect size.

Factorial design – A trial design where two or more different interventions are evaluated using the same participant sample. Has the advantage that two trials for the price of one can be undertaken. The simplest 2 x 2 factorial design results in four different groups.

ITT analysis – intention to treat analysis – This is where all participants are analysed in their original randomised groups; it is the most robust analytical method.

Minimisation – A non-random method that can form comparator groups. Groups are formed in such a way as to ensure that they are balanced on known covariates. If undertaken properly minimisation is as effective, and often better, at eliminating selection bias as random allocation.

Multi-variate analysis – In an RCT most known and unknown variables affecting outcome will be balanced at baseline. Nevertheless, particularly in small studies, imbalance in prognostic variables can still affect the precision of the results. This is particularly the case with the pre-test variable, which will strongly predict outcome. A more precise estimate of the effect size (i.e., with smaller confidence intervals) can be obtained by undertaking a multivariate analysis with the pre-test score as a covariate as well as the group allocation.

Numbers needed to treat or teach (NNT) – This is a method of converting the effects of an intervention into an easily understood metric. Thus, a NNT of 5 means that five people need to be taught in order that one extra person passes an important threshold (e.g., an exam).

Observational data or study – Data generated from a non-randomised study where estimates of effectiveness are gathered by comparing people exposed to an intervention with those unexposed.

Paired randomisation – Participants are formed into matched pairs on the basis of important covariates (e.g., gender). Once the study group has been formed into pairs a random member of each pair is allocated to the intervention.

Pairwise randomisation – A method of allocating participants that ensures numerical balance within a centre but avoids the problem of predictability that occurs with blocked randomisation. Randomisation takes place only when two participants are eligible and then one is selected, at random, for the intervention.

Participant preference – A type of trial where preferences of participants are recorded and sometimes only participants with no preference are randomised.

Pilot study – A type of study that precedes the definitive trial; can be an internal or external pilot. Characteristics are: small sample size and or incompletely developed intervention.

Per-protocol analysis – Participants not complying with the treatment protocol are excluded from the analysis. Violates randomisation and can lead to bias.

Placebo – Commonly used in drug trials for the control treatment. The placebo is an inert substance that looks and tastes like the real drug and blinds or masks the participant, doctor and assessor as to the treatment group.

Power – Given a pre-specified hypothesised difference between intervention groups the power of a study relates to the chances of observing any difference between groups as being statistically significant if it exists. Power is commonly set at either 80 per cent or 90 per cent.

Preference trial – A trial design that takes participants' preferences into account by either asking them before randomisation (fully randomised preference design) or by only randomising those who do not have a preference and letting those with a preference have their preferred treatment.

Quasi-alternation – A biased method of constructing group membership that uses some characteristic of the participant, such as month of birth, first letter of surname to determine allocation.

Quasi-randomisation – Usually used to refer to alternation or other systematic methods of forming comparator groups, such allocating by date of birth.

Random sampling – A sampling method to allow an estimation of a parameter within a stated population. This allows generalisation of parameter estimates. Sometimes confused with randomisation.

RCT – Randomised controlled trial. This is where groups have been formed through random allocation (or a similar method). This is the main method that ensures that allocation bias is eliminated at baseline.

Regression analysis – A statistical method that is sometimes used on trial data to adjust for chance imbalances between two groups and to improve the precision of estimates of any treatment effect.

Regression discontinuity design – A quasi-experimental alternative to the RCT. This design selects people into their intervention groups on some pre-test variable with a pre-defined cut-off; if properly implemented this approach can produce unbiased estimates of effect sizes – albeit less efficiently than an RCT.

Resentful demoralisation – Participants who have a preference for an intervention and who are assigned to the opposite intervention may become demoralised and this may bias the trial's results.

Selection bias – This occurs when groups are formed by a process other than randomisation and important factors that are associated with outcome differ between the groups *before* they are exposed to the intervention.

Significance – This can be statistical, clinical, educational, economic. Statistical significance is usually 5 per cent ($p = 0.05$) or 10 per cent ($p = 0.10$) and relates to replication of a trial. Replication of an identical trial, where there is no treatment effect, many times will result in 5 per cent of the trials showing a difference as being statistically significant if the 5 per cent level is adopted. Other forms of significance relate to whether or not a difference between groups is worth having in terms of policy or practice.

Simple randomisation – This is the easiest form of randomisation akin to tossing a coin. A disadvantage with simple randomisation is that with small studies (<100) there is a high probability of having large chance imbalance between the groups. More importantly, there can be imbalance in important covariates. Restricted forms of randomisation are often used to prevent this.

Stratification – This is a process whereby randomisation is restricted (e.g., by blocking) such that any important known confounders are balanced between the groups.

Zelen's method – A trial design whereby participants are randomised *before* consent to take part in the study is obtained. The single consent method is where consent is only sought from those allocated to the novel intervention.

1
Background to Controlled Trials

1.1 Background

A key reason for undertaking any research is to increase certainty in an uncertain world. We all directly or indirectly consume research. We hope the treatment we are prescribed by our doctor will be effective in improving our condition. We want to know which educational interventions, curricular innovations and teaching methods are effective in increasing knowledge, skills and understanding. Policy-makers and practitioners are interested in the relative effectiveness of crime and justice interventions, for example rehabilitation programmes and sentencing policies.

Health and social science research can provide the knowledge that enables us to determine what does and does not work. The 'gold-standard' research method for addressing the 'what works?' question in 'evidence-informed' policy-making and practice is the randomised controlled trial (RCT).

The aims of this book, therefore, are: to introduce the RCT; to describe its methodology and design, focusing on when and how to undertake an RCT; to describe examples of high quality and weak application of the method; and to introduce critical appraisal of published RCTs. We do not include in the book detailed statistical justification for using the RCT or describe detailed statistical approaches for its analysis. Statistical theory and analysis are more than adequately covered by other authors (e.g., Altman, 1991; Bland, 2000). If the research design of an RCT is adequate and applied rigorously, then relatively simple statistical analysis is required. Even the most heroic form of statistical analysis cannot compensate for a poorly designed, poorly conducted trial. Consequently, it is the design aspect of a trial that is the most important issue relating to an RCT, and this is the focus of the book.

The randomised controlled trial (RCT) is a simple research method of elegant design. Two or more groups are formed through random allocation; one or more of the groups is exposed to an intervention (experimental group), while the other group(s) receive(s) an alternative treatment or no treatment (comparison or control group). The effects of the intervention are observed by comparing the outcomes of both groups. If the groups assembled through randomisation are sufficiently large, we can be confident that any differences observed between the groups will be a consequence of the intervention, rather than a result of some other known or unknown variable.

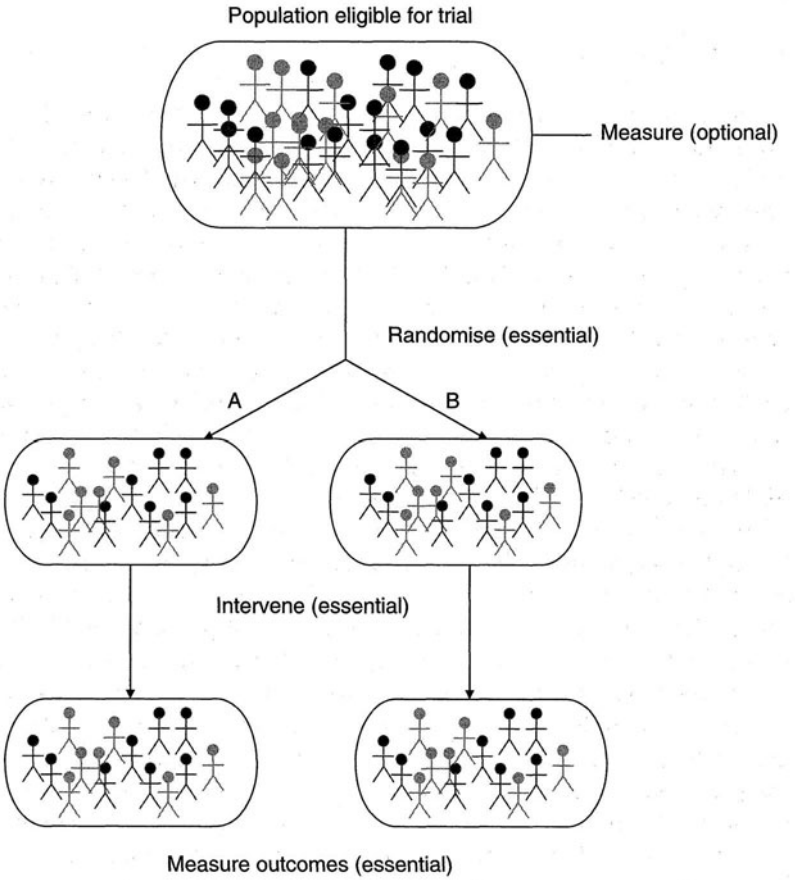

Figure 1.1: Schematic outline of a randomised trial

In Figure 1.1 we show the basic outline of the randomised trial. In essence the design is as follows: we assemble a population for whom the intervention is appropriate (this population may then be measured, although this step is not a pre-requisite); we then allocate the participants to two or more groups and apply the intervention(s) to the groups formed by randomisation; at some pre-specified time in the future we measure the groups in terms of their outcome – if there is a difference between the groups, and assuming that the difference and the sample size are sufficient, we can infer a causal relationship between our intervention and the group differences.

1.2 The randomised trial

Social interactions in the fields of health care, education, crime and justice, and other social sciences involve complex phenomena, including relationships between doctors and patients, teachers and students, social workers and clients. The best method for evaluating any proposed changes in health care, education, crime and justice, and other areas of public policy is the RCT, because it is able to deal adequately with the level of complexity inherent in these fields (Sheldon and Oakley, 2002) by 'teasing out' from the background 'noise' whether or not an intervention is actually effective.

The RCT has developed considerably since its inception in the middle of the last century. Whilst some disciplines use the method more often than others (health care research compared with, for example, educational research), the breathtaking simplicity of the design means, for questions of effectiveness, it could be used more often in place of other less rigorous evaluative approaches.

In this book we use a variety of examples to illustrate the design of trials. Due to historical reasons, many of these examples are from health care research; however, we include examples from other disciplines, in particular education and crime and justice. This is primarily because, as health care researchers have applied the RCT away from drug trials, they have had to grapple with numerous problems that threaten its internal validity, including how to deal with participants' strong preferences for a given treatment. Education and other social science researchers, whilst dealing with similar problems, have undertaken fewer RCTs in recent years. Health care trial research is funded more generously than in other areas, allowing more methodological research activity to take place in the design and use of the RCT. Nevertheless lessons from the design and conduct of trials in education and other social sciences (e.g., the

development of cluster or group randomised trials) have been enthusiastically adopted by health care researchers.

The RCT has long been recognised as the 'gold-standard' research method in health care research (Pocock, 1983), although this has not always been the case. Silverman (2004) described entrenched opposition he encountered from both clinicians and clinical researchers to the use of the method in the 1940s and 1950s. It is still sometimes argued, even in health care, that issues of effectiveness can be resolved through the use of other research methods, such as basic science, qualitative enquiry or through before and after approaches (Penston, 2007). However, the use of other methods to infer causality has led, and continues to lead, to the implementation of ineffective or harmful interventions.

1.3 Health care disasters

At this point it is worth noting some deadly examples from health care research of inappropriate implementation of interventions not previously having been adequately exposed to a randomised trial. Mistakes in health care research can be counted in mortality or morbidity, and this has led to the realisation that, morally and ethically, patients need to be protected from potentially hazardous new treatments by first evaluating the treatments in RCTs. In contrast, in other areas, for instance social welfare, any potentially hazardous effect of an intervention does not manifest itself with such direct or obvious consequences.

One of the earliest health care disasters involved administering oxygen to premature infants. In the 1940s and 1950s the incidence of blindness seemed to be increasing among premature babies. The cause of this was not discovered until an RCT evaluating the 'routine' practice of supplementing premature babies with oxygen showed that babies allocated to oxygen supplementation had significant increases in blindness, compared with un-supplemented infants (Silverman, 1977, 1997). Similarly, in the late 1940s and early 1950s, and on the basis of evidence from case reports, some premature babies were given prophylactic antibiotics. It was only later, during a randomised trial published in 1954, that this routine practice was shown to lead to brain damage and death in significantly more babies than those who had received an alternative treatment (Silverman and Altman, 1996).

One of the biggest catastrophes in terms of actual numbers of deaths was the routine use of anti-arrhythmia drugs for post-myocardial infarction patients. Many cardiologists were opposed to the use of RCTs on

Box 1.1: The CAST trial

From about 1978, hundreds of thousands of patients were given flecainide and other, similar anti-arrhythmia drugs. In 1987 an adequately powered trial was begun. In 1989 the trial was terminated abruptly due to *increased* deaths in the active treatment groups.

All cause death was 7.7 per cent in the treatment group compared with 3.0 per cent in the placebo group (relative risk of death of taking active treatment = 2.5, 95 per cent confidence interval 1.6 to 4.5) (Cardiac Arrhythmia Suppression Trial (CAST) Investigators, Preliminary Report, 1991).

It has been estimated that tens of thousands of people died as a result of uncontrolled use of these agents (Silverman, 1997).

ethical grounds. They believed that such drugs were beneficial and that to withhold them, therefore, would be unethical (see Box 1.1).

Trials were eventually started in the 1980s, but stopped early because of significantly increased mortality among patients allocated to the active treatment (CAST Investigators, 1991). Indeed, the trialists were so confident the anti-arrhythmia drugs would prove to be beneficial or, at worst, have no effect, that the trial was designed to enable early stoppage once an important benefit had been found. Instead, the interim analysis found that mortality was significantly elevated in the active treatment groups. As these drugs had not previously been evaluated using large RCTs, it has been estimated that tens of thousands of patients died through their unrestricted use in routine clinical practice (Silverman, 1997).

In another disaster, thousands of pregnant women were given a synthetic hormone to prevent miscarriage. Randomised trials later failed to show that this treatment – diethylistilboestrol (DES) – was effective. Unfortunately, it transpired that some female children whose pregnant mothers were exposed to DES later developed rare vaginal cancers and other serious health conditions (Oakley, 2000).

In 2004 a randomised controlled trial evaluating a 'standard' therapy for head injured patients (high dose steroids) was terminated half way through (CRASH Trial Collaborators, 2004). After recruiting half of the 20 000 participants across the world it was found that two-week mortality was significantly elevated among the steroid-treated patients. It has been estimated that the failure to evaluate this treatment promptly probably caused the deaths of more than 10 000 people (Sauerland and

Maegele, 2004). In the field of head injuries alone, several 'standard' treatments still remain unevaluated (CRASH Trial Collaborators, 2004).

A systematic review and meta-analysis of all the trials of antioxidants (e.g., betacarotene, vitamins A, C, E and selenium) showed that vitamins A and E and betacarotene supplementation can actually *increase* mortality (Bjelakovic et al., 2007). For vitamin C and selenium supplementation no evidence was found for harm or benefit.

Medical regulators now require evidence from properly conducted randomised controlled trials before drug treatments are implemented. Non-tested older drug treatments, vitamins, herbal supplements, 'natural' remedies and many non-pharmaceutical treatments (e.g., novel surgical therapies) can still be given to patients.

It is important to note that clinicians who have used harmful interventions probably did so with the best of intentions. Silverman, himself an early advocate of the use of RCTs in paediatric medicine, describes the case of a premature infant under his care. On the basis of data derived from animal experiments, he gave the child a drug treatment to prevent blindness only to discover later (from an RCT) that this drug increased mortality among infants and did not prevent blindness (Silverman, 2004).

Whilst new treatments are often seen as better than old interventions or 'standard care' this may not be necessarily true. In a review of RCTs comparing the efficacy of new drugs for childhood cancers with usual care it was found that in around half the trials the new drug was superior and in half it was inferior (Kumar et al., 2005). It is important, therefore, that all novel interventions are tested in rigorous RCTs.

1.4 Social science trials

Education and other social science trialists are able to point to fewer clear examples of adverse effects accruing through the lack of an RCT, to counter the arguments of those who oppose the wider use of trials. Despite this, however, these areas are not without their equivalent of anti-arrhythmia disasters and one is the 'Scared Straight' programme.

'Scared Straight' is a widely used intervention in North America. Juvenile offenders are taken to meet long-term prisoners in order to deter them from further crime. A recent version being offered in the UK is to take juvenile drug users to prisons to meet jailed drug offenders. A series of RCTs from North America was undertaken and summarised in a systematic review. The review demonstrated that the 'Scared Straight' programme actually *increased* the risk of offending in the juveniles in the

Box 1.2: Scared straight (Petrosino et al., 2002)

This initiative originated in the USA in the 1970s. The aim was to take juvenile offenders and expose them to presentations from prisoners serving life sentences, in order to deter them from further offending behaviour. Uncontrolled evaluations (i.e., before and after studies) suggested it had had a 94 per cent success rate of preventing juveniles from recidivism and the programme was widely implemented in the USA. Similar programmes have been used in the UK, Australia, Norway and Canada.

A systematic review of the randomised trials of Scared Straight found that all but one indicated a harmful effect of the programme and *increased* offending among participants. A meta-analysis of the trials showed that the odds of offending were 1.68 (95 per cent CI 1.20 to 2.36) for juveniles allocated to the intervention group.

intervention group compared with juveniles in the control group (Petrosino et al., 2002) (see Box 1.2).

Similar examples include a trial, undertaken in the UK, testing the effectiveness of social work supervision of school truants, which showed an increase in the risk of truancy compared with no supervision (Berg et al., 1978). In the USA a trial was undertaken to look at the use of routine arrests for people who were suspected of intimate partner abuse (Hirschel et al., 1992). Contrary to expectations, arrests did not lead to a reduction in future partner abuse.

An interesting example is the use of driver education among older school children to reduce vehicle accidents among young drivers on the basis of survey evidence. However, a systematic review of RCTs showed that, contrary to expectations, driver education programmes led to an *increase* in young driver deaths and road accidents (Cochrane Injuries Group Driver Education Reviewers, 2001).

1.5 Conclusions

The randomised controlled trial is the most effective method of assessing causality. Other approaches can give misleading results, and there are many examples, particularly from health care research, where practitioners and policy-makers have implemented ineffective or harmful interventions on the basis of evidence derived from non-randomised study designs. Whilst the basic format of the randomised trial (Figure 1.1) is