Research Methods in Education

Fifth edition

Louis Cohen, Lawrence Manion
and Keith Morrison

London and New York
Experiments, quasi-experiments and single-case research

Introduction

The issue of causality and, hence, predictability has exercised the minds of researchers considerably (Smith, 1991: 177). One response to the problem has been in qualitative research that defines causality in the terms of the participants (Chapter 6). Another response has been in the operation of control, and it finds its apotheosis in the experimental design. If rival causes or explanations can be eliminated from a study, then, it is argued, clear causality can be established, the model can explain outcomes. Smith (1991: 177) claims the high ground for the experimental approach, arguing that it is the only method that directly concerns itself with causality; this, clearly, is contestable, as we show in Chapters 6–9, and 13 of this book.

In Chapter 11, we described ex post facto research as experimentation in reverse in that ex post facto studies start with groups that are already different with regard to certain characteristics and then proceed to search, in retrospect, for the factors that brought about those differences. We then went on to cite Kerlinger’s description of the experimental researcher’s approach:

If \( x \), then \( y \); if frustration, then aggression … the researcher uses some method to measure \( x \) and then observes \( y \) to see if concomitant variation occurs.

(Kerlinger, 1970)

The essential feature of experimental research is that investigators deliberately control and manipulate the conditions which determine the events in which they are interested. At its simplest, an experiment involves making a change in the value of one variable — called the independent variable — and observing the effect of that change on another variable — called the dependent variable.

Imagine that we have been transported to a laboratory to investigate the properties of a new wonder-fertilizer that farmers could use on their cereal crops, let us say wheat (Morrison, 1993: 44–5). The scientist would take the bag of wheat seed and randomly split it into two equal parts. One part would be grown under normal existing conditions — controlled and measured amounts of soil, warmth, water and light and no other factors. This would be called the control group. The other part would be grown under the same conditions — the same controlled and measured amounts of soil, warmth, water and light as the control group, but, additionally, the new wonder-fertilizer. Then, four months later, the two groups are examined and their growth measured. The control group has grown half a metre and each ear of wheat is in place but the seeds are small. The experimental group, by contrast, has grown half a metre as well but has significantly more seeds on each ear, the seeds are larger, fuller and more robust.

The scientist concludes that, because both groups came into contact with nothing other than measured amounts of soil, warmth, water and light, then it could not have been anything else but the new wonder-fertilizer that caused the experimental group to flourish so well. The key factors in the experiment were:

- the random allocation of the whole bag of wheat into two matched groups (the control and the experimental group), involving the
initial measurement of the size of the wheat to ensure that it was the same for both groups (i.e. the pretest);
- the identification of key variables (soil, warmth, water, and light);
- the control of the key variables (the same amounts to each group);
- the exclusion of any other variables;
- the giving of the special treatment (the intervention) to the experimental group whilst holding every other variable constant for the two groups;
- the final measurement of yield and growth (the post-test);
- the comparison of one group with another;
- the stage of generalization – that this new wonder-fertilizer improves yield and growth under a given set of conditions.

This model, premised on notions of isolation and control of variables in order to establish causality, may be appropriate for a laboratory, though whether, in fact a social situation either ever could become the antiseptic, artificial world of the laboratory or should become such a world is both an empirical and a moral question respectively. Further, the ethical dilemmas of treating humans as manipulable, controllable and inanimate are considerable (see Chapter 2). However, let us pursue the experimental model further.

Frequently in learning experiments in classroom settings the independent variable is a stimulus of some kind, a new method in arithmetical computation for example, and the dependent variable is a response, the time taken to do twenty problems using the new method. Most empirical studies in educational settings, however, are quasi-experimental rather than experimental. The single most important difference between the quasi-experiment and the true experiment is that in the former case, the researcher undertakes his study with groups that are intact, that is to say, the groups have been constituted by means other than random selection. We begin by identifying the essential features of pre-experimental, true experimental and quasi-experimental designs, our intention being to introduce the reader to the meaning and purpose of control in educational experimentation.

**Designs in educational experimentation**

In the outline of research designs that follows we use symbols and conventions from Campbell and Stanley (1963):

1. $X$ represents the exposure of a group to an experimental variable or event, the effects of which are to be measured.
2. $O$ refers to the process of observation or measurement.
3. $X$s and $O$s in a given row are applied to the same persons.
4. Left to right order indicates temporal sequence.
5. $X$s and $O$s vertical to one another are simultaneous.
6. $R$ indicates random assignment to separate treatment groups.
7. Parallel rows unseparated by dashes represent comparison groups equated by randomization, while those separated by a dashed line represent groups not equated by random assignment.

**A pre-experimental design: the one group pretest–post-test**

Very often, reports about the value of a new teaching method or interest aroused by some curriculum innovation or other reveal that a researcher has measured a group on a dependent variable ($O_1$), for example, attitudes towards minority groups, and then introduced an experimental manipulation ($X$), perhaps a ten-week curriculum project designed to increase tolerance of ethnic minorities. Following the experimental treatment, the researcher has again measured group attitudes ($O_2$) and proceeded to account for differences between pretest and post-test scores by reference to the effects of $X$.

The one group pretest–post-test design can be represented as:
Experimental $O_1 \ X \ O_2$

Suppose that just such a project has been undertaken and that the researcher finds that $O_2$ scores indicate greater tolerance of ethnic minorities than $O_1$ scores. How justified is she in attributing the cause of $O_1$–$O_2$ differences to the experimental treatment ($X$), that is, the term's project work? At first glance the assumption of causality seems reasonable enough. The situation is not that simple, however. Compare for a moment the circumstances represented in our hypothetical educational example with those which typically obtain in experiments in the physical sciences. A physicist who applies heat to a metal bar can confidently attribute the observed expansion to the rise in temperature that she has introduced because within the confines of her laboratory she has excluded (i.e. controlled) all other extraneous sources of variation (this example is suggested by Pilliner, 1973).

The same degree of control can never be attained in educational experimentation. At this point readers may care to reflect upon some possible influences other than the ten-week curriculum project that might account for the $O_1$–$O_2$ differences in our hypothetical educational example.

They may conclude that factors to do with the pupils, the teacher, the school, the classroom organization, the curriculum materials and their presentation, the way that the subjects' attitudes were measured, to say nothing of the thousand and one other events that occurred in and about the school during the course of the term's work, might all have exerted some influence upon the observed differences in attitude. These kinds of extraneous variables which are outside the experimenters' control in one-group pretest–posttest designs threaten to invalidate their research efforts. We identify a number of such threats to the validity of educational experimentation in Chapter 5.

A 'true' experimental design: the pretest–posttest control group design

A complete exposition of experimental designs is beyond the scope of this chapter. In the brief outline that follows, we have selected one design from the comprehensive treatment of the subject by Campbell and Stanley (1963) in order to identify the essential features of what they term a 'true experimental' and what Kerlinger (1970) refers to as a 'good' design. Along with its variants, the chosen design is commonly used in educational experimentation.

The pretest–posttest control group design can be represented as:

Experimental | RO_1 X O_2
Control

It differs from the pre-experimental design that we have just described in that it involves the use of two groups which have been constituted by randomization. As Kerlinger observes, in theory, random assignment to E and C conditions controls all possible independent variables. In practice, of course, it is only when enough subjects are included in the experiment that the principle of randomization has a chance to operate as a powerful control. However, the effects of randomization even with a small number of subjects is well illustrated in Box 12.1.

Randomization, then, ensures the greater likelihood of equivalence, that is, the apportioning out between the experimental and control groups of any other factors or characteristics of the subjects which might conceivably affect the experimental variables in which the researcher is interested. It is, as Kerlinger (1970) notes, the addition of the control group in our present example and the random assignment of subjects to E and C groups that radically alters the situation from that which obtains in the pre-experimental design outlined earlier. For if the groups are made equivalent, then any so-called 'cloning' effects should be present in both groups.
If the mental ages of the children of the experimental group increase, so should the mental ages of the children of the control group... If something happens to affect the experimental subjects between the pretest and the post-test, this something should also affect the subjects of the control groups.

(Kerlinger, 1970)

So strong is this simple and elegant true experimental design, that all the threats to internal validity identified by Campbell and Stanley (1963) are controlled in the pretest-post-test control group design.

One problem that has been identified with this particular experimental design is the interaction effect of testing. Good (1963) explains that whereas the various threats to the validity of the experiments listed in Chapter 5 can be thought of as main effects, manifesting themselves in mean differences independently of the presence of other variables, interaction effects, as their name implies, are joint effects and may occur even when no main effects are present. For example, an interaction effect may occur as a result of the pretest measure sensitizing the subjects to the experimental variable. Interaction effects can be controlled for by adding to the pretest-post-test control group design two more groups that do not experience the pretest measures. The result is a four-group design, as suggested by Solomon. Later in the chapter, we describe an educational study which built into a pretest-post-test group design a further control group to take account of the possibility of pretest sensitization.

A quasi-experimental design: the non-equivalent control group design

Often in educational research, it is simply not possible for investigators to undertake true experiments. At best, they may be able to employ something approaching a true experimental design in which they have control over what Campbell and Stanley (1963) refer to as 'the who and to whom of measurement' but lacks control over 'the when and to whom of exposure', or the randomization of exposures – essential if true experimentation is to take place. These situations are quasi-experimental and the methodologies employed by researchers are termed quasi-experimental designs. (Kerlinger (1970) refers to quasi-experimental situations as 'compromise designs', an apt description when applied to much educational research where the random selection or random assignment of schools and classrooms is quite impracticable.)

One of the most commonly used quasi-experimental designs in educational research can be represented as:

\[
\begin{array}{cccc}
O_1 & X & O_2 & O_3 & O_4 \\
\end{array}
\]

Experimental Control
The dashed line separating the parallel rows in the diagram of the non-equivalent control group indicates that the experimental and control groups have not been equated by randomization – hence the term ‘non-equivalent’. The addition of a control group makes the present design a decided improvement over the one group pretest-posttest design, for to the degree that experimenters can make E and C groups as equivalent as possible, they can avoid the equivocality of interpretations that plague the pre-experimental design discussed earlier. The equivalence of groups can be strengthened by matching, followed by random assignment to E and C treatments.

Where matching is not possible, the researcher is advised to use samples from the same population or samples that are as alike as possible (Kerlinger, 1970). Where intact groups differ substantially, however, matching is unsatisfactory due to regression effects which lead to different group means on post-test measures. Campbell and Stanley put it this way:

If [in the non-equivalent control group design] the means of the groups are substantially different, then the process of matching not only fails to provide the intended equation but in addition insures the occurrence of unwanted regression effects. It becomes predictably certain that the two groups will differ on their post-test scores altogether independently of any effects of X, and that this difference will vary directly with the difference between the total populations from which the selection was made and inversely with the test–retest correlation. (Campbell and Stanley, 1963)

Procedures in conducting experimental research

In Chapter 11, we identified a sequence of steps in carrying out an ex post facto study. An experimental investigation must also follow a set of logical procedures. Those that we now enumerate, however, should be treated with some circumspection. It is extraordinarily difficult (and indeed, foolhardy) to lay down clear-cut rules as guides to experimental research. At best, we can identify an ideal route to be followed, knowing full well that educational research rarely proceeds in such a systematic fashion. (For a detailed discussion of the practical issues in educational experimentation, see Evans (1978), Chapter 4, ‘Planning experimental work’, Riecken and Boruch (1974), and Bennett and Lumsdale (1975).)

First, the researcher must identify and define the research problem as precisely as possible, always supposing that the problem is amenable to experimental methods.

Second, she must formulate hypotheses that she wishes to test. This involves making predictions about relationships between specific variables and at the same time making decisions about other variables that are to be excluded from the experiment by means of controls. Variables, remember, must have two properties. First, they must be measurable. Physical fitness, for example, is not directly measurable until it has been operationally defined. Making the variable ‘physical fitness’ operational means simply defining it by letting something else that is measurable stand for it – a gymnastics test, perhaps. Second, the proxy variable must be a valid indicator of the hypothetical variable in which one is interested. That is to say, a gymnastics test probably is a reasonable proxy for physical fitness; height on the other hand most certainly is not.

Third, the researcher must select appropriate levels at which to test the independent variables in order for differences to be observed. The experimenter will vary the stimuli at such levels as are of practical interest in the real-life situation. For example comparing reading periods of forty-four minutes, or forty-six minutes, with timetabled reading lessons of forty-five minutes is scarcely likely to result in observable differences in attainment.

Fourth, in planning the design of the experiment, the researcher must take account of the population to which she wishes to generalize her results. This involves her in decisions over sample sizes and sampling methods.

Fifth, with problems of validity in mind, the
researcher must select instruments, choose tests and decide upon appropriate methods of analysis.

Sixth, before embarking upon the actual experiment, the researcher must pilot test the experimental procedures to identify possible snags in connection with any aspect of the investigation (Simon, 1978).

Seventh, during the experiment itself, the researcher must endeavour to follow tested and agreed-on procedures to the letter. The standardization of instructions, the exact timing of experimental sequences, the meticulous recording and checking of observations – these are the hallmark of the competent researcher.

With her data collected, the researcher faces the most important part of the whole enterprise. Processing data, analysing results and drafting reports are all extremely demanding activities, both in intellectual effort and time.

Borg and Gall (1979: 547) set out a useful series of steps in the planning and conduct of an experiment:

**Step 1** Carry out a measure of the dependent variable.

**Step 2** Assign participants to matched pairs, based on the scores and measures established from Step 1.

**Step 3** Randomly assign one person from each pair to the control group and the other to the experimental group.

**Step 4** Administer the experimental treatment/intervention to the experimental group and, if appropriate, a placebo to the control group. Ensure that the control group is not subject to the intervention.

**Step 5** Carry out a measure of the dependent variable with both groups and compare/measure them in order to determine the effect and its size on the dependent variable.

Borg and Gall indicate that difficulties arise in the close matching of the sample of the control and experimental groups. This involves careful identification of the variables on which the matching must take place. They suggest (p. 547) that matching on a number of variables that correlate with the dependent variable is more likely to reduce errors than matching on a single variable. The problem, of course, is that the greater the number of variables that have to be matched, the harder it is actually to find the sample of people who are matched. Hence the balance must be struck between having too few variables such that error can occur, and having so many variables that it is impossible to draw a sample.

Further, the authors draw attention to the need to specify the degree of exactitude (or variance) of the match. For example, if the subjects were to be matched on, say, linguistic ability as measured in a standardized test, it is important to define the limits of variability that will be used to define the matching (e.g. ± 3 points). As before, the greater the degree of precision in the matching here, the closer will be the match, but the greater the degree of precision the harder it will be to find an exactly matched sample.

One way of addressing this issue is to place all the subjects in rank order on the basis of the scores or measures of the dependent variable. Then the first two subjects become one matched pair (which one is allocated to the control group and which to the experimental group is done randomly, e.g. by tossing a coin), subjects three and four become the next matched pair, subjects five and six become the next matched pair, and so on until the sample is drawn. Here the loss of precision is counterbalanced by the avoidance of the loss of subjects.

The alternative to matching that has been discussed earlier in the chapter is randomization. Smith (1991: 215) suggests that matching is most widely used in quasi-experimental and non-experimental research, and is a far inferior means of ruling out alternative causal explanations than randomization. Randomization, he argues, produces equivalence over a whole range of variables, whereas matching produces equivalence over only a few named variables. The use of randomized controlled trials (RCTs), a method used in medicine, is a putative way of establishing causality and generalizability (though, in medicine, the sample sizes for some RCTs are necessarily so small – there being limited sufferers from a particular complaint – that randomization is seriously compromised).
A powerful advocacy of RCTs for planning and evaluation is provided by Boruch (1997). Indeed he argues (p. 69) that the problem of poor experimental controls has led to highly questionable claims being made about the success of programmes. Examples of the use of RCTs can be seen in Maynard and Chalmers (1997).

Mitchell and Jolley (1988: 103) pose three important questions that researchers need to consider when comparing two groups:

- Are the two groups equal at the commencement of the experiment?
- Would the two groups have grown apart naturally, regardless of the intervention?
- To what extent has initial measurement error of the two groups been a contributory factor in differences between scores?

**Examples from educational research**

**Example 1: a pre-experimental design**

A pre-experimental design was used in a study involving the 1991–2 Postgraduate Diploma in Education group following a course of training to equip them to teach social studies in senior secondary schools in Botswana. The researcher wished to find out whether the programme of studies he had devised would effect changes in the students’ orientations towards social studies teaching. To that end, he employed a research instrument, the Barth/Shermis Studies Preference Scale (BSSPS).

The BSSPS provides measures of what port to be three social studies traditions or philosophical orientations, the oldest of which, Citizenship Transmission, involves indoctrination of the young in the basic values of a society. The second orientation, called the Social Science, is held to relate to the acquisition of knowledge-gathering skills based on the mastery of social science concepts and processes. The third tradition, Reflective Inquiry, emphasizes the process of inquiry. Forty-eight Postgraduate Diploma students were administered the BSSPS during the first session of their one-year course of study. At the end of the programme, the BSSPS was again completed in order to determine whether changes had occurred in students’ philosophical orientations. Briefly, the ‘preferred orientation’ in the pretest and post-test was the criterion measure, the two orientations least preferred being ignored. Broadly speaking, students tended to move from a majority holding a Citizenship Transmission orientation at the beginning of the course to a greater affirmation of the Social Science and the Reflective Inquiry traditions. Using the symbols and conventions adopted earlier to represent research designs, we can illustrate the Botswana study as:

```
  Experimental
   O₁  X  O₂
```

The briefest consideration reveals inadequacies in the design. Indeed, Campbell and Stanley describe the one group pretest–post-test design as ‘a “bad example” to illustrate several of the confounded extraneous variables that can jeopardize internal validity. These variables offer plausible hypotheses explaining an $O₁$–$O₂$ difference, rival to the hypothesis that caused the difference’ (Campbell and Stanley, 1963). The investigator is rightly cautious in his conclusions: ‘it is possible to say that the social studies course might be responsible for this phenomenon, although other extraneous variables might be operating’ (Adeyemi, 1992, emphasis added). Somewhat ingenuously he puts his finger on one potential explanation, that the changes could have occurred among his intending teachers because the shift from ‘inculcation to rational decision-making was in line with the recommendation of the Nine Year Social Studies Syllabus issued by the Botswana Ministry of Education in 1989’ (Adeyemi, 1992).

**Example 2: a quasi-experimental design**

Mason, Mason and Quayle’s longitudinal study took place between 1984 and 1992. Its principal aim was to test whether the explicit teaching of linguistic features of GCSE textbooks, coursework and examinations would produce an improvement in performance
across the secondary curriculum. The design adopted in the study may be represented as:

```
Experimental
Control
\[ \begin{array}{c}
O_1 \quad X \quad O_2 \\
O_1 \quad \quad O_2
\end{array} \]
```

This is, of course, the non-equivalent control group design outlined earlier in this chapter in which parallel rows separated by dashed lines represent groups that have not been equated by random assignment.

In brief, the researchers adopted a methodology akin to teaching English as a foreign language and applied this to Years 7–9 in the study school and two neighbouring schools, monitoring the pupils at every stage and comparing their performance with control groups drawn both from the three schools. Inevitably, because experimental and control groups were not randomly allocated, there were significant differences in the performance of some groups on pre-treatment measures such as the York Language Aptitude Test. Moreover, because no standardized reading tests of sufficient difficulty were available as post-treatment measures, tests had to be devised by the researchers, who provide no details as to their validity or reliability. These difficulties notwithstanding, pupils in the experimental groups taking public examinations in 1990 and 1991 showed substantial gains in respect of the percentage increases of those obtaining GCSE Grades A–C. The researchers note that during the three years 1989 to 1991, ‘no other significant change in the policy, teaching staff or organization of the school took place which could account for this dramatic improvement of 50 per cent’ (Mason et al., 1992).

Although the researchers attempted to control extraneous variables, readers may well ask whether threats to internal and external validity (see Chapter 5) were sufficiently met as to allow such a categorical conclusion as, ‘the pupils . . . achieved greater success in public examinations as a result of taking part in the project’ (Mason et al., 1992).

**Example 3: a ‘true’ experimental design**

Another investigation (Bhadwal and Panda, 1991) concerned with effecting improvements in students’ performance as a consequence of changing teaching strategies used a more robust experimental design. In rural India, the researchers drew a sample of seventy-eight pupils, matched by socio-economic backgrounds and non-verbal IQs, from three primary schools that were themselves matched by location, physical facilities, teachers’ qualifications and skills, school evaluation procedures and degree of parental involvement. Twenty-six pupils were randomly selected to comprise the experimental group, the remaining fifty-two being equally divided into two control groups. Before the introduction of the changed teaching strategies to the experimental group, all three groups completed questionnaires on their study habits and attitudes. These instruments were specifically designed for use with younger children and were subjected to the usual item analyses, test–retest and split-half reliability inspections. Bhadwal and Panda’s research design can be represented as:

```
Experimental
First control
Second control
\[ \begin{array}{c}
RO_1 \quad X \quad RO_2 \\
RO_1 \quad RO_4 \\
RO_3 \quad RO_5
\end{array} \]
```

Recalling Kerlinger’s discussion of a ‘good’ experimental design, the version of the pretest–post-test control design employed here (unlike the design used in Example 2 above) resorted to randomization which, in theory, controls all possible independent variables. Kerlinger adds, however, ‘in practice, it is only when enough subjects are included in the experiment that the principle of randomization has a chance to operate as a powerful control’ (Kerlinger, 1970). It is doubtful whether twenty-six pupils in each of the three groups in Bhadwal and Panda’s study constituted ‘enough subjects’.

In addition to the matching procedures in drawing up the sample, and the random allocation of pupils to experimental and control
groups, the researchers also used analysis of covariance, as a further means of controlling for initial differences between E and C groups on their pretest mean scores on the independent variables, study habits and attitudes.

The experimental programme involved improving teaching skills, classroom organisation, teaching aids, pupil participation, remedial help, peer-tutoring, and continuous evaluation. In addition, provision was also made in the experimental group for ensuring parental involvement and extra reading materials. It would be startling if such a package of teaching aids and curriculum strategies did not effect significant changes in their recipients and such was the case in the experimental results. The Experimental Group made highly significant gains in respect of its level of study habits as compared with Control Group 2 where students did not show a marked change. What did surprise the investigators, we suspect, was the significant increase in levels of study habits in Control Group 1. Maybe, they opine, this unexpected result occurred because Control Group 1 pupils were tested immediately prior to the beginning of their annual examinations. On the other hand, they concede, some unaccountable variables might have been operating. There is, surely, a lesson here for all researchers!

**Single-case research: ABAB design**

Increasingly, in recent years, single-case research as an experimental methodology has extended to such diverse fields as clinical psychology, medicine, education, social work, psychiatry, and counselling. Most of the single-case studies carried out in these (and other) areas share the following characteristics:

- they involve the continuous assessment of some aspect of human behaviour over a period of time, requiring on the part of the researcher the administration of measures on multiple occasions within separate phases of a study.

- they involve 'intervention effects' which are replicated in the same subject(s) over time.

Continuous assessment measures are used as a basis for drawing inferences about the effectiveness of intervention procedures.

The characteristics of single-case research studies are discussed by Kazdin (1982) in terms of ABAB designs, the basic experimental format in most single-case researches. ABAB designs, Kazdin observes, consist of a family of procedures in which observations of performance are made over time for a given client or group of clients. Over the course of the investigation, changes are made in the experimental conditions to which the client is exposed. The basic rationale of the ABAB design is illustrated in Box 12.2. What it does is this. It examines the effects of an intervention by alternating the baseline condition (the A phase), when no intervention is in effect, with the intervention condition (the B phase). The A and B phases are then repeated to complete the four phases. As Kazdin says, the effects of the intervention are clear if performance improves during the first intervention phase, reverts to or approaches original baseline levels of performance when the treatment is withdrawn, and improves again when treatment is recommenced in the second intervention phase.

An example of the application of the ABAB design in an educational setting is provided by Dietz (1977)'s whose single-case study sought to measure the effect that a teacher could have upon the disruptive behaviour of an adolescent boy whose persistent talking disturbed his fellow classmates in a special education class.

In order to decrease the unwelcome behaviour, a reinforcement programme was devised in which the boy could earn extra time with the teacher by decreasing the number of times he called out. The boy was told that when he made three (or fewer) interruptions during any 55-minute class period the teacher would spend extra time working with him. In the technical language of behaviour modification theory, the pupil would receive reinforcing consequences.
Box 12.2
The ABAB design

The solid lines in each phase present the actual data. The dashed lines indicate the projection or predicted level of performance from the previous phase.

Source Adapted from Kazdin, 1982

when he was able to show a low rate of disruptive behaviour (in Box 12.3 this is referred to as ‘differential reinforcement of low rates’ or DRL).

- When the boy was able to desist from talking aloud on fewer than three occasions during any timetabled period, he was rewarded by the teacher spending fifteen minutes with him helping him with his learning tasks. The pattern of results displayed in Box 12.3 shows the considerable changes that occurred in the boy’s behaviour when the intervention procedures were carried out and the substantial increases in disruptions towards baseline levels when the teacher’s rewarding strategies were withdrawn. Finally, when the intervention was reinstated, the boy’s behaviour is seen to improve again.

By way of conclusion, the single-case research design is uniquely able to provide an experimental technique for evaluating interventions for the individual subject. Moreover, such interventions can be directed towards the particular subject or group and replicated over time or across behaviours, situations, or persons. Single-case research offers an alternative strategy to the more usual methodologies based on between-group designs. There are, however, a number of problems that arise in connection with the use of single-case designs having to do with ambiguities introduced by trends and variations in baseline phase data and with the generality of results from single-case research. The interested reader is directed to Kazdin (1982), Borg (1981) and Vasta (1979).6

Meta-analysis in educational research

The study by Bhadwal and Panda (1991) is typical of research undertaken to explore the effectiveness of classroom methods. Often as not, such studies fail to reach the light of day, particularly when they form part of the research requirements for a higher degree. Meta-analysis is, simply, the analysis of other analyses. It involves aggregating the results of other studies into a coherent account. Among the advantages of using meta-analysis, Fitz-Gibbon cites the following:
Box 12.3
An ABAB design in an educational setting

![Graph showing an ABAB design with baseline, treatment, reversal, and treatment phases.]

Source: Kazdin, 1982

- Humble, small-scale reports which have simply been gathering dust may now become useful.
- Small-scale research conducted by individual students and lecturers will be valuable since meta-analysis provides a way of coordinating results drawn from many studies without having to coordinate the studies themselves.
- For historians, a whole new genre of studies is created – the study of how effect sizes vary over time, relating this to historical changes.

(Fitz-Gibbon, 1985: 46)

McGaw (1997: 371) suggests that quantitative meta-analysis replaces intuition, which is frequently reported narratively (Wood, 1995: 389), as a means of synthesizing different research studies transparently and explicitly (a desideratum in many synthetic studies (Jackson, 1980)), particularly when they differ very substantially. Narrative reviews, suggest Jackson (1980), Cook et al. (1992: 13) and Wood (1995: 390) are prone to:

- lack comprehensiveness, being selective and only going to subsets of studies;
- misrepresentation and crude representation of research findings;
- over-reliance on significance tests as a means of supporting hypotheses, thereby overlooking the point that sample size exerts a major effect on significance levels, and overlooking effect size;
- reviewers' failure to recognize that random sampling error can play a part in creating variations in findings amongst studies;
- overlook differing and conflicting research findings;
- reviewers' failure to examine critically the evidence, methods and conclusions of previous reviews;
- overlook the extent to which findings from research are mediated by the characteristics of the sample;
- overlook the importance of intervening variables in research;
- unreplicability because the procedures for integrating the research findings have not been made explicit.

Over the past few years a quantitative method for synthesizing research results has been developed by Glass et al. (1978; 1981) and others (e.g. Hedges and Olkin, 1985; Hedges, 1990;
Rosenthal, 1991) to supersede narrative intuition. Meta-analysis, essentially the 'analysis of analysis', is a means of quantitatively (a) identifying generalizations from a range of separate and disparate studies, and (b) discovering inadequacies in existing research such that new emphases for future research can be proposed. It is simple to use and easy to understand, though the statistical treatment that underpins it is somewhat complex. It involves the quantification and synthesis of findings from separate studies on some common measure, usually an aggregate of effect size estimates, together with an analysis of the relationship between effect size and other features of the studies being synthesized. Statistical treatments are applied to attenuate the effects of other contaminating factors, e.g. sampling error, measurement errors, and range restriction. Research findings are coded into substantive categories for generalizations to be made (Glass et al., 1981), such that consistency of findings is discovered that, through the traditional means of intuition and narrative review, would have been missed.

Fitz-Gibbon (1985: 45) explains the technique by suggesting that in meta-analysis the effects of variables are examined in terms of their effect size, that is to say, in terms of how much difference they make rather than only in terms of whether or not the effects are statistically significant at some arbitrary level such as 5 percent. Because, with effect sizes, it becomes easier to concentrate on the educational significance of a finding rather than trying to assess its importance by its statistical significance, we may finally see statistical significance kept in its place as just one of many possible threats to internal validity. The move towards elevating effect size over significance levels is hugely important (see also Chapter 10), and signals an emphasis on 'fitness for purpose' (the size of the effect having to be suitable for the researcher's purposes) over arbitrary cut-off points in significance levels as determinants of utility.

The term 'meta-analysis' originated in 1976 (Glass, 1976) and early forms of meta-analysis used calculations of combined probabilities and frequencies with which results fell into defined categories (e.g. statistically significant at given levels), though problems of different sample sizes confounded rigour (e.g. large samples would yield significance in trivial effects, whilst important data from small samples would not be discovered because they failed to reach statistical significance) (Light and Smith, 1971; Glass et al., 1981; McGaw, 1997: 371). Glass (1976) and Glass et al. (1981) suggested three levels of analysis: (a) primary analysis of the data; (b) secondary analysis, a re-analysis using different statistics; (c) meta-analysis analysing results of several studies statistically in order to integrate the findings. Glass et al. (1981) and Hunter et al. (1982) suggest several stages in the procedure:

**Step 1** Identify the variables for focus (independent and dependent).

**Step 2** Identify all the studies which feature the variables in which the researcher is interested.

**Step 3** Code each study for those characteristics that might be predictors of outcomes and effect sizes. (e.g. age of participants, gender, ethnicity, duration of the intervention).

**Step 4** Estimate the effect sizes through calculation for each pair of variables (dependent and independent variable) (see Glass, 1977), weighting the effect size by the sample size.

**Step 5** Calculate the mean and the standard deviation of effect sizes across the studies, i.e. the variance across the studies.

**Step 6** Determine the effects of sampling errors, measurement errors and range of restriction.

**Step 7** If a large proportion of the variance is attributable to the issues in Step 6, then the average effect size can be considered an accurate estimate of relationships between variables.

**Step 8** If a large proportion of the variance is not attributable to the issues in Step 6, then review those characteristics of interest which correlate with the study effects.

Cook et al. (1992: 7–12) set out a five stage model for an integrative review as a research process, covering:
• problem formulation (where a high quality meta-analysis must be rigorous in its attention to the design, conduct and analysis of the review);
• data collection (where sampling of studies for review has to demonstrate fitness for purpose);
• data retrieval and analysis (where threats to validity in non-experimental research – of which integrative review is an example – are addressed). Validity here must demonstrate fitness for purpose, reliability in coding, and attention to the methodological rigour of the original pieces of research;
• analysis and interpretation (where the accumulated findings of several pieces of research should be regarded as complex data points that have to be interpreted by meticulous statistical analysis).

Fitz-Gibbon (1984: 141–2) sets out four steps in conducting a meta-analysis:

Step 1 Finding studies (e.g. published, unpublished, reviews) from which effect sizes can be computed.
Step 2 Coding the study characteristics (e.g. date, publication status, design characteristics, quality of design, status of researcher).
Step 3 Measuring the effect sizes (e.g. locating the experimental group as a z-score in the control group distribution) so that outcomes can be measured on a common scale, controlling for ‘lumpy data’ (non-independent data from a large data set).
Step 4 Correlating effect sizes with context variables (e.g. to identify differences between well-controlled and poorly controlled studies).

Wood (1995: 393) suggests that effect-size can be calculated by dividing the significance level by the sample size. Glass et al. (1981: 29, 102) calculate the effect size as:

\[
\frac{X_{\text{of experimental group}} - X_{\text{of control group}}}{\text{Standard deviation of the control group}}
\]

Hedges (1981) and Hunter et al., (1982) suggest alternative equations to take account of differential weightings due to sample size variations.

The two most frequently used indices of effect sizes are standardized mean differences and correlations (ibid.: 373), though nonparametric statistics, e.g. the median, can be used. Lipsey (1992: 93–100) sets out a series of statistical tests for working on effect sizes, effect size means and homogeneity. It is clear from this that Glass and others assume that meta-analysis can only be undertaken for a particular kind of research – the experimental type – rather than for all types of research; this might limit its applicability.

Glass et al. (1981) suggest that meta-analysis is particularly useful when it uses unpublished dissertations, as these often contain weaker correlations than those reported in published research, and hence act as a brake on misleading, more spectacular generalizations. Meta-analysis, it is claimed (Cooper and Rosenthal, 1980), is a means of avoiding Type II errors (discussed in Chapter 5 – failing to find effects that really exist), synthesizing research findings more rigorously and systematically, and generating hypotheses for future research. However Hedges and Olkin (1980) and Cook et al. (1992: 297) show that Type II errors become more likely as the number of studies included in the sample increases.

Further, Rosenthal (1991) has indicated a method for avoiding Type I errors (finding an effect that, in fact, does not exist) that is based on establishing how many unpublished studies that average a null result would need to be undertaken to offset the group of published statistically significant studies. For one example he shows a ratio of 277:1 of unpublished to published research, thereby indicating the limited bias in published research.

Meta-analysis is not without its critics. Since so much depends upon the quality of the results that are to be synthesized, there is the danger that adherents may simply multiply the inadequacies of the data base and the limits of the sample (e.g. trying to compare the incomparable). Hunter et al. (1982) suggest that sampling error and the influence of other factors has to be addressed, and that it should account for less than 75 per cent of the variance in observed
effect sizes if the results are to be acceptable and able to be coded into categories. The issue is clear here: coding categories have to declare their level of precision, their reliability (e.g. intercoder reliability – the equivalent of inter-rater reliability, see Chapter 5) and validity (McGaw, 1997: 376–7).

To the charge that selection bias will be as strong in meta-analysis – which embraces both published and unpublished research – as in solely published research, Glass et al. (1981: 226–9) argue that it is necessary to counter gross claims made in published research with more cautious claims found in unpublished research.

Because the quantitative mode of (many) studies demands only a few common variables to be measured in each case, argues Tripp (1985),' cumulation of the studies tends to increase sample size much more than it increases the complexity of the data in terms of the number of variables. Meta-analysis risks attempting to synthesize studies which are insufficiently similar to each other to permit this with any legitimacy (Glass et al., 1981: 22; McGaw, 1997: 372) other than at an unhelpful level of generality. The analogy here might be to try to keep together oil and water as ‘liquids’; meta-analysts would argue that differences between studies and their relationships to findings can be coded and addressed in meta-analysis. Eysenck (1978) suggests that early meta-evaluation studies mixed apples with oranges! Though Glass et al. (1981: 218–20) refute this charge, it remains the case (McGaw, 1997) that there is a risk in meta-analysis of dealing indiscriminately with a large and sometimes incoherent body of research literature.

It is unclear, too, how meta-analysis differentiates between ‘good’ and ‘bad’ research – e.g. between methodologically rigorous and poorly constructed research (Cook et al., 1992: 297). Smith and Glass (1977) suggest that it is possible to use study findings, regardless of their methodological quality, though Glass and Smith (1978) and Slavin (1984a, 1984b), in a study of the effects of class size, indicate that methodological quality does make a difference. Glass et al. (1981: 220–6) effectively address the charge of using data from ‘poor’ studies, arguing, amongst other points, that many weak studies can add up to a strong conclusion (p. 221) and that the differences in the size of experimental effects between high-validity and low-validity studies are surprisingly small (p. 226).

Further, Wood (1995: 296) suggests that meta-analysis oversimplifies results by concentrating on overall effects to the neglect of the interaction of intervening variables. To the charge that, because meta-analyses are frequently conducted on large data sets where multiple results derive from the same study (i.e. that the data are non-independent) and are therefore unreliable, Glass et al. (1981) indicate how this can be addressed by using sophisticated data analysis techniques (pp. 153–216). Finally, a practical concern is the time required not only to use the easily discoverable studies (typically large-scale published studies) but to include the smaller-scale unpublished studies; the effect of neglecting the latter might be to build in bias in the meta-analysis.

It is the traditional pursuit of generalizations from each quantitative study which has most hampered the development of a data base adequate to reflect the complexity of the social nature of education. The cumulative effects of ‘good’ and ‘bad’ experimental studies is graphically illustrated in Box 12.4.

**An example of meta-analysis in educational research**

Glass and Smith (1978) and Glass et al. (1981: 35–44) identified seventy-seven empirical studies of the relationship between class size and pupil learning. These studies yielded 725 comparisons of the achievements of smaller and larger classes, the comparisons resting on data accumulated from nearly 900,000 pupils of all ages and aptitudes studying all manner of school subjects. Using regression analysis, the 725 comparisons were integrated into a single curve showing the relationship between class size and achievement in general. This curve revealed a definite inverse relationship between class size and pupil learning. When the
Box 12.4
Class size and learning in well-controlled and poorly-controlled studies

Regression lines for the regression of achievement (expressed in percentile ranks) onto class-size for studies that were well-controlled and poorly-controlled in the assignment of pupils to classes.

Source Adapted from Glass and Smith, 1978

researchers derived similar curves for a variety of circumstances that they hypothesized would alter the basic relationship (for example, grade level, subject taught, pupil ability etc.), virtually none of these special circumstances altered the basic relationship. Only one factor substantially affected the curve – whether the original study controlled adequately in the experimental sense for initial differences among pupils and teachers in smaller and larger classes. Adequate and inadequate control curves are set out in Box 12.4.