Learning lessons: Enhancing evaluation through research review

Prepared for Road Safety Division, Department of the Environment, Transport and the Regions

R Pawson and A Myhill (University of Leeds)
This report has been produced by TRL Limited, under/as part of a Contract placed by the Department of the Environment, Transport and the Regions. Any views expressed are not necessarily those of the Department.

TRL is committed to optimising energy efficiency, reducing waste and promoting recycling and re-use. In support of these environmental goals, this report has been printed on recycled paper, comprising 100% post-consumer waste, manufactured using a TCF (totally chlorine free) process.
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive Summary</td>
<td>1</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>3</td>
</tr>
<tr>
<td>2 Materials reviewed</td>
<td>4</td>
</tr>
<tr>
<td>3 The analysis - 21 lessons to enhance evaluation</td>
<td>6</td>
</tr>
<tr>
<td>3.1 Embeddedness</td>
<td>7</td>
</tr>
<tr>
<td>3.2 Programmes are theories</td>
<td>8</td>
</tr>
<tr>
<td>3.3 Process and outcomes</td>
<td>9</td>
</tr>
<tr>
<td>3.4 Experimental trials and tribulations</td>
<td>12</td>
</tr>
<tr>
<td>3.5 What is it about a programme which works, for whom and in what circumstances, and in what respects?</td>
<td>15</td>
</tr>
<tr>
<td>3.6 Learning the rules</td>
<td>16</td>
</tr>
<tr>
<td>3.7 Normative change: Peer-pressure and clubbing together</td>
<td>19</td>
</tr>
<tr>
<td>3.8 User-involvement and peer-research</td>
<td>21</td>
</tr>
<tr>
<td>3.9 Identifying risk</td>
<td>23</td>
</tr>
<tr>
<td>3.10 Targeting programmes</td>
<td>24</td>
</tr>
<tr>
<td>3.11 Age groups</td>
<td>25</td>
</tr>
<tr>
<td>3.12 Localities</td>
<td>28</td>
</tr>
<tr>
<td>3.13 What outcomes?</td>
<td>30</td>
</tr>
<tr>
<td>3.14 Hierarchy of outcomes</td>
<td>31</td>
</tr>
<tr>
<td>3.15 Approval ratings</td>
<td>32</td>
</tr>
<tr>
<td>3.16 Self-report</td>
<td>33</td>
</tr>
<tr>
<td>3.17 Educational gains</td>
<td>35</td>
</tr>
<tr>
<td>3.18 Simulations</td>
<td>36</td>
</tr>
<tr>
<td>3.19 Retrospective search for outcomes</td>
<td>37</td>
</tr>
<tr>
<td>3.20 Unintended outcomes</td>
<td>39</td>
</tr>
<tr>
<td>3.21 Meta-analysis and review</td>
<td>41</td>
</tr>
<tr>
<td>4 Conclusion</td>
<td>43</td>
</tr>
<tr>
<td>5 Acknowledgements</td>
<td>44</td>
</tr>
<tr>
<td>6 References</td>
<td>44</td>
</tr>
<tr>
<td>Abstract</td>
<td>49</td>
</tr>
<tr>
<td>Related publications</td>
<td>49</td>
</tr>
</tbody>
</table>
Children are making their way to the forefront of policy thinking. Clearly, they have always taken pride of place in educational provision and maintained their fair share of attention in the health fields. In recent years, however, specialist programmes for and about children have sprung up in every nook and cranny of the policy arena - from drugs-education to sex-abuse prevention to anti-smoking initiatives to anti-litter campaigns to fire-safety programmes to healthy-eating crusades to anti-bullying interventions to poison-prevention lessons to sun-safety procedures to road safety provision. Our review examines evaluations of programmes in all of these areas and more, with the aim of pulling lessons together on how best to conduct such research in the future. This summary presents a preview of what would normally be called our 'recommendations'. We have been struck, however, by just how much of our investigation is about lesson-learning - programmes face children with lessons to be learned, the success of which researchers have tried to glean, which findings the policy-maker has to interpret in order to inform subsequent programmes, which process should in time provide better lessons for children. It seems appropriate, therefore, to begin with a summary of our main lessons:

**Lesson 1.** Evaluation research should always begin with the recognition that programme efficacy depends on the ‘ideas’, ‘individuals’, ‘institutions’, and ‘infrastructure’ which make up the intervention. In any particular investigation, it is unlikely that all such features may be controlled and measured but efforts should always be made to observe, monitor and record basic information on all these features. It is particularly important that collective wisdom about the ‘four I’s’ should accumulate across investigations.

**Lesson 2.** Programmes are not things but conjectures, and the evaluator needs to test them as conjectures. The typical programme under review takes the form of a theory which speculates, ‘if we do X, we may persuade children to do Y’. The ideas embodied in a programme, however, may be met with acceptance, incomprehension, indifference or resistance on the part of children. A ‘theory-driven approach’ to evaluation is thus recommended, which considers the congruence of the conceptions of all concerned.

**Lesson 3.** Programme outcomes are the result of programme processes and evaluations are always enhanced to the extent that the study of one supports the understanding of the other. Programmes dependent on elaborate social process will always generate a complex footprint of outcome. Anticipating and understanding these patterns demands a partnership of ‘process’ and ‘outcome’ evaluation, and requires the use of both quantitative and qualitative methods.

**Lesson 4.** Randomised control trials are often considered the gold-standard method in evaluation since they offer great certainty about whether a particular programme has worked in a particular instance. But they throw up inconsistent results because they provide no understanding of the exchange of ideas within a programme nor of the context in which it takes place. They are not an absolute priority for the policy-maker.

**Lesson 5.** Policy-makers should desist from charging evaluation research with discovering ‘what works?’, for no initiative will work for all subjects in all circumstances. The really instructive evaluation question is ‘what is it about a programme which works for whom in what circumstances and in what respects?’.

**Lesson 6.** Programmes which aim to teach children sets of behavioural codes/rules/guidelines have to appreciate that children do not just ‘follow’ rules but adapt them to their existing understanding. These processes of assimilation are quite uneven because certain rules will clash with other expectations created during their upbringing. The researcher’s task is to uncover and explain the patterns of acceptance and resistance to rules.

**Lesson 7.** Programmes which aim to inform children via ‘peer-education’ and ‘club-activity’ have to anticipate wide variation in the ‘credibility of’ and a sense of ‘belonging to’ such manufactured social formations. Recipients will differ in the extent to which they find support and shared-identity. Researchers should expect considerable variability in the success of such networks and concentrate on discovering the successful combinations of message, messenger and subject.

**Lesson 8.** Programmes which aim to establish children as ‘partners’ in the development of the daily activity which makes up an initiative have to establish the boundaries of the child’s expertise. These programmes tend to be researched using a ‘formative approach’, which focuses on the day-to-day interpretations of those involved. This method, however, often lacks range in terms of coverage of viewpoints and the ability to trace outcomes. It frequently ends in advocacy rather than evaluation.

**Lesson 9.** Programmes should be developed with the aim of identifying and counteracting ‘risk factors’. But the antecedents of social problems are always multiple. Evaluation thus has the task of understanding what risks a programme can and cannot confront, and to ensure that the tractable problems are pinpointed and attacked. The key priority is to ‘get the grease to the squeak’.
Lesson 10. Before any intervention or programme commences, there should be intelligence on the persons and situations at risk. And after any evaluation, such understanding should be refined considerably. Such discernment of risk and response is best gathered by a close analysis of the behaviour of particular sub-groups under the initiative.

Lesson 11. Having children as programme subjects and research subjects provides evaluation with a special concern with ‘age’. The aptness of an intervention never depends simply on age per se, but on the child’s developmental stage and their exposure to the contributory problem and their potential control over prevention strategies. The key strategy for promoting children’s participation in the research process lies in the adaptation of standard instruments, presenting questions in frameworks and formats with which particular age-groups are already familiar.

Lesson 12. Safety problems congregate in hazardous environments. Many of the risks children face are specific to local environments. Programmes aimed at children’s safety thus need to tap into local knowledge of the conditions which generate the problems they are set to tackle. Evaluation research needs to map differences in peoples’ everyday wisdom about their communities and measure success in terms of growth in understanding of, and responses to, the local hazards.

Lesson 13. Programme effects are always diverse and evaluators should never rely on singular measures to monitor programme success (or failure). Evaluators should always use multiple measures thereby anticipating an uneven footprint of outcomes associated with the many processes involved in constructing, implementing and experiencing initiatives.

Lesson 14. A hierarchy of outcomes should be established for each family of programmes, following through the changes in understanding, attitudes and behaviour that interventions seek to affect. Outcome measurement is more secure the further it travels along the understanding \(\rightarrow\) attitudes \(\rightarrow\) behaviour chain. It is often infeasible to measure long-term behavioural outcomes of a programme and actual outcome measures should be chosen to optimise the balance between the ‘feasibility’ and ‘strength of proof’ of potential measures.

Lesson 15. Programmes cannot be considered properly tested if they are assessed solely through the usage of ‘approval ratings’. Such measurement methods are often seriously contaminated with error due to ‘subject effects’. Sentiments of ‘approval’ for a programme are so distantly related from the question of whether it has changed behaviour that it is probably a waste of research effort to attempt to perfect such measures.

Lesson 16. Programmes cannot be considered properly tested if outcomes are measured solely through the usage of self-reported behaviour. The accuracy of self-report varies from topic to topic and from child to child in any programme area. Each field of programming needs to build up its own intelligence on the validity of self-report by considering the power of recall, level of stigma, degree of anonymity and capacity for deception associated with the task.

Lesson 17. Any education programme worth its salt will produce educational gains and less evaluation time and effort is required to show that this is the rule. The more ‘proximate’ to reality the assessment of knowledge-gain measure the better. Evaluations without the resources to carry through investigation to behavioural outcomes should concentrate on assessing the variability in knowledge gains across different subjects and different aspects of the programme curriculum.

Lesson 18. Measuring programme outcomes by simulating the situations in which they are intended to have effects is an important, if limited, method for the evaluator. Provided ethical and practical issues can be dealt with, the more true-to-life the outcome measures the better.

Lesson 19. Retrospective surveys can provide broad-brush information on the coverage of programmes and some evidence on their (otherwise obscure) longer-term impact. Caution must be observed in making causal inferences about programme impacts from survey data which is not able to monitor closely the subject’s involvement with a programme.

Lesson 20. Evaluation research should always seek a multiplicity of outcomes, including those which fall outside the intentions and expectations of programme-architects. Unanticipated outcomes of programmes flow from programme complexity, subject over-reaction and bloody-mindedness, technological failure, legal loopholes and, occasionally, from nowhere.

Lesson 21. Researchers should never evaluate a programme ab ovo, they should always inspect previous broken eggs. Policy-makers attempt to produce new responses to old problems but the chances are some very similar initiative will have been put into place sometime, somewhere. Periodic reviews of existing research are thus vital. The process of review itself, however, takes on many forms from the catalogue-of-abstracts to quantitative-meta-analysis. The most beneficial way to conduct a review is to study each piece of research in the light of its contribution on ‘what works for whom in what circumstances.’
‘It seems as if there is no happy medium. You either let them out and you’re careless, or you keep them in and you’re over-protective, and your lad’s gonnae have an accident because you cannae teach it to be street-wise if you’ve got him in the house’. (Glaswegian parent – quoted in Roberts et al., 1995:66)

1 Introduction

Evaluation research is charged with the task of weighing up the pros and cons of social interventions in order to provide a more rational basis for policy decisions. It is thus supposed to provide the empirical sheet-anchor in the process of ‘evidence-driven policy-development’. But evaluation has proved a difficult business and opinions differ as to whether it has achieved this goal. Proponents of evaluation regard meticulously won and judiciously chosen evaluation data as furnishing the only safe mooring point to programme-development, whilst opponents perceive a slipping anchor, with highly selective and partisan research findings being dredged up as soon as the wind or current swing round to the latest policy offensive.

The goal of this paper is to provide a modest contribution to improving the conduct of evaluation research in an attempt to tip the balance to a more rational utilisation of empirical evidence in policy-decisions. It aims to do so by reviewing previous efforts to build an evidence-base in a particular corner of the policy field: interventions aimed at children. Throughout our review, we acknowledge the mixed fortunes of evidence-driven policy-development. Accordingly, our method consists of an examination of how previous inquiries have confronted the many pitfalls of applied research. Our dual goal is to select out those examples of ‘best practice’ which might sustain future research and to weed out some ‘worst practice’ by showing how it might be eradicated or improved. Evaluation research is a quintessentially practical form of inquiry and thus we deliver the lessons of evaluation research using a ‘show and tell’ approach, in the belief that methodological progress is best advanced through the provision of clear exemplars.

Our review, however, aims to provide more than a balance sheet of the ‘strengths’ and ‘weaknesses’ of the specific methods in the evaluation tool-box. The review process itself affords the potential of recognising and bringing together the benefits that accrue from particular technical approaches. Evaluation’s stock-in-trade is the case study of new programme X in place of Y in time-frame Z. Its evidence-base is, therefore, composed of rather piecemeal, partial and fragmented studies. Any particular study will allow particular inferences to be drawn but will be silent on a range of other matters pertinent to policy-development. By collecting together the pieces of the jigsaw, the process of review may suggest ways of synthesesing the varied approaches in order to provide a more powerful evaluative framework. To illustrate, let us consider some of the standard pitfalls that beset the typical evaluation study:

- Problems with the ‘scope’ of evaluation - most evaluation studies are ‘one off’ and the specifics of context and locality often make it tough to draw transferable lessons.
- Problems with the ‘rhythm’ of the typical evaluation - the policy-making timetable demands results tabled tomorrow and this makes for ‘breathless’ and ‘quick and dirty’ evaluation.
- Problems with the ‘artificiality’ of evaluation - concentration on evaluating ‘demonstration’ projects means that initial success is often hard to replicate due to ‘showcasing’ effects.

In all these cases, solution lies in trying to learn and incorporate lessons from previous evaluations:

- The cumulation of evaluation research depends on charting and understanding the successes and failures in previous initiatives across a whole family of programmes.
- The same programme theory might have utility across different intervention fields, and a significant borrowing of evidence may occur through policy departmental ‘glass-walls’.
- An evidence-base created over time will anticipate some developments in policy-making and so encourage researchers and policy-makers to cease wheel re-invention with the over-duplication of similar, one-off inquiries.

In short, our task in this report is to begin exploration of how we might begin to achieve some of these advantages and so explore the potential of ‘joined-up’ thinking in evaluation research.

So much for our broad approach, let us now consider our specific task in more detail. Whilst this is a monograph about evaluation methodology, it is, first and foremost, a commissioned piece of research with a quite specific mission. Our main brief here is to produce our review in an attempt to clear the ground for future Transport Research Laboratory (TRL) research for the Department of the Environment, Transport and the Regions (DETR) on the evaluation of road safety education. The report attempts to ‘read-across’ from evaluations of educational/health/safety initiatives aimed at children for pointers and guidelines for future evaluation in the field of children’s road safety. Our task is one of ‘learning lessons’ and, indeed, the motif of lesson learning is vital throughout the entire domain of evidence-driven policy-development. For it to be successful, the whole process needs to be imbued with a flow of teaching and learning. Thus we aim to learn lessons from previous evaluations - in order to inform research to come - whose findings can be incorporated within subsequent programme building - which in time should provide better road safety lessons for children.

Our quest, therefore, is to facilitate such a sequence. Our aim is the relatively modest one of ‘starting the ball rolling’, trying to stimulate the eventual contribution evaluation will make to the process of protecting children from one of the commonest causes of death and injury. Such high hopes need to be balanced by prefacing the report with some rather more down-to-earth remarks about its limitations - of which three should be made quite clear.
One. To repeat, this is a ‘monograph about methodology’ but it is NOT an evaluation text-book. We shall not, therefore, be engaging in the ‘paradigm wars’ over first principles and plotting the downfall of ‘experimentalism’ by ‘constructivism’ (or vice-versa, or similar). Nor are we seeking to produce a technical manual, teaching lessons on the finer points of the specific methods which have found their way into evaluation such as ‘focus groups’ or ‘longitudinal study’ or ‘log-linear statistics’ etc. Our methodological domain is more modest and lies with the intermediate, strategic concerns outlined above.

Two. Although the lessons to be learned are predominantly methodological, we do not believe in methodology-for-its-own-sake. Thus, in making judgements about what constitutes good evaluation practice, we will inevitably stray into the issue of what makes a good programme. Whilst we will not, therefore, refrain from making certain limited observations on ‘what works’ in interventions aimed at children, readers should note that this is the by-product and not the end-product of our review. Our powder is saved for consideration of ‘what works’ in evaluation.

Three. This is a report for road safety initiatives with children but it is not about road safety initiatives with children. It is being produced in a series alongside other research on road safety and its readership, hopefully, will contain both denizens and newcomers to the world of evaluation initiatives for children. That readership should note, therefore, that the present authors make no claims to a command of the aforementioned substantive areas. Our expertise lies in methodology and evaluation. We are commissioned here to search for transferable lessons about the evidence-base across policy fields, the whole point being to apply a fresh look through the fresh eyes of the outsider. These fresh eyes, however, will be guided by some old principles and plotting the downfall of ‘experimentalism’ by ‘constructivism’ (or vice-versa, or similar).

Two Materials reviewed

This section provides a brief description of ground covered and ground cleared in our selection of previous inquiries. As explained above, our initial remit was to examine evaluations of existing educational and health interventions aimed at children. Material was collected using the customary information science techniques of ‘prodding’ a selection of data bases with groups of ‘key words’, and waiting for the computers to spill forth reference after reference. Such a process captures many irrelevant references as well, but it is preferable to edit from an embarrassment of riches and rags than to miss relevant sources. Keywords were truncated using ‘wildcard’ symbols (*), so as to pick up all possible forms of a term from its basic root (particularly important in relation to trans-Atlantic spellings: e.g. the keyword ‘program*’ will pick up both ‘program’ and ‘programme’ and all derivatives).

Searches were organised in concept-groups as follows:

1 words such as methodolog*, evaluat*, analy*, meta-analy*, review*, assess*, research*, good practice*;
2 words such as program*, campaign*, initiative*, interven*, safe*, scheme*, project*, prevent*, skill*, promot*, educat*;
3 words such as child*, young*, youth*, juvenile*, bab*, teen*, paediatric*, pediatric*, kid*, infant*, pre-school*.

Data bases searched were as follows:

1 Blaise-line - the on-line retrieval service operated by the British Library.

Files accessed were:

- BNB (British National Bibliography) contains records for books and first issues of periodicals published in Britain since 1950.
- CONF (Conference Proceedings Index) contains records for proceedings of conferences, symposia, seminars and workshops acquired by BLDSC since 1964.
- DOE (Department of Environment Library) consists of records for books and journals of interest to the DETR and external bodies with similar interests.
- DSCM (Document Supply Centre Monographs) lists books in Western European languages held at the BLDSC. Coverage is of items published 1980 onwards.
- SIGLE (System for Information on Grey Literature in Europe) contains records of semi-published documents since 1981, e.g. Research Reports, working and discussion documents, some translations and conference reports, theses, some official publications and local authority documents.
- SO (Stationery Office) A collection of all government publications included in HMSO/SO catalogues since 1976. SO also publishes on behalf of other national bodies and it is a sales agent for a number of British and International organisations.
- WHITAKER lists out-of-print, in-print and forthcoming books published in the UK, and English language titles published overseas but available in the UK.
2 Journals contents pages:

- BIDS (British Library Inside Information) provides access to the bibliographic records of articles taken from 10,000 journals of the British Library’s most requested titles since 1992.
- IBSS (International Bibliography of the Social Sciences) a British Library database of journals and books at the London School of Economics.

3 Library catalogues:

- University of Leeds Library Catalogue.
- Policy Research Institute Library Catalogue.
- Leeds Metropolitan University Library Catalogue.

4 Internet

- A preliminary search of the Internet was undertaken but the results were disappointing. The main Internet search engines, such as Alta Vista and Infoseek, generate a long list of documents, which are automatically ranked in order of relevance to the keyword search and this allows for less complicated searches than with the Boolean terms above. Despite this, the information generated from the Internet searches was not particularly accessible, being either in the form of a mania of public-service ‘briefings’ or making references to ‘grey literature’ which was not easily accessible from this side of the Atlantic.

5 EVALTALK

- This is the ‘discussion list’ of the American Evaluation Association. All e-mail received by the LISTSERV is distributed to all members. We posted two messages. Initially, we presented a broad request for any interventions involving children but this, presumably, was rather open-ended, as we received only replies that sought further clarification. Subsequently, we considered EVALTALK may be of more use in retrieving references to specific issues. However, our second, focused, request also failed to produce any useful responses. It may be that we were merely unfortunate in the particular requests we posted. We were also unable to access the ‘strings’ of EVALTALK, a catalogue of previous discussions. So, although we failed to generate useful material from EVALTALK, we recommend further exploration for anyone following in our footsteps.

   With the exception of the latter, these sources provided a mass of material on a massive range of interventions and evaluations thereof. Next, came the task of turning database references into hard-copy, source materials. We were fortunate in that the local Leeds libraries have massive educational and medical collections, so we had immediate access to much useful material. We also collected certain crucial items directly from the National Reference Library and, for the rest, we burned a large hole in an inter-library loans budget. The collection was topped-up by ‘cold-calling’ a range of institutions and authors for some of the more inaccessible reports. For the record, it should be stated that the traditional, manual search method of discovering some key book or paper, and then searching for references in bibliographies therein, was still able to provide some of our most useful leads.

   The next task was to create some sense of order in the piles of discharging material. The initial categories of ‘health’ and ‘education’ programmes were soon breached and thus we created a basic typology of ‘programme families’ to guide the gathering of materials as in Table 1. Save for one (to be discussed in the next paragraph), the content of these classifications is self-evident and the framework served well enough as a first-base for organising material. Our aim was to achieve a balanced portfolio of programmes and research across these areas. Our files were in fact lopsided, with a vast preponderance of material relating to health education and relatively little to, say, environmental programmes. This, no doubt, reflects the preponderance of initiatives and evaluations in certain policy-fields and we ended with a range of studies which we consider a decent representation of the work available. We ceased collection of materials early in the third and final month of the project in order to concentrate fully on the analysis. This guillotine landed somewhat arbitrarily but this is the same-as-it-ever-was in such research, leaving us with no shortage of materials with which to research.

   We also gathered, as our ninth category, a small amount of material on research on children per se for which we searched outside the programme evaluation field. Many readers will be aware that the topic of ‘childhood’ is burgeoning - as represented by new journals, new academic departments and institutes, new national research and governmental initiatives. Considerable methodological thinking has occurred in the wake of these developments, devoted to the topic of the ‘child as research subject’. (Qvortrup et al., 1994; James and Prout, 1990) Clearly there is much that is of relevance in this literature to our study - but also much that is not. We note, for instance, the turf-wars and paradigm-posturing between the so-called ‘BIG 3’ - the developmental perspective, the socialisation perspective and the emerging social construction perspective, as they compete for pride of place in the new discipline. The consequence is that we have only cherry-picked at this material in search of direct pointers to evaluation research. There is probably more that can be culled as this new material develops - a point that should be added to the list of project limitations above.

   Table 1 lists some typical examples of the types of interventions we reviewed. Within each intervention family, of course, yet further distinctions could be made in respect of:

| Differences in evaluation methodology utilised - randomised control trial, quasi-experimentation, process evaluation, action research etc. |
| Differences in type of research materials produced - journal papers, official reports, grey literature, promotional literature, existing reviews (more than we imagined!) etc. |
The significance of these distinctions will become evident as we move to analysis and the evaluation of evaluations. It is nevertheless useful to stress the diversity of interventions under review using the following elegant summary of children’s programmes from Weissberg and Bell (1997):

“We can identify at least four distinct prevention approaches:

a those concerned only with changing the cognitive-behavioural-affective skills of target children without reference to the environment;

b others that focus on changing both children’s skills and their interactions with those in their surrounding environment (e.g., family and classroom based interventions);

c certain multi-component interventions that emphasise simultaneously transforming the organisational policies, structures, and practices of settings that affect adults as they attempt to enhance children’s skills; and

d other multi-component programs that change structures and policies at a more macro level without explicit person-centred efforts to change children’s behaviours”.

This initial framework of programme types and subtypes comprised our stock of materials. A final adjustment to the sample was made by the removal of a significant amount of material that was analytically slight, entirely descriptive or merely promotional. Thereafter, the aim was to find those studies that exemplify particular evaluation and programming issues. And at this stage analysis came into play. As we clarified our ideas on what makes for good evaluation and strong programming, we returned again and again to these records in search of examples and exemplars of good practice.

3 The analysis - 21 lessons to enhance evaluation

We submit our analysis in the form of 21 lessons to guide future evaluation of programmes aimed at children. Such a form of presentation may seem somewhat presumptive and prescriptive, so we should make clear that they are intended to facilitate further discussion rather than end it. Evaluation is no different from any other area of social science in that its methods evolve. Principles are laid down, found wanting in practice, and then revised. Our attempt to provide some guidance for future practice is part of that process.

Indeed, readers should note that these ‘lessons’ have already been through the mill of revision and adjustment. A first draft of this report was presented at a DETR seminar in March 1999 and contributions here have lead to substantial reshaping of the arguments in two sections and a refinement of our conclusion.
As a final preliminary, let us offer a reminder of the scope of these lessons. What follows are our answers to a particular range of evaluation questions - of which the reader is entitled to know, ‘who set them?’. The 21 propositions are in fact a distillation from three sources. Most arise directly from the original DETR/TRL commission which framed the broad agenda for this monograph. Inevitably one is guided by one’s own research experiences and so our schedule is further shaped by matters that have proved central in evaluation research and the methodological writing of one of the present authors (Pawson and Tilley, 1997; Duguid and Pawson, 1998). Also present is a class of problems which were not anticipated in advance but simply shouted at us in the materials reviewed.

Recall, also, that this is no evaluation treatise, thesaurus, text-book, or technical-manual and our lessons are directed only at strategic issues which are commonplace in evaluations of programmes aimed at children. We begin in our first five lessons with certain ‘common-to-all’ issues and then work into more specific evaluation problems associated with the sub-types of programme identified in the paragraphs above. For each lesson, we provide some essential conceptual background and give working examples of the problem (and, where applicable, the solution). Each section ends with a boxed summary which attempts to encapsulate the main issue in the form of guidelines on some do’s and don’ts in future research.

3.1 Embeddedness

One needs a clear understanding of the character of social programmes in order to know how they may ‘work’. And in this, our first and most general lesson, we try to capture something essential about the nature of all the initiatives we encountered in our review. Programmes never operate in a social vacuum, they are always inserted into an existing flow of social conduct. Extant social arrangements conspire to generate behaviour deemed to be damaging, dangerous or deviant and the policy-maker intervenes with programmes to generate behaviour deemed to be damaging, dangerous or deviant and the policy-maker intervenes with programmes to induce change? First of all, it may be something to do with the effectiveness of the ideas which are exchanged; improvements in understanding, skill, and aptitude lay at the base of much behavioural, personal and social change, and a programme’s ideas are usually reckoned to be the engine of these developments. Secondly, the background of the individuals will make a difference, the intervention’s ideas will stick somewhat better according to the balance of the characteristics of the teachers and the learners involved in the programme. Thirdly, the institution which is charged with delivering the programme will be significant; the choice of whether learning environment is a school, colleague, community centre, prison, CD-rom or distance-learning manual may be crucial to its uptake and acceptability. Fourthly, the societal infrastructure will be vital in enabling the programme to work; unless the world outside the classroom is receptive to the ideas and individuals emerging from the initiative, any positive change may still be stymied.

Our review detected a growing emphasis on this multi-layered approach across the various domains of evaluation. The concept of embeddedness is, for instance, termed the ‘ecological approach’ by some researchers in the health fields. Ecological theories of health education thus take account of the effects of several levels of influence - individuals, families, local neighbourhoods, cultural groups and social mores. Typical of several appeals considering how an ecological approach could enhance mental health service evaluations was the following:

‘Most child mental health evaluation research has examined the effects of treatment on individual children with specific disorders. Only rarely does evaluation incorporate the family and even less the characteristics of the system of care, such as ways in which use of one treatment modality, for instance in-patient treatment, influences other treatments such as outpatient family therapy or the provision of respite care. Yet treatment of children with emotional and behavioural symptoms usually requires a complex multi-system strategy with collaboration between a number of professionals. Several administrative bodies may be involved; hospital based mental health services; special education; social services; the courts or prisons; and community health.’ (McGuire, 1997)

Further calls for the ‘multi-level’ approach come in the area of ‘child abuse’ (Cox 1997, 1998 and from the Consortium on the School-Based Promotion of Social Competence, 1994). We find it difficult to conceive of any area of social policy (including road safety) to which such a ‘depth’ approach would not apply and so submit that the notion of ‘embeddedness’ be embedded routinely as the framework for constructing the evidence-base. Road safety initiatives need to be understood ‘up and down’ as well as ‘across’ the following matrix:

- **Ideas** - highway codes, crossing points, judging distances, parked cars etc.
- **Individuals** - children (across age groups), parents, teachers, road safety officers, police, instructors etc.
- **Institutions** - schools, clubs, parent’s organisations, media campaigns etc.
- **Infrastructure** - traffic conditions, legislation, road design, public transportation policy etc.
Programme ideas mounted at any one of these levels (or at any of the levels from Weissberg’s typology above) are not inoculated from influences at other levels. Successful ‘top-down’ change, such as that introduced by fresh legislation, requires the will of individuals and the support of institutions in gaining everyday acceptance of new law. Effective ‘bottom-up’ change, such as that introduced via classroom ideas, needs to be nurtured by supportive teaching, in a conducive forum, to a plan that is accommodated within existing social norms. Policymakers may or may not anticipate such depth in their endeavours but the evaluator cannot ignore these levels, for programme success and failure lies betwixt them all. We read, for instance, of one account of a ‘five-a-day fruit-and-veg’ health education initiative (Boaz and Ziebland, 1998) which reported success in the learning environment of the classroom but that, over in the canteen, the same authority’s meal service was busily responding to the demand for chips and chocolate cake.

Coming to the ramifications of embeddedness of programmes, we begin with a warning to avoid ‘flat evaluation’. Research which seeks to understand programme effectiveness at only one level - the commonest form of which is attempting to understand behavioural change as cognitively inspired and driven by individual differences - will often fail to see that the wood is surrounded by trees. The notion of embeddedness thus provides the background but also the challenge to evaluation research. Whilst it is plain that the four I’s will carry significant sway on the outcome of all educationally-oriented initiatives, it is also clear that they may constitute a threat for researchers. The notion of embeddedness may conjure up a research task in which everything is connected to everything else and, accordingly, the researcher may feel under-resourced when charged with evaluating ‘a programme’, when in fact they have to deal with all of these different levels of reality.

Our first lesson does not in fact require an omnipotent researcher able to control and measure everything from societal cultural flows to the brain’s cognitive flashes and all in between. We will come to actual ways of slicing the onion in subsequent lessons. For now it is sufficient to place the notion of depth-evaluation on the agenda.

**Lesson 1.** Evaluation research should begin with the recognition that programme efficacy depends on the ideas, individuals, institutions, and infrastructure which make up the intervention. In any particular investigation, it is unlikely that all such features may be controlled and measured but efforts should always be made to observe, monitor and record basic information on all these features. It is particularly important that collective wisdom about the four I’s should accumulate across investigations.

### 3.2 Programmes are theories

The rhetoric about programmes often makes them sound like things that are set in motion mechanically to be followed automatically by their results. Thus we are often inclined to take for granted that smaller classes lead to greater educational achievement, that skilling the workforce will lead to greater employment, that warning of health risks will lead to more wholesome behaviour, that more imprisonment will lead to less crime opportunities and so on. But, as we shall see, there is many a slip between programme cup and outcome lip, due to the diverse ways in which initiatives can be interpreted by different stakeholders in different circumstances.

At this stage we introduce an important distinction often made in the literature - between programmes which are ‘passive’ (in the sense that their intended effects work regardless of the ideas and preferences of their intended subjects) and programmes which are ‘active’ (in the sense that their intended effects work through the reasoning and predilections of their intended subjects). In the first category are interventions such as fluoridation of the water supply, and in the second campaigns giving advice to brush-twice-a-day. Passive programmes are normally, and correctly, considered somewhat easier to evaluate - the initiatives are singular and their processes simple, and so it is easy to monitor their presence and absence; interventions tend to come on tap instantaneously and before-and-after comparisons are therefore normally easy to monitor; impacts are typically sought at the aggregate level and thus significant subject numbers are easily come by. Whilst passive programmes may seem to promise an easier life for the evaluator and thus a safer bet for the policy-makers, our mention of them here offers only passing comfort, since they are most, most uncommon.

Passive programmes are often thought of as depending on some environmental or physical or technological intervention for their impact but these activities are not as impervious to human judgement as they may first appear. For example, CCTV surveillance to prevent (among other things) disturbances caused by the presence of adolescent gangs in town centres does not provide some impenetrable barrier to such gatherings. Success depends on how the images are deciphered by the operator, how quickly the police can mobilise in response to a call, how aware the youths are of the surveillance, and how apprehensive they are of its consequences. This chain of reasoning can get very long and convoluted, as illustrated by Norris’s (1998) research which observed youths staging mock-fights in front of city-centre cameras in order to invite police activity for the pleasure of wasting police time. Following the same argument, the barrier formed by a well-designed, well-fixed stair-gate may well work because it is impenetrable for those who crawl or toddle, but for older children who jump and climb, the obstacle has effectiveness only according to whether it is perceived as a warning, or a nuisance, or a piece of playground equipment. In the same manner, smoke detectors, school salads and street lighting depend on the use made of them, a point put rather vividly on the former by Malonee et al. (1996) who reported that in a scheme involving free distribution of smoke alarms, only 50% of them were properly installed and functioning after 2 years. Indeed some alarms were never installed, with the battery being...
perceived as more valuable than the bleep, and often being purloined for other uses.

As far as the evaluator goes the distinction between active and passive programmes should be treated as a continuum, the entirely passive subject being something of a rarity. We proceed on the basis that, in the main, the programmes we have reviewed depend for their efficacy on their interpretation by programme subjects. This places a second, general demand on the evaluator to be aware of the hunches, folk-theories and everyday wisdom of all participants to a programme and to ensure that they figure in the evaluation.

There are a range of custodians of relevant within-programme theories. These can be put into three main categories: the programme architects - policy-makers formulating and agreeing to fund initiatives; the programme practitioners - those whose job it is to implement the programme on the ground; and the programme participants - those whose behaviour may alter as an outcome of partaking in the programmes. Each will have a different ‘take’ on the programme, reflecting differing forms of experience of it, had at different stages. All may be more or less inchoate, and will generally require reconstruction and formalisation by the researcher (of which more anon). The key point here is that the evaluator should always look for match and mismatch between the various theories in play, for here lies basic evidence of the difference between programme success and failure.

As an example, let us consider the fate of an extremely simple hypothesis which we might term ‘dishy-David-Beckham-theory’. Mitchell’s (1997) analysis of health education messages in teenage girl’s magazines begins with the discovery of one theory at work in the guise of the editor’s decision to increase coverage of ‘positive sporting role models’, since this might persuade her readers to equate ‘desirability’ with ‘fitness’. Mitchell went on to pose this theory directly to the readers in the following exchange:

‘Interviewer: ‘But do you think the fact that these good-looking blokes are footballers has any effect on girls’ attitude to playing football?’

Girl: ‘No, I think it has more effect on them to watch football, well not the football - the guys.’ (general laughter and agreement), (1997)

Here then is a pointer to a somewhat misdirected theory, which is perceived as rebounding, rather substantially, as follows:

‘There was a noticeable lack of sporting female role models. In contrast, male sporting heroes, particularly ‘dreamy’ professional footballers, received substantial coverage. These articles reinforce the idea of the female role as spectator/admirer and therefore implicitly discourage girls from participation.’ (ibid.)

Programmes often begin life with policy-makers and practitioners assuming one thing and programme subjects thinking another. For example, seat belt laws may be seen as ‘unwarranted intrusions of state power into a realm that ought to remain the preserve of an individual’ (Adams, 1995). The mark of a successful programme is the extent to which behaviour draws towards a common understanding. Our second guiding rule for evaluation makes the point that programmes are speculative theories about how individuals will respond to ideas. One vital task in evaluation lies in tracing the trajectory of these theories through different stakeholders and across the duration of the programme.

Lesson 2. Programmes are not things but conjectures, and the evaluator needs to test them as conjectures. The typical programme under review takes the form of a theory which speculates, ‘if we do X, we may persuade children to do Y’. The ideas embodied in a programme, however, may be met with acceptance, incomprehension, indifference or resistance on the part of children. A ‘theory-driven approach’ to evaluation is thus recommended, which considers the congruence of the conceptions of all concerned.

### 3.3 Process and outcomes

In this section, we move from broad lessons based on the nature of programmes to a consideration of the competing approaches to evaluation itself. Part of our appointed task is to search for the most objective methodologies for conducting evaluation and this is the first of several sections which will build to a cumulative judgement on that question.

The most common distinction within evaluation methodology literature is between ‘outcome’ and ‘process’ inquiries. The former seeks direct evidence on whether a programme works by measuring its impact - by charting changes in behaviour brought about by the programme. The latter seek to understand what the programme actually does to bring about change - by examining the activities and reasoning of stakeholders in the programme. There are a variety of techniques associated with each goal, but we can say as a first approximation that outcome analysis is ‘quantitative’ and that process inquiry is ‘qualitative’. This division of labour has sown seeds of ‘positivism versus phenomenology’ rivalry into some of the evaluation literature. Our starting point on the matter is quite plain and quite the opposite, and can be stated thus - ‘to ‘qual’ or to ‘quant’: that ain’t the question’. Such an appeal for mixed-method research is hardly revolutionary - it forms, for instance, the clarion call in the editorial in the first edition of the new Evaluation journal (Stern, 1995). Our purpose in this section is to make plain the case for a combination of process and outcome approaches but also to begin to explore the rather more vexed question about how one combines qualitative and quantitative information. We pursue the argument by looking at two pieces of research on similar schemes - one an outcome study and one a process evaluation.
The initiatives examined here have a long history in prison and corrections services and go by different names - ‘juvenile awareness programmes’, ‘youth aversion schemes’, ‘public speaking panels’, ‘prison inmate encounters’, ‘youth visits’, but most famously as the ‘scared-straight initiative’. The first study we will refer to is Finckenauer’s (1982) research on the New Jersey Juvenile Awareness Programme. This inquiry is famous for its frankness about bullying, skulduggery and anguish - and that is only about the correction’s officials and researchers! Here we stick to the methodological lessons.

The programme took shape in the endeavours of the lifer’s group at Rahway State Prison. They were concerned to dispel their stereotype image as sadistic, inhuman creatures and so devised an initiative to ‘keep kids out of trouble’. With some imaginative support from corrections staff they hit on the idea of ‘bringing them [young offenders] in to find what a life of crime is all about - straight from us’. Over a period, the initiative turned from a form of counselling (‘rap sessions’) into a form of shock therapy. The inmates presented a picture of violence, rape and self-mutilation in prison life and did so in a particularly graphic and confrontational way. The idea was (and still is) particularly newsworthy and was brought to national attention in the US by brilliantly evocative depictions of the visits, first in the Reader’s Digest and then an Oscar winning documentary Scared Straight. Publicity then overtook the programme. The imagination of the public and public figures was captured, especially in relation to the fanfare accompanying the film:

Fact: Virtually all adult criminals have been juvenile offenders. IF ONLY THEY COULD HAVE BEEN STOPPED THEN.

Fact: 80-90% of kids on this unique programme are ‘Scared Straight’ - Take an hour and watch this powerful approach work.

Against this background, but before the main onslaught of publicity, Finckenauer had set himself the task of trying to discover whether the Juvenile Awareness Programme (JAP) actually worked. The 80-90% figure in the publicity above had been created in a flimsy, do-it-yourself attitude survey carried out by the inmates and corrections staff (more of which later). Finckenauer tried to gather some rather more dispassionate evidence by way of an outcome evaluation that followed up recorded offences of the programme subjects subsequent to the visits. The intended research design was to obtain a sample of local juvenile offenders and to divide them randomly, with fifty being put through the programme, and fifty acting as control group. A range of problems beset the randomisation procedures, most notably with one police agency taking their entire batch of eleven juveniles to Rahway instead of the random 5 or 6 whom they were supposed to select. In the course of the research 19 of the sample were lost through further bungles and kerfuffles between researchers and officials. The trial was then pushed successfully through to outcome data with no further alarms. Finckenauer’s overall conclusion was that the initiative was counterproductive in that the experimental subjects fetched up with a higher rate of failure (as measured by the presence of a recorded offence within six-months of the intervention) than the controls - as in Table 2.

Table 2 Outcomes of the JAP programme

<table>
<thead>
<tr>
<th></th>
<th>Success</th>
<th>Failure</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiments</td>
<td>27 (58.7%)</td>
<td>19 (41.3%)</td>
<td>46 (56.8%)</td>
</tr>
<tr>
<td>Controls</td>
<td>31 (88.6%)</td>
<td>4 (11.4%)</td>
<td>35 (43.2%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>58</strong></td>
<td><strong>23</strong></td>
<td><strong>81</strong></td>
</tr>
</tbody>
</table>

Source: Finckenauer 1982, p135

These unwelcome results arrived at the same time as national acclaim for the programme was building. Accordingly, the evaluation was put to unusually close scrutiny and met with lay and academic criticism galore. Methodological debate revolved around the small sample size and the faulty randomisation procedures. Finckenauer conceded, as he must, that since his sample was drawn from one jurisdiction, that the research could not tell us if the same result would apply to other similar programmes mounted elsewhere. But on all other accounts he was unrepentant and set in train a variety of further analyses to support his negative finding on the Rahway programme. Using ‘worst cases scenarios’ for his ‘lost cases’, namely that the missing experimental subjects would have all ‘succeeded’ and the absent controls all ‘failed’, he was still able to calculate that the programme had a null effect (p138). Furthermore, he used some background data (sex, risk, prior offence, age) collected on the subjects in regression and variance analysis, and calculated that whilst there were significant differences between the attendees and non-attendees, if these were partialled out statistically, that the experimental group still does worse than the control group.

Now we are left here with a rather typical state of affairs for such outcome analysis. Finckenauer’s inquiry goes some way to (im)proving his case that JAP didn’t work, but his method is open to familiar claims that there are further unreckoned, undiscovered, unmeasured differences between the two groups that could account for the differences in subsequent offences. Probably the best that can be claimed for the research was that there was no solid evidence that the particular programme worked. The mists descended further when along came a follow up study of Rahway (Langer, 1979) using longer time frames and a differently constructed sample, which showed that there was a significant tendency for non-JAP juveniles to commit more serious crime over a two-year follow-up period. Langer’s effort, arguably, had even less control over the matching of experimental subjects and controls, but we note the collapse into inconclusiveness after several years of effort.

And thus we come to our point about ‘outcome-only’ inquiry, which does not rest on the technical limitations of studies such as these - but is a criticism of the very
question that is posed. The outcome question is simply - 'did it work?' and outcome analysis allows the researcher to answer that question without recourse to the question 'why did it (or didn’t it) work?'. Now if one actually consults Finckenauer’s monograph it is clear that the reasons for the potential failure of JAP came to him quite independently of the experimental trial.

Finckenauer records in detail a long conversation with practitioner-lifer Frank Bindhammer in which the inmate recalls how his group’s initial aim was to act as the wiser, older brother and how this changed to confrontational mode under the glow of publicity. He also mentions the lack of control over the referral system with some kids been admitted without arrest records, some for ‘light stuff’, with some under medication, and others deeply embedded in the system (one youth attended the programme four times - ‘you guys are cool’). Analysis of the subject’s testimony also reveals process differences. One session had left a group of young boys so frightened that they exaggerated what had happened and claimed they had been kissed and fondled by the inmates. One of the girls attending the programme refers to the mixed signals she got from the programme - ‘they yelled at you, then shook your hand’, also adding - ‘I think it would be better if we [girls] went to a women’s prison’. Finckenauer also witnessed further departures from the ‘tell-it-from-experience’ programme philosophy, such as when a particularly young group of girls were subjected by the men to tales of the lesbian encounters they would be put through should they end up in prison. In short, we can see that the programme has the potential to work in several different ways. Some subjects might indeed be ‘scared straight’, some were apparently ‘scared crooked’, others were evidently ‘scared silly’, and still others seemed ‘simply confused’.

What is beginning to be built here is a picture of the internal complexity of the Juvenile awareness initiative, and the discovery of these internal workings of a programme is, of course, the very stuff of ‘process evaluation’. We can see how this latter method operates by looking at a recent Home Office study of some UK versions of ‘juvenile awareness’ initiatives. Lloyd (1995) examined such schemes at Garth, Risley and Maidstone. His methodology involved observation of ‘events’ at each prison and in-depth interviews with inmates, prison staff, referral agencies and programme subjects. No conviction/reconviction follow up was possible due to ‘time and funding’ constraints. The three schemes differed a great deal, one being perceived as basically educational, one more confrontational and one targeted specifically at car crime. Subjects ranged from social service and probation referrals to children from ‘school visits’. Space constraints forbid us from reproducing in detail here the range of viewpoints noted in interview. It will not surprise the reader to learn that they echo the brief selection of reactions to JAP noted above, being if anything more diverse because of the different character of the interventions.

Process-only evaluations normally have a role clarifying, formalising and weighing-up the diverse currents of practice and opinion flowing through programmes, in order to suggest recommendations which draw out, prioritise and develop their best practices. We can see this function clearly in the following brief paraphrase of some of Lloyd’s conclusions:

- The selection of the target group will have great implications for the nature of the project - the confrontational approach should be restricted to cautioned or convicted prisoners.
- Pre-sentence referrals should be discontinued - the highly negative and threatening portrayal of prison life could lead to self-injury or suicide if a sentence is imminent.
- Formal assessment should be put in place - sensitive and vulnerable young people should be filtered out.
- Adolescents with little criminal involvement might reject the advice offered [by lifers and long term prisoners] because they cannot imagine themselves ever committing very serious crimes and serving long prison sentences.
- School children will gain educational benefits and the non-confrontational aspects of the schemes should be developed to inform children about the criminal justice system.

Having portrayed the two modes of evaluation at work, we are now in a position to draw our conclusions. Lloyd’s findings are presented as recommendations but looked at closely are in fact ‘theories’ about how and for whom the programme might and might not work. This is quite inevitable and quite proper, for a good process evaluation will be extremely sensitive to the appropriateness of particular programme approaches for particular subjects. Typically, therefore, such research will perceive a patterned outcome emanating from a programme, suggesting precisely where it will and will not reach. This understanding of within-programme diversity is the great strength of process evaluation.

But Lloyd’s propositions are also theories in the negative sense that they are entirely untested hypotheses. Process evaluation provides us with some potential pros and cons about different programme pathways but does not follow them through to their destinations. It cannot say whether present shortcomings or newly suggested arrangements are supported in fact. Recommendations are made on the basis of advice from the participants and the judgement of the researcher. Such judgements may carry conviction because they are based on long experience. But the seasoned eye is in fact no substitute for empirical evidence that actually confirms that only the cautioned and convicted respond to confrontation, that educational gains only flow to the criminally uninitiated, that initial vulnerability in the subject or too close a time to actual sentence results in harm, that very serious offenders may not be effective as presenters, and so on.

These are all good empirical questions, at the very root of understanding how such programmes work. The important thing about them is that the overall programme outcome is actually the sum of these various pathways to success and failure. A mêlée of confrontational, educational and experiential mechanisms flow though such schemes and each process will trigger different responses.
according to backgrounds of the participants. Programmes do not work or fail per se, they flourish or founder according to the balance of ideas and subjects. Lloyd’s speculations about the inner workings of the Garth, Risley and Maidstone schemes shadow Finckenhauer’s doubts about the efficacy of the Rahway programme - but neither produces a definitive piece of research. Finckenhauer’s and Langer’s work provide us with two aggregate, snapshots of the outcomes of a scheme in action. The mixed messages of these inquires can probably be explained by a changing balance of approaches and clientele as the programme matured. Lloyd’s research produces some insightful but untested hunches about the inner mechanics of such programmes. Rather than relying either on fragile output data or a speculative feel for best practice, what is needed is an approach that combines the understanding of outcome and process. Rather than only tracing the differences between programme and non-programme youths, we need quantitative data that is sensitive to a whole range of conditions for the success of a programme, supported by some well-grounded qualitative observations about the processes which have generated the differences.

All programmes are constituted in a spiral of process and outcomes. Road safety lessons may be found useful, novel, obvious, boring or even intimidating - according to who teaches and who learns them. And the clearer the understanding of such generative mechanisms, the more who teaches and who learns them. And the clearer the understanding of outcome and process. Rather than only tracing the differences between programme and non-programme youths, we need quantitative data that is sensitive to a whole range of conditions for the success of a programme, supported by some well-grounded qualitative observations about the processes which have generated the differences.

All programmes are constituted in a spiral of process and outcomes. Road safety lessons may be found useful, novel, obvious, boring or even intimidating - according to who teaches and who learns them. And the clearer the understanding of such generative mechanisms, the more directed and precise is the study of outcomes. Once again, understanding these patterns will require the use of elaborate social process will always generate complex outcome patterns. Anticipating and understanding these patterns will require the use of both quantitative and qualitative methods.

Lesson 3. Programme outcomes are the result of programme processes and evaluations are always enhanced to the extent that the study of one supports the understanding of the other. Programmes dependent on elaborate social process will always generate complex outcome patterns. Anticipating and understanding these patterns will require the use of both quantitative and qualitative methods.

3.4 Experimental trials and tribulations

This section examines what is often considered the ‘gold standard’ in evaluation. Experimental evaluation, especially in the form of the randomised controlled trial (RCT), is normally considered to offer the securest method of discovering whether a particular programme has worked. We do not wish to challenge the proposition that RCT means are best if research ends consist of applying pass/fail verdicts on a particular programme trial. But we do want to question the question, and ponder whether classical, experimental-versus-control comparisons serve particularly well the broader requirements of policy-making.

Let us begin with a reminder of the familiar framework of experimental evaluation. Take two more or less matched groups (if they are really matched through random allocation, you call it real experimentation; ‘quasi-ness’ following from the impracticality of this in many cases). Treat one group and not the other. Measure both groups before and after the treatment of the one. Compare the changes in the treated and untreated groups, and lo and behold, you have a clear measure of the impact of the programme. The practitioner, policy-adviser, and social scientist are at one in appreciating the elegance of this strategy. The basic (untreated control group design with pre-test and post-test) is set down as Figure 1 using Campbell’s classic OXO notation (Cook and Campbell, 1979). The key stroke of logic in the design is that, being identical to begin with, the only difference between experimental and control group lies in the application of the initiative. Any difference in behavioural outcomes between the groups is thus accounted for in terms of the action of the treatment.

We found many strong advocates of RCTs in the papers we reviewed. Oakley et al. (1995) produced a widely influential set of principles in a BMJ paper on evaluating sexual health interventions for young people. The ‘key messages’ could not be clearer: journal editors and funders are exhorted to refuse the methodologically flawed, and flaws follow unless RCTs-with-long-term-behavioural-outcome-observation are utilised. Roberts has also produced a string of authoritative papers on the benefits of RCTs in evaluating child-care practice. The argument (Newman and Roberts, 1997) is the familiar one about ‘ruling out competing explanations’. The authors pose the typical example of a reported emotional ‘improvement’ following social worker visits to families. Invariably it is the case, the authors argue, that factors other than the intervention itself could be responsible for the observed progress - features such as ‘... parents’ emotional state improving simply with the passage of time; ... [receiving] help from relatives or other domiciliary services; parents
liking the workers and not wishing to disappoint them by reporting no progress; parents learning the ‘right answers’ to questions about their mental state through being exposed to such questions on a regular basis; and parents who took up the offer of service being the one’s who were going to improve anyway.’ If we suppose that these ‘other factors’ are distributed equally through the families who may be offered such a service, then assigning that population randomly to treatment and control conditions ensures a comparison in which the only difference between the groups is the exposure of one to the programme, which then should show up in post-treatment differences.

Does this logic hold? Does OXO set the standard? We think not and we think the studies we have reviewed demonstrate that it does not. From the outset, we should make clear the basis of our doubts. Often pitted against RCTs are ‘emancipatory researchers’ who see programme building as a matter of wrestling political control in favour of the oppressed and see experimental research on subjects as part of that oppression. Such baby-with-the-bathwaterism is not pursued here.

Another source of objections to the RCT lies in pragmatic and ethical scruples concerning the mechanics of the experimental process. A review by Towner et al. (1996), covering some of the same territory as ours, notes that only 18% of studies used RCTs and points to some practical reasons for their scarcity. Legislative interventions tend to come all of a piece and so can rarely be randomly allocated to one test location and not another. Setting up community-based interventions involves lengthy negotiations and researchers cannot easily dump a control sample from the negotiatees for the sake of science. The alternative of ‘matching’ communities is fraught with difficulties, they have untold complexities and unique singularities and are not referred to as ‘localities’ for nothing. Matching, in any case, works best with weight of numbers in the experimental and control conditions, and community interventions tend to be funded only by the handful. Even the random allocation of individuals (children) can sometimes be difficult, especially if the programme is carried out within a tightly-knit institution (where the mechanics of isolating the groups is often difficult) or in a dispersed intervention (like a television broadcast in which it is difficult to know who has received the ‘treatment’ and who has not). On top of these are the ethical problems involved in allocating some children to potentially advantageous conditions whilst allowing others to remain in disadvantage. Macdonald and Roberts (1995) respond robustly to the latter charge, showing that high ethical principles are hardly jeopardised if one apportions controls to a ‘waiting-list’, so that they can enter a programme later, when in fact we have some knowledge of its efficacy.

We take a pragmatic line on these pragmatic objections. Sometimes it is impracticable to mount an RCT, sometimes it is not. But again this is not our problem with the method. The real problem with RCTs was spotted almost two decades ago by Chen and Rossi (1983).

‘The domination of the experimental paradigm in the program evaluation literature has unfortunately drawn attention away from a more important task in gaining understanding of social programmes, namely, developing theoretical models of social interventions. A very seductive and attractive feature of controlled experiments is that it is not necessary to understand how a social programme works in order to estimate its net effects through randomised experiments.’

RCTs fail to register how programmes work in two important respects. First, is that the intervention itself is regarded as a singular ‘treatment’ and the point of the experiment is to register the difference between its presence and absence. One might get away with such assumptions in some programmes in agriculture and medicine but, as we have seen, social programmes are not things but theories. Programmes involving education, peer-group pressure, social clubs and so on are ‘active’ and designed to work through the reasoning of the subject. The ‘treatment’, in short, is invariably a complex social process and a crucial part of evaluation must always be that of trying to ascertain what it is about the programme that influences change.

One can see the slip in reasoning involved here by going back to one of Newman and Roberts’s ‘confounding variables’ in the social work visits example above. They ponder, in one instance, whether it is the ‘intervention’ or ‘liking’ for the workers which produced emotional improvement? But these are quite indistinguishable in an experimental and control comparison. Experimental subjects may get ‘both’, controls ‘neither’. Or, to put the matter more accurately, the intervention in such a case is the totality of the exchange between workers and subjects. Thus, generally speaking, programmes are the ideas that are covered and the manner in which they are expressed and the predilections of the participants and the relationships that are developed and so on. The ‘same’ programme will inevitably vary from application to application as these features unfold in the real world. What gold standard evaluation methods need is to get into the black box of a programme, by developing a way of understanding how the internal mechanisms of a programme lead to outcomes.

The second vital weakness of RCTs is the failure to appreciate that programmes are embedded within ideas, individuals, institutions, and infrastructure and that the impact of a programme depends on its reception at all these levels. As we have seen, the basic idea of the RCT is to flatten out all differences between experimental and control groups in order that we can say the programme and the programme alone was responsible for any change. In the effort to hold ‘everything constant’, the RCT attempts to create a contextless situation, thus overlooking vital circumstances which carry a programme to success or failure. Let us imagine the same, equally-effective, youth training programme offered in two quite different sets of circumstances. In a period of economic growth the experimental subjects may pick up more jobs (the outcome measure) than the controls. In a recession, with no jobs
available, the same programme may show no differences between the two groups. Different trials would emerge with inconsistent results.

All the programmes we looked at are surrounded by institutional and infrastructural circumstances which can be hostile or favourable to the programme but which are not part of the programme. Programmes are organised in actual communities, real schools, tangible streets, concrete environments. They are not dispersed generally through society. They are set amidst a balance of contextual conditions which may differ in terms of traffic-danger, drug-culture, dietary-opportunities, smoking-tolerance and so on. The RCT can offer no understanding of how such contextual conditions constrain the effectiveness of a programme because the experiment takes these conditions as a given and gets on with ensuring that the individuals within any current application are properly matched or randomised. But programmes work by offering subjects advice, knowledge, incentives, pressure to change their behaviour. Programmes are ideas and whether these ideas are taken up very much depends on the circumstances of the subject. Each real world programme location will thus draw in a different balance of individuals and gold standard evaluation methods need to incorporate some systematic method for discovering for whom and in what circumstances a programme will work.

In short, experimental methods in evaluation research offer a weak understanding of the internal mechanisms of a programme and of external contextual constraint (Pawson and Tilley. 1997). These drawbacks are often revealed in the production of inconsistent results across different experimental trials with the same programme. What was noticeable in our review was that even some long-standing initiatives, which are borrowed and adapted over the world over, manage to do so on the basis of surprisingly fickle rates of success. Let us look at a real example in some depth.

Royse (1998) reports on a recent RCT evaluation of a Brothers Project in Lexington, Kentucky. This is a ‘mentoring’ project, offering voluntary adult support for vulnerable youths, whose roots go back to the Big Brothers/Big Sisters movement that began in 1902. In the Lexington trial, African-American teenagers with low education achievement from poor, female-headed households were referred to the programme by counsellors and school principals. The mentors were African-American men, ‘typically college graduates in their thirties who had moved into the community as a result of employment’. With parental consent, the youths were assigned randomly to experimental and control conditions. Experimental subjects and controls were pre- and post-tested on a self-esteem test, a drug and alcohol questionnaire, and self-report on school performance and behaviour. Both groups were given an incentive ($15) to return (up to six times) for the assessments. The programme is described as follows:

‘In addition to providing the mechanism that brought together the mentor with the mentorial subjects, the Brothers staff also pursued a strategy of providing organised group recreational outings on a monthly basis. Such activities allowed mentors to compare notes with other mentors and receive social support, while taking some pressure off them for planning ‘fun’ events. These outings also gave mentorial subjects an opportunity to broaden their horizons, get out of their neighbourhoods, and develop new friends’.

In RCT terms the programme was a failure. No significant differences were found between the experimental and control groups on any of the outcome measures (except in some of the later assessments - by which time the response rate had dropped dramatically to render differences unreliable). Royse, having dutifully reported pages of null results, laments that mentoring programmes have a history of mixed results, and then goes on to speculate on the ‘host of reasons’ that may be offered for his disappointing findings. It is interesting that what he highlights here are potential differences in the inner workings of the programme and contextual constraints limiting the effectiveness of mentoring - precisely those factors which are not attended to in an RCT.

It is clear from the programme description that mentors have a free hand to develop their role and Royse knew this was a problem:

‘Hamilton and Hamilton (1992) observed that mentors often have different views of their purpose. What they call Level 1 mentors saw their primary purpose as developing a relationship. Level 2 mentors spoke of introducing options as their major purpose. Level 3 mentors stressed developing character, while Level 4 mentors focused on developing competence..... Hamilton and Hamilton concluded that an emphasis on learning to do something produced the most functional pairs.’

Royse also mentions previous work which had suggested that mentoring works better in the following conditions: the more the parents had contact with relationship with big brother/big sister; the more a big sister supports a lone-parent female as well as her child; with younger children who are more receptive to big brother/big sister authority; with constant, long-term meetings between mentor and mentoral subject.

It is downright curious that Royse drew on these explanations from previous research only to explain the failure of his trial, rather than incorporate them into his research. Clearly we needed to know about the character of the mentoring in the Brothers project. It might be that the efforts of types 1 and 3 mentors balanced out the efforts of types 2 and 4. It might be that further (type 5?) expectations were at work. It might be that insufficient Lexington mentors attempted to establish contacts with parents. It might be that the system relied too much on ‘fun events’. The black-box RCT method cannot help us with any of these hypotheses and it is clear that the
methodological failure in this instance is the lack of a supplementary theory-driven, process evaluation.

We also need to know if the context of the Lexington initiative placed limitations on its success. One notes that it was carried out with male mentors on 14 to 16 year-olds in female-headed households. The programme is thus automatically shielded from two potential success mechanisms - the greater openness of younger children to a big brother’s bidding and the sisterly support of female mentor to female parent. The RCT design cannot help us in understanding the significance of the initial choices on programme structure and it is clear that the methodological omissions in this instance are to do with the use of meta-analysis in cumulative programme building, and better exploration and targeting of sub-group progress.

We conclude that evaluators need to clear their heads of talk of gold-standard methods. Very few programme-builders believe any more in magic-bullets. Panacea programmes that work all the time for all comers do not exist. Much the same goes for evaluation methods. RCTs are the best method for giving an overall judgement on the efficacy of a particular programme trial. But the programme-architects and policy-makers need to know much more than this. Evaluation’s gold standard thus consists of the deployment of a variety of strategies and techniques. Our preliminary examples have demonstrated the necessity of utilising both outcome and process analysis. We have already shown the need for the guiding hand of theory in evaluation and demonstrated the requirement for studying programmes across their different incarnations. And so, whilst it is not the most attractive term, we are on our way to arguing that ‘methodological pluralism’ is the touchstone for evaluation research. Its promise and perils are explored in forthcoming lessons.

Lesson 4. Randomised control trials are often considered the gold-standard method in evaluation since they offer great certainty about whether a particular programme has worked in a particular instance. But they throw up inconsistent results because they provide no understanding of the exchange of ideas within a programme nor of the context in which it takes place. They are not an absolute priority for the policy-maker.

3.5 What is it about a programme which works, for whom and in what circumstances, and in what respects?

This is the shortest section in the paper but also the pivotal one. We introduce no new illustrations from evaluation research here but instead draw together the initial lessons in an attempt to capture the ‘basic question’ of evaluation research. Whilst there is no single, gold-standard method in the evaluation toolbox, we think that there is a gold-standard question and an overall research-into-policy strategy to go with it.

The venerable evaluation question ‘what works?’ has been teased into remarkably many forms, some more much useful to the policy-maker than others (Tilley, 1998). Tilley lists, amongst others: the opportunist question - ‘what can be made to seem to work?’; the practical question - ‘what worked for this problem here?'; the historical question - ‘what did work?’; and the constant-conjunction question - ‘what always works?’. Whilst the first and fourth may seem to offer instant attractions for the policy-maker, one soon discovers that opportunist, talking-up of evidence lasts only until someone else tries to emulate the study and that the search for panaceas is a speech-makers pipe-dream rather than a policy option. The second and third questions (the ones mostly asked in the studies reviewed) acknowledge that programmes work conditionally but, with eyes fixed firmly on success, they tend to under-report the ever-present constraints on effectiveness. However, it is of relatively little use to the policy-maker to know that a programme has worked on occasions in the past and in specific localities. They need to know what are the chances for a programme out there - in general and in the future. This requires a knowledge of what works, set against a rigorous understanding of the conditions and constraints. There is, in short, a rather longer version of evaluation’s basic question (see Figure 2) that captures better the vicissitudes of the policy-initiative - what is it about a programme which works for whom in what circumstances and in what respects?

‘WHAT WORKS’ really means:

(what is it about a programme)

which works

(for whom in what circumstances)

(and in what respects)

Lessons 6-9 Lessons 9-12 Lessons 13-20

Figure 2 Asking the right question: an organising framework

One can illustrate the power of this question by returning briefly to Rahway and Lexington and the two evaluations we have looked at in some detail. It is clear that there is no simple answer to the question of whether the Scared Straight or Brothers projects worked because they involved complex encounters which were capable of triggering ‘fright’, ‘admiration’, ‘terror’, ‘learning’ and ‘distancing’ in the case of the former, and ‘friendship’, ‘lifestyle-choice’, ‘character-development’, ‘practical skills’ and ‘laissez-faire’ in the case of the latter. Evaluation does not get off the runway unless an attempt is made to discover what it is about a programme which can affect change. And so it is in the further programmes we are about to review. Road safety programmes involve complex learning situations and the same rule applies about understanding precisely what is learned in the classroom and at the roadside.

Scared Straight and Brothers clearly varied in their impact on different types of subjects. Encounters with prisoners were considered to have a different effect on the
cautioned and convicted as opposed to the sensitive and the vulnerable. Encounters with mentors vary in influence according to the age of the subject, the gender of mentor and parent, and whether a solid triangle is formed between all three. Evaluation cannot hope to discover transferable lessons unless an attempt is made to discover for whom and in what circumstances a programme works. And so it is in the further programmes we are about to review. Road safety programmes target children of different ages, include diverse personnel and are applied in varied localities, and the same rule applies about understanding the optimal balance of participants and their place in the community.

Another reason for the lack of simple answers to the questions of whether the Scared Straight or Brothers projects worked is because they embody a diffuse set of expectations. Youth visits to prisons are aimed at preventing offending, reducing re-offending and broadening education. Mentoring was targeted at self-esteem, drug and alcohol misuse, and school performance and discipline. Ambitions for programme success also varied from the short to the long-term. Evaluation cannot hope to provide policy guidance unless it provides clear evidence about the different respects in which a programme may work which are revealed across a variety of outcome measures. And so it is in the further programmes we are about to review. Road safety schemes, seek to shift knowledge, attitudes and behaviour, and equip children for different periods in their lives, and the same rule applies about the need for comprehensive but targeted measures on programme effectiveness.

Lesson 5. Policy-makers should desist from charging evaluation research with discovering ‘what works?’, for no initiative will work for all subjects in all circumstances. The really instructive evaluation question is ‘what is it about a programme which works for whom in what circumstances and in what respects?’.

The remaining lessons break this master-proposition into sub-clauses and move into detail with respect to methods required to answer each separate question, as in Figure 2.

3.6 Learning the rules

One important class of programmes in our review is those which aim to instruct children in proper/appropriate/safe behaviour by having them learn and internalise a set of codes/rules/guidelines about how to conduct themselves. In one sense there is no mystery about what it is that makes such programmes work - children learn the rules and follow them. Alas, the first step does not always lead to the second. Rules are devised and learned in the abstract, and then applied in the concrete, so there is inevitably some slippage between the two phases. This so-called ‘fringe of incompleteness’ (Garfinkel, 1967) in the coverage of rules is captured by the everyday phrase the ‘rule of thumb’. A police officer might learn from experience how to recognise suspicious behaviour and we might attempt to formalise one aspect of such furtive behaviour under the rule-of-thumb watch out for people ‘moving quickly, carrying large bags, avoiding your eyes, disappearing into crowds’. Bright cops, however, do not apply this guideline to people ‘scurrying along, hefting suitcases, scanning the departures board, cutting through queues’ in train stations. As the saying goes - there are exceptions to every rule.

This gap between the conception of a rule and its application is thus a vital topic for the evaluator. Most safety codes for children tend to be based on absolutes - ‘try not to cross between parked cars’, ‘say no to strangers’, and so on - but these propositions confront children with persistent borderline dilemmas, as when the whole street is parked or when they are not quite sure about who actually counts as a stranger. Scott et al. (1998) note that it is not just children who struggle with such rules. ‘A Lothian and Borders Police Stranger-Danger talk to 9 year-olds identified police officers, shopkeepers, and ‘ladies with children’ as ‘safe-strangers’’. This predicament has been particularly taxing for programme-builders attempting to fine-tune rules for avoiding child sex abuse as we can see in the following:

‘Central to almost all sexual abuse prevention programmes is the touch continuum or the need for the child to differentiate between what is a safe and a not-safe touch. A fundamental problem here, which could prove problematic when a child attempts to use his/her knowledge, is that often abuse starts not with touching, but with the building up of a relationship between the victim and the perpetrator. Offenders have reported that they only gradually build up to physical touching after the child has grown to trust them. The reasons for this are so that the child does not recognise what is going on as abusive and, therefore, does not tell; and further, so that they feel a partner in the ‘relationship’ which makes the child feel responsible and, therefore, unlikely to tell. Therefore, when prevention programmes attempt to teach children the difference between safe and unsafe touching, the issue becomes very complex, given that there is the added factor that initially the touch or the abuse may feel good and the child may feel a partner in what is going on. Children will, therefore, find it difficult to make distinctions between what is a good and a bad touch, forcing us to look at the whole issue of whether attempts should even be made to try to teach young children these concepts’. (Mayes et al., 1992)

This dilemma signals an urgent need for researchers to get to grips with the complexity and contextualisation of safety rules and we move to a description of some intriguing work by Warden et al. (1997) evaluating the Kidscape safety training programme. This programme was designed to test out how learning rules ‘generalised’ beyond the content of the programme. Warden’s team presented a whole range of ‘safety situations’ for subjects to assess, with the scenario of ‘being approached by a stranger’ being represented in four picture-stories:
Further variation in the drawings was created by the presence or absence of other participants or witnesses to the interactions. ‘This manipulation was introduced to examine the extent to which children discriminated in their perceptions of the relative safety of situations. Thus, for example, in the ‘Cinema’ situation (Figure 3) there are five variations to the situation: (i) the child is alone, hereafter the alone version, (ii) the child is accompanied by a friend of the same age, the friend version, (iii) the child is accompanied by his/her mother, the mother version (iv) the strange man is accompanied by a woman, the confederate version and (v) the strange man is accompanied by a woman and two children, the friendly witness version’.

Further depth was added to the encounters using a short ‘contextualising story’ which included variations to draw attention to the bystanders. When narrating these stories, the interviewer referred to the child in the story as ‘he’ or ‘she’ to coincide with the sex of the child being interviewed. The drawings of the ‘cinema’ situation were thus introduced according to the appropriate version of the following narrative:

‘Here is a girl (boy) walking past the cinema (with her mother/friend). This man (and woman/and their two children) is standing outside, and he (the man) starts talking to the girl (boy). The man is very friendly; he tells the girl (boy) his name and says, ‘Have you seen this film? I hear it’s really good. Would you like to come and see it with me (us)?’

The picture-stories were used to measure children’s ‘safety assessments’ within an experimental trial. The experiment was carried out in two age groups, 6 year-olds and 10 year-olds and in three stages. Experimental subjects were tested pre-programme, 4 weeks after training, and 2-3 months after training. Matched controls were tested at the same intervals, without the application of the training programme. Safety judgements were made using coloured plastic discs to categorise the situations as safe (green), uncertain (yellow) or dangerous (red), as well as by some open-ended questions on what the boy/girl should do.

Figure 3 Variations of a stranger approach: invitation to the cinema

Source: Warden et al., 1997
The results of the experimental trial were not as expected. The Kidscape groups did increase in caution across trials, but ‘trained and untrained children did not differ in the extent to which their judgements changed across the three test situations; in other words both the trained and untrained children became more wary across the three testings.’

Much is to be learned from these negative and surprising findings and Warden et al. (1997) apply some instructive hypotheses in this respect. The improvement of both experimental subjects and controls is explained in terms of the ‘learning effect of testing’. The safety assessments themselves are long, complex, subtle and therefore instructive. The investigation as a whole concerns itself with an arena in which there is already considerable activity; most schools will have introduced some form of safety-awareness into their existing routines and parents, police and media will have brought the topic to the fore periodically. This rather engaging ‘red, yellow, green’ exercise may be supposed to have enhanced this flow of discussion, so that the research instruments themselves must be considered as teaching tools of substantial significance.

A second feature of the results, the pattern of judgements across the various safety scenarios, is also informative. Children’s safety judgements varied considerably as, for instance, across the five, stranger-approach, picture-stories. Aggregate results, combining all trials and all groups, are summarised as follows (% perceiving situation as red-for-dangerous) - ‘alone’ 75%, ‘confederate’ 74%, ‘friend’ 73%, ‘friendly witness’ 56% and ‘mother’ 28%. There are, of course, no ‘correct’ answers to be supplied in respect of situations, which are still basically imaginary and open-ended, but this particular pattern does display a satisfying perspicacity on the children’s part about the variations in stranger-approach. Some other patterns were rather more disconcerting. One of the picture-stories was about adults demanding a kiss/hug in return for a favour and the responses. These children’s responses to situations involving physical intimacy with an adult, already less cautious, became even more unwary when an adult imposed secrecy, as though the collusion of the secrecy contract somehow imbued the interaction with safety.’

‘Particularly noticeable was the disarming effect that secrecy imposition had on the younger children’s responses. These children’s responses to situations involving physical intimacy with an adult, already less cautious, became even more unwary when an adult imposed secrecy, as though the collusion of the secrecy contract somehow imbued the interaction with safety.’

A third, and again unanticipated, feature of the results, was a considerable difference in the safety-judgements between experimental and control groups at first assessment, with the so-called ‘control children’ usually demonstrating less caution. Warden et al. (1997) discuss this in terms of the fearsome difficulties of matching schools in their inquiry. Their preferred solution would be to obtain a list of volunteer schools for programmes such as Kidscape and then operate ‘waiting-list randomisation’ to rid the comparisons of subject differences. But as we have previously argued, such information on preliminary differences in the child’s perception of danger is vital in our understanding of whether programmes will work. According to Warden et al. (1997), the essential difference between the experimental and control schools was a ‘willingness to participate in safety-training’. So what is being uncovered here are institutional and community differences in understanding, awareness, ability and urgency in dealing with the problem. The key point is that these selfsame differences in understanding, awareness and ability will show up in the children’s heads. What Warden’s research suggests, quite inadvertently, is that the very locations and the very children needing more programmes (and research and subsequent discussion) might well be the ones that normally will get less.

We conclude this section with a brief selection of findings from further studies exploring this matter of how children interpret and discriminate between safety-rules. The first comes from another study by Warden, with a different set of colleagues (Moran et al., 1997). No ‘programme’ was involved in this research, instead 6, 8 and 10 year olds were presented with a series of video sequences depicting various encounters between children and adult males. These professionally-recorded scenarios exhibited a series of ‘stranger-danger’ permutations by manipulating ‘location’ (school, street, private-garden), ‘degree of familiarity’ (stranger, very familiar, unknown but familiar) and ‘type of approach’ (request, offer, demand). Each encounter was filmed up to the point where the child-actor had to make a decision and the research data was obtained by asking the subjects ‘what will the boy/girl do?’ and ‘why?’ The results, again, showed significant differences in rates of ‘compliance’ according to details of the scenario and age of the respondent. There was also considerable variation in the reasons given for the decision - with thoughts about ‘stranger-danger’ being balanced in the children’s minds against considerations about ‘what the child might gain’, about ‘the need to be kind to others’, about the importance of ‘taking responsibility for one’s own actions’, about the likelihood that ‘parents know where their children will be’ and so on. The researchers are not able to trace back to the source of these decision dilemmas but we suspect that differences in patterns of upbringing and socialisation would establish quite different norms in the way sub-groups of children would respond and reason.

Further differences in the understanding of safety-instruction in 3-8 year-olds are uncovered in research by Coppens (1986). This study used photographs of safe and unsafe situations to chart the children’s knowledge. She found a steady improvement with age in the responses, but whilst more than half of the six to eight year-olds had a perfect ‘safety’ score, only one child in the group had a perfect ‘prevention’ score. Similarly, in a study of risks likely to cause personal injury, this time for 9-12 year olds, Dickson and Hutchinson (1988) show that, although the group were aware of such risks, they were unclear as to how they would respond in practice.

All of the studies in this section show that children do not just ‘follow’ rules but they adapt them to existing understanding. It is quite impossible, therefore, to teach them how to behave in every single situation they
encounter. The various ‘scenarios’ in the above inquiries pick out and manufacture via sketches, likenesses and video-tapes just a few features of a real situation. In the real world, of course, many, many other cues, prompts, distractions and enticements would confront children attempting to follow a rule and may override that rule.

This is a tough lesson but no counsel of despair. Moran and colleagues draw the following moral about programming from their study:

‘to learn to keep themselves safe, children must be given the opportunity to explore, through discussion and role play, a wide variety of safety situations and their potential risk and resolutions .... educationalists must remain alert to the potential conflict children may experience in learning to keep themselves safe from harm and a socialisation process which exhorts them to be kind, helpful and obedient.’

The lesson for the researcher is that they must unearth and explain the patterns whereby rules are sometimes assimilated easily and sometimes with difficulty. The focus should be on disjunctions as above, which explain that children have adapted more easily to rules about ‘stranger-danger’ than to those about ‘physical intimacy’ or ‘secrecy pacts’, that ‘demands’ from strangers are harder to cope with than ‘requests’ or ‘offers’, that ‘knowledge’ of what is unsafe often outstrips understanding of ‘response’. All of these patterns will themselves have sub-variations according to the backgrounds of the children and it is for the researcher to track down ambiguities, blind-spots and contradictions, so that the teaching of rules can be adapted to accommodate them.

**Lesson 6.** Programmes which aim to teach children sets of behavioural codes/rules/guidelines have to appreciate that children do not just ‘follow’ rules but adapt them to their existing understanding. These processes of assimilation vary because certain rules will clash with other expectations created during their upbringing. The researcher’s task is to uncover and explain the patterns of acceptance and resistance to rules.

### 3.7 Normative change: Peer-pressure and clubbing together

Partly as a reaction to the difficulties of teaching via codes and the possibility that imposition of ‘rules from above’ will be met with resistance, another class of programmes with children attempts to use the mechanism of ‘peer-pressure’ to get the message across. Young people are generally quick to share personal experiences, and health educators, in particular, have tried to harness this naturally-occurring process as the engine of change in social programmes as follows:

‘Peer health education is the teaching or sharing of information, values and behaviours by members of similar age or status groups.’ (Sciacca, 1987)

This is yet another ‘big idea’ which has spread its roots across the policy-fields and before evaluating its potential, it is appropriate to acknowledge its diversity:

‘Despite its widespread use, it is clear that peer education means different things to different people. While peer education has generally been taken to mean education of young people by young people - that is with the emphasis being on equality of age - it is important the word peer should not be too narrowly defined. In a number of projects, for example, the use of ‘peer leaders’ clearly implies older rather than same age ‘peers’ .... On occasion, peer is also taken to mean - as is its literal meaning - equal status. Thus, the term has recently been used, inter alia, to describe crack-user-lead research into crack-use, drug-user-led education for drug-users, and the use of drug-using prostitutes to educate other drug-using prostitutes’ (Shiner and Newburn, 1996)

As peer-education projects multiply, so do the number of theories about why they might work. Milburn (1995) provides us with a useful guide to ‘rationales’, which can be summarised as follows:

- **Peer-interactions** tend to be more frequent, intense and diverse than with other people - and thus provide an arena for empathy and imitation.
- **Peer social networks** tend to be protective - and thus to offer conditions of co-operation and mutual support.
- **Peer education** is participatory - ‘helpers’ and the ‘helped’ give and take, and both are deemed to grow and be empowered.
- **Peer-education** depends upon the similarities and shared-identities of influencer and recipient - and this increases the persuasiveness of any message.

Whether peer-education ‘works’, depends on which, if any, of these mechanisms is unleashed in the activities of the clubs, groups and outreach programmes promoted in its name. The suggestion here is that in some contexts the sharing of identities will be beneficial and in others it will be a hindrance. Some clubs will offer many opportunities for mutual support and in others these will be overwhelmed by out-of-club activities. Some matters may fall under the understanding of peers and some will not. The goal, once again, is for evaluation to trace the patterns of accord and discord and show how these find their way to programme outcomes.

The task can be demonstrated with two evaluations, which we consider half-finished in respect of this goal. **Smokebusters** is a much-imitated initiative based on forming clubs whose purpose was to provide a peer-group who felt positive about not smoking and would provide mutual-help to members in developing the confidence to refuse cigarettes. The Grampian Smokebusters club was formed in Scotland in 1987, inviting all 10-13 year olds in the region to join and for £1 they received a membership card, discount card, stickers, posters, newsletters, discos, social events - all promoting a ‘stop-starting-smoking’ message. The programme was evaluated by van Teilingen
et al. (1996) using a one-in-ten, sample survey of the target population. The results demonstrated the popularity of the club, which attracted half the eligible children, and its failure - for after four years ‘smoking prevalence in the Grampian region was similar to Scottish levels and former members were just as likely to have started smoking as those who had never joined’.

The researchers explain the results in terms of David being swamped by Goliath. Due to financial constraints only a relatively small number of members could be admitted to the events. Due to the wide geographical catchment of Grampian Smokebusters it was ‘likely that only 15 percent of all members ever attended such events’. What is more, ‘the club did not exist in a social vacuum, as the children continued to live in an environment where adults and older children smoked, where tobacco advertising continued, and where it was still possible, despite legislation, for children to buy cigarettes without great difficulty’.

It is hardly surprising that peer-pressure could not be sustained by newsletters and discount cards and we note that the programme theory has not really been put to test in this particular evaluation. There is, however, a missed opportunity here, for some of the Smokebusters members will have attended a wide range of events and will have had the opportunity for empathy, modelling, mutual support, give-and-take, partnership etc. What we needed to know from this inquiry was how smoking rates varied with participation. The authors respond somewhat weakly that ‘a more detailed process evaluation might have been helpful in defining the more effective components of the analysis’ and that ‘further studies should be designed to establish how young people associate themselves and others with Smokebusters’. Really though, the opportunity is there in the outcome data, since the survey could have easily tracked ‘attendance levels’, ‘length of membership’, ‘card usage’, ‘friendship networks’ etc. and correlated these with smoking behaviour.

Another example of the subtleties of peer education comes in a Home Office drugs education programme aimed at children (Youth Awareness Project, YAP). The research was formative in nature and consisted of visits to YAP workshops and youth clubs, during and after which subjects were interviewed at length about their experiences. The focus of Shiner and Newburn’s (1996) evaluation is on the ‘credibility’ of the YAP workers who offered advice to stay clear of drugs:

‘The study identifies and examines three dimensions of credibility: person-based credibility, arising from a perceived (and often shared) personal characteristic, for example age, sex or ethnic origin; experience-based credibility, arising from a perceived (and often shared) experience, for example drug-use, contact with the criminal justice system or bereavement; and message-based credibility, arising from what is said and how it is said.’

The study uncovered the importance of the latter two types and suggests that the usage of age-equivalent peers is of little value in the particular field of drugs education. Those youths who have made the first steps down the road to misuse are particularly clear that direct experience of drug-use on the part of the advice-worker is a necessary part of providing a credible drug information service. On which matter, one participant fumes:

‘I hate people who mock, who say things and they’ve never done it, and they’ve never been through it, and never know what they’re going on about. I hate people like that, it’s just like I just go ‘no, forget it’. I walk out if I find someone like that, never done something like it, something I don’t like’.

Whilst such findings are an important part of programme building, they remain, in this form of research, only conjectures. Shiner and Newburn are able to show that the message from such a source has credibility, since the subjects also reported clear understanding of the ‘harm reduction’ emphasis of YAP and a liking for not having been issued the pointless instruction (for them) to ‘just say no’. But understanding and a liking for a message are one thing, acting upon it is another. The ultimate question of whether YAP subjects go on to actually reduce harm caused by their drugs intake or get off drugs altogether is not addressed in formative research of this sort. The possibility always remains that subjects have ‘acted up’ their fondness for message and messenger rather that ‘acted upon’ them. These distinctions of types of credibility are nevertheless an important pointer to researching peer education. Different subjects will find different approaches credible, and if good outcome data can be obtained, it is important to discover which combinations of subject and message-bearer work best.

Further variations in the circumstances in which ‘peer-pressure’ and ‘clubbing-together’ work (or not) are brought home by two further brief examples quoted from Milburn (1995), which we recommend as an excellent overview of the ‘peer-education’ literature:

‘West and Mitchell (1995) conclude that the case for coercive peer pressure as an influence on adolescent behaviour is not proven, and is much more complex than studies indicate. For example, they suggest that peer influence may operate differently with respect to different health related behaviours. Furthermore, they argue that peer influence may, in fact, only be seen to have any discernible effects on more ascribed or visible adolescent health issues, such as being overweight. Understanding the social context of health relevant behaviours and their associated concepts of identity are vital to the analysis of these peer processes’.

‘De Paulo et al. (1989) explored issues of age differences in reactions to help in a school based peer tutoring context. They pointed out that there may be self-threatening implications of receiving help which could produce negative and defensive reactions on the part of the recipient. They found that amongst children of 8-10 years help was more threatening in dyads in
which the children were similar to each other in both age and achievement. There were better supportive outcomes in dyads where tutors were ‘older and smarter’ than their tutees. This underlines the importance of understanding perception of the peer educator by the recipient in any peer intervention’.

Levels of involvement and participation rates might well be fragile in the more ‘manufactured’ clubs - people do not know what is involved in an anti-smoking or road safety club in the way that they understand the purpose of a bingo or cricket or youth club and this ambiguity may dilute effectiveness. Effectiveness will also vary according to who and what is the mainstay of club events and in ‘multi-activity’ clubs such as Smokebusters it is hard to perceive which elements have credibility.

**Lesson 7.** Programmes which aim to inform children via ‘peer-education’ and ‘club-activity’ have to anticipate wide variation in the ‘credibility of’ and sense of ‘belonging to’ such social formations. Recipients will also differ in the extent to which they find support and shared-identity. Researchers should anticipate large amounts of internal variation in the success of such networks and concentrate on discovering the successful combinations of message, messenger and subject.

### 3.8 User-involvement and peer-research

In the previous two sections we have stressed the importance of having a close understanding of the inner workings of an initiative and for researchers to trace how programme theories are actually translated into the ideas and choices of subjects. One strategy of evaluation research has taken this lesson closely to heart and places the subject’s hopes and expectations at the very core of programme implementation, development and research.

We were able to trace this shift in the basic focus of interventions in several sub-disciplines throughout our review. In the mental health field ‘de-institutionalised policies and an increasing emphasis on social treatment, give greater impetus to the need for evaluation studies to find more empowering ways to involve service users’ (Drewett, 1977). Drewett’s paper goes on to give a useful review of the strengths and weakness of various measures (such as ‘health panels’ and ‘citizen’s juries’) designed to give greater voice to the adult user. Rather closer to our interests here is the spawning of a variety of projects seeking to ‘increase the control that young people have over their own lives and to increase the involvement they have individually and collectively in the decision making of agencies, institutions and systems in the provinces that effect them’ (Kearney and Keenan, 1988). France’s (1996) research on a UK Youth Action Project designed to reduce young people’s involvement in crime provides thoughtful reflection upon the fine line that exists between ‘empowerment’ and ‘exploitation’ in attempting to hand control to those with no particular stake or trust in the system.

Our concern with such initiatives is, as ever, based in the methodology. Such participatory programmes are usually partnered with a research strategy known as ‘formative evaluation’. Gittelsohn (1998) supplies us with the rational as applied in her study of Pathways - a Native American family-based, anti-obesity programme:

‘Formative information gathering is essential whenever developing health interventions in new cultural settings where information is limited. People’s understanding of their health and what makes their bodies healthy is embedded in personal experiences, and culture-based meanings and belief systems. Formative research methods help unlock these meanings and gain insight into developing relevant intervention strategies’

Often, the second string to this particular research bow is to involve programme participants as ‘peer-researchers’. This goes a step further than the use of peers in the delivery of an intervention (as described in the previous section). Co-opting users in the actual research process, rather than them being ‘passive’ subjects, is claimed to bring verisimilitude to the findings. The idea is that subjects are consulted in every stage of the project from sampling, through data collection and analysis, to dissemination. In practice this often means incorporating young people as members of the ‘research steering group’ as was the case in the East Surrey, Stepping Out project. In this six young care leavers joined researchers and health and social service staff in a project evaluating the needs of young people leaving the ‘looked after’ system (Broad and Saunders, 1998).

We arrive at the key and contemporary question - how far should children’s views be taken into account when evaluating a programme designed for them? We steam into somewhat contentious waters here since although it is acknowledged that children’s competencies are not the same as those of adults, there is no agreement on how these differences constrain initiatives and their research.

We begin by noting that there are any number of examples of projects in which children take part in the decision-making process and take on responsibility for themselves and others. A good summary of these participatory initiatives is to be found in Miller’s Never Too Young (1996). This Save the Children handbook lists many examples of routines, sessions and whole programmes based on incorporating the child’s viewpoint. One example is to move from nursery ‘meal provision’ to a model that allows children to make choices:

‘At the Wood End Family Project in Coventry, crèche staff set up a drinks and fruit bar, available at all times, so that the children could decide for themselves when to have a drink and something to eat. There were 20 children in the group, aged 0 - 4 years. At first the children kept asking if they could have fruit and water rather than helping themselves, but they soon realised that it was there for them to have when they wanted. Initially the children kept spilling water, but as they
got used to pouring they became proficient and less was spilt. The children helped staff to wipe up spillages when they occurred. The children told the staff that they really enjoyed eating the fruit. Staff commented that the activity went very well and feel that in time the children will get used to this as a regular part of the session.' (Miller, 1996)

Another illustration, this time with older children, is contained in Hart’s (1994) review of children’s roles in primary environmental care. Activities at the Notting Dale Urban Studies Centre in London carry the trade mark of ‘user evaluation’ in that children not only learn but engage in programme research and development. The idea is for the children to add to the archival and survey resources of the centre, helping it to become more and more a community resource:

‘Groups of children from the surrounding schools go to the Notting Dale centre to conduct investigations about the local environment. The children go out with tape recorders, cameras, pencils and paper to document existing conditions. They interview residents, local officials and people who work in the neighbourhood. On their return to the Center [American spelling in original], they transcribe recordings, print photographs and type reports. Materials assembled by previous groups of children are retrieved from files for reference and comparison. Teachers and Center staff assist children with their tasks, engage them in discussion and offer guidance in making decisions when requested. Working together in small groups, the children sift through, discuss and interpret their materials, and put them in the form of a newspaper to take back for printing and circulation around their school. Over the course of time, much material has been collected by groups working at the Center. It has become a repository of perceptions about the local environment through the work of young residents. In effect, the Center houses a vast array of archives - statistics, minutes, briefs, case studies, correspondents, newspapers, and the students’ own documents - that describe life on the public housing estates in the area and reflect upon environmental changes, why they are happening and how they could be improved.’ (Hart 1994)

We have quoted these two examples at some length because we detect a tinge of the rose-coloured lens in each description. Both are driven by the conviction that programmes for children work better if youngsters are consulted about the development of the initiative and are given scope to participate in its running. This is a principle with widespread currency and we have little doubt that it is applied to many road safety programmes and could be applied to more. The inclusion of a bit of local ‘savvy’ about what children actually get up to on their journeys to school could add realism to such projects.

Promising a prospect as this is, we regard the gains mooted in the above examples as untested hypotheses and we suspect that a failure to recognise this has become a feature of some applications of the participatory approach. Relinquishing ‘ownership’ of a programme becomes an end in itself and the evaluation becomes somewhat messianic in establishing that the children have a voice. Our view is that programmes are always locked into multiple-stakeholding, being shaped by many ideas, all of them with the potential to be fallible. Thus of the ‘drinks and fruit bar’, we would indeed want to know if it increased the choice and independence of the children, but also if there were any other unintended effects apart from ‘spillage’ - such as the need to guard against choking amongst the very young, the need to prevent eating slices dropped onto the floor or the replacement of unfancied, chewed bits to the bar, the need to settle disputes when stocks run low or when a favourite fruit is featured, the need to convince parents who do not allow such free access at home, and so on. Of the ‘Urban Studies Centre’, we would need to check out - how many children survey and how many skive, how many initiate and how many are directed, whether the community practised what the archive preached, and whether the local residents think ‘how interesting’ or ‘not again’ when approached by the tyro investigators - before we jumped to Hart’s conclusion.

Broad and Saunders’ (1998) aforementioned research on the needs of care leavers provides what we consider to be a better balanced view of the promises and pitfalls of young people’s participation in formative evaluation. Their paper ends with the classic ‘benefits and costs’ listing, which is alas too long to reproduce here, so let us place the dilemmas in the head of the single novice researcher. He or she might well generate a more trustful response from peers, might uncover points otherwise unspoken, might galvanise a sense of conviction about the research, and might be enormously persuasive in embodying the research findings when it comes to their dissemination. The same young person might, however, face emotional costs in confronting his/her own problems on a wider scale, struggle with some of the technical demands of research however much these were simplified, forefront personal understanding over that of other youngsters who become their subjects, and be ill-equipped to discern the longer-term and broader-range implications of their findings.

For us, then, ‘partnership’ is just one of a number of potential mechanisms through which a programme may work, and one which stands alongside ‘leadership’, ‘instruction’, ‘care’, ‘discipline’, ‘peer-education’, ‘play’ and ‘affection’ as an engine of success, and one with the capacity to foul-up as well as free-up a child’s imagination. To be sure, too many programmes are aimed at children, rather than being conducted with children and too many evaluations have regarded children as subjects rather than stakeholders. Nevertheless, there comes a point where the pendulum can swing too far and research becomes advocacy rather than evaluation. Formative research is a relatively popular and low-budget method. It is an essential part of the evaluation tool-kit since getting together and sorting out day-to-day problems just about exhausts the evaluation ambitions of many initiatives. There is however this tendency, when the method goes
public and into print, for such evaluation to stress certain voices and particular coalitions. We conclude, appropriately enough, with a warning from the major advocate of children’s participation, which could have been written with evaluation in mind:

“The views of children will need to be balanced against those of others involved and against such considerations of finance, health and practicality” (Miller 1996).

Lesson 8. Programmes which aim to establish children as ‘partners’ in the development of the daily activity which makes up an initiative have to establish the boundaries of the child’s expertise. These programmes tend to be researched using a ‘formative approach’, which focuses on the day-to-day interpretations of those involved. This method, however, often lacks range in terms of coverage of viewpoints and the ability to trace outcomes. It frequently ends in advocacy rather than evaluation.

3.9 Identifying risk

In this and the following three lessons, we begin exploration of the issue how the researcher is best able to discover ‘for whom’ and ‘in what circumstances’ a programme works. Two concepts that are much used in relation to this aim are the notions of ‘risk group’ and ‘target group’ and we begin with a series of commentaries upon these from the literature. The most common theme we discovered was fierce rebuttal of ‘single-cause, single- upon these from the literature. The most common theme we discovered was fierce rebuttal of ‘single-cause, single-solution’ assumptions about problems and programming. Risks were generally seen as multiple and, more to the point, as being located through the different layers of social reality, being seated in dysfunctional ideas, individuals, institutions and infrastructures (recall lesson 1). Programmes developed specifically for ‘high risk’ children often located their target group via a catalogue of problems. For instance, St Pierre et al., (1997) evaluation of a drug-prevention scheme follows the initiative’s own definition of the eye of the storm: ‘Poor bonding between children and parents, low parental involvement in activities with children, maternal isolation, poor family management practices and family history, and parental approval of drug use’. Ross’s (1992) evaluation of an after-school programme for ‘latchkey’ children defines these youngsters in terms of, ‘Lack of parental supervision (especially during after-school hours), solitary risk taking, economic depression, absence of moderating extra-curricular activities, lack of social influence (in relation to parents social spheres in particular), lack of homework completion and other problems in school performance, stress indicators (including anger and isolation)’.

Acknowledgement of the complexity of the problem-base is by now commonplace, risk has been added to risk, and several analyses and meta-analyses (Lazenblatt, 1997; Culbertson and Schellenbach, 1992; Utting, 1996; Hawkins, 1992; Bruvold, 1993) have scoured the social problems literature in attempting to give an overall picture of risk factors as they apply to their own domain. Table 3, adapted from Dryfoos, provides an overview. The author considers that risks, or ‘antecedents’ as she calls them, can be spilt into four main domains.

Table 3 Framework of antecedents of adolescent high-risk behaviour

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Household composition</td>
</tr>
<tr>
<td>Sex</td>
<td>Income, poverty status</td>
</tr>
<tr>
<td>Race and ethnicity</td>
<td>Parental education</td>
</tr>
<tr>
<td>Personal</td>
<td>Parental role bonding</td>
</tr>
<tr>
<td>Expectations for education</td>
<td>Parental practice of</td>
</tr>
<tr>
<td>Perception of life options</td>
<td>high risk behaviours</td>
</tr>
<tr>
<td>School grades</td>
<td>Culture in home</td>
</tr>
<tr>
<td>Conduct, general behaviour</td>
<td>Community</td>
</tr>
<tr>
<td>Religiosity</td>
<td>Neighbourhood quality</td>
</tr>
<tr>
<td>Peer influence</td>
<td>Segregation</td>
</tr>
<tr>
<td>Peer use</td>
<td>School quality</td>
</tr>
<tr>
<td>Conformity-rebelliousness</td>
<td>Employment situation</td>
</tr>
<tr>
<td>Involvement in other</td>
<td></td>
</tr>
<tr>
<td>high risk behaviours</td>
<td></td>
</tr>
<tr>
<td>Psychological factors</td>
<td></td>
</tr>
<tr>
<td>Self-esteem</td>
<td></td>
</tr>
</tbody>
</table>

Adapted from Dryfoos, 1990

In long lists lie uncomfortable lessons. The range of the issues which gives rise to problems stands as a rather humbling measure of what can be done in response. Programme designers need to inspect closely which of the above factors is the source of the problem under scrutiny and which of the elements they may be able to influence in the course of an initiative. The answer, typically, is that most of these factors can be said to influence the problem and very few are in the control of the programme. If we consider road traffic victims as a case in point, it is easy to see that accidents are demographically unbalanced and occur much more in specific types of communities, that victims and culprits are more likely to be of a certain character and share certain personal traits, and tend to come from particular family, educational and social backgrounds (O’Reilly, 1994). And it is impossible to expect one-off initiatives to get to grips with them all.

Across the various policy-fields we have examined, such thinking has hastened calls for comprehensive multi-level, multi-agency programmes. Our reading also indicates, alas, that having multiple-goals and multiple-agencies often leads to competition and confusion as well as co-operation (Bloxham, 1996; Kemp et al., 1998). Thus the lesson we tend to retain from typologists of the multiple-antecedents of social problems is that one has to understand what aspects of the overall problem that a programme can and cannot confront. Multiple risk factors always apply but they don’t apply equally to all programme subjects. It is very easy, as we shall see, for programme architects to mistake one risk for another. Programmes need to prioritise and research needs to help it prioritise. The
key task of evaluation is thus, in a memorable phrase we owe to Pease, ‘to ensure that the grease gets to the squeak’ (quoted in Hough and Tilley, 1998).

Lesson 9. Programmes should aim to identify and counteract ‘risk factors’. But the antecedents of social problems are always multiple. Evaluation thus has the task of understanding what risks a programme can and cannot confront, and to ensure that the tractable problems are pinpointed and attacked. The key priority is to get the grease to the squeak.

3.10 Targeting programmes

The human equivalent of getting the grease to the squeak is to target programmes at appropriate subjects in the appropriate situations. Customarily, interventions are built with at least approximate notions of type of persons and circumstances that may benefit from the activities of in the programme, and one of the key tasks of evaluation is to refine that understanding. Some programmes, of course, have near universal targets - such as general road safety education for children. Some are risk-focused - such as drug-education programmes for the ‘latchkey youth’s’ as defined above. But in both cases, good evaluation follows down the same path of discovering how the programme has succeeded for particular sub-groups and sub-situations. This section examines examples showing the importance of the progressive focusing-down on programme targets. Knowledge of ‘for whom’ and ‘in what circumstances’ an initiative will work builds from day-one of a programme. Ideas on initial target-groups are assembled from the working knowledge of practitioners, from the demographics of a problem, from existing social and psychological theory, and from common sense.

All these sources have the capacity to be built-upon, which we illustrate with the seemingly mundane example of ‘litter’ control. Common-sense tells us of a sub-group and a sub-situation which will cause the problem. We reckon that we knew that children litter more than adults and that littered areas attract more litter. It came as something of a surprise to us that such matters were on the receiving end of research, but it transpires that both of these mini-theories have been put to refinement (La Hart and Bailey, 1975; Huffman et al., 1995). Adults do in fact litter less than children and are indeed more responsive than children to the introduction of signs in public places requesting proper disposal of the stuff. This rule does not apply uniformly across either group, however, for when people congregate in groups rather than in twos and threes, they litter more (per head) and notice the signs less. Litter does indeed attract litter, for when an area is seeded with litter more is dropped, and when an area is cleared subsequent littering is much reduced. Even the intermediate act of sweeping the rubbish into piles, it seems, is enough to make a positive difference. Children do not respond much to normative appeals on litter but they do go for incentives. The simple device of placing deposits on glass bottles achieved the prized goal in littering policy of having children deal with other folk’s rubbish (coke bottles) as well as their own. Giving litter a value has its limits however. La Hart’s research recounts the sad tale of a nature trail in which wardens introduced some secretly marked litter on the pathways. The incentive was that children should collect as much rubbish as they could and those who discovered the special ‘seeded’ pieces would win a prize. As anticipated, the children entered the competitive fray with relish and much litter was collected. Knowing children, alas, soon realised that their own lunch-box litter was not in with a chance and often dumped it so that it could be replaced with the stuff with potential value. The example is a microcosm for evaluation. We begin with some basic ideas about people and places which may be prone to a problem and the research provides a sharper focus.

What is true for litter also goes for safer sex. Close discernment subject characteristics and circumstantial features also figure in Buyssse and Van Oost’s (1997) evaluation of a sexual risk reduction programme. This Belgian programme aimed at adolescents covered ‘safer sex negotiation’, ‘resistance to peer pressure’, ‘knowledge about condoms’, ‘HIV/AIDS awareness’ and a ‘practical’ (in condom purchase). A quasi-experimental design was used but, again, the results with policy-implications are to be found in differences within the programme subjects.

The researchers point out that since ‘safer sexual decisions are to some extent influenced by perceptions of socially expected behaviour, it is expected that subjects differing in gender role expectations, differ in intention in relation to safer sex’. Thus they anticipated that the need for, and reactions to, such a programme would differ according to the ‘gender scripts’ internalised by different sub-groups, four of whom were identified as follows:

‘(1) subjects with an intolerant expectation of both male and female condom-related behaviour, labelled ‘conservatives’,(2) subjects who have internalised the traditional gender role script with it being acceptable for males and unacceptable for females to carry and introduce condoms, labelled ‘traditionalis’, (3) subjects with inverse role expectations with it being acceptable for females and unacceptable for males to carry and introduce condoms, labelled ‘non-traditionals’, and (4) subjects with tolerant expectations of both male and female condom-related behaviour, labelled ‘permissionis’.

The programme as a whole had a liberating effect on discussion about sexual matters, after the sessions there was markedly ‘more frequent safer sex-related communication with friends in treatment subjects’. This effect was not, however, found for ‘traditional’ treatment adolescents. These youths expect the male to take the lead in sexual liaisons, the sexual script of these (‘regular’ guys / ‘shyer’ girls?) rarely turns to condom talk and the programme, it seems, did not make them wish to do so. The programme was valued highly by the treatment group as a whole for its ‘information provision’, but once again this rating was not shared by all, with the ‘non-traditionals’ registering ‘dissatisfaction’, wanting more information on safe sex practices other than male condom use.
In short, what we have here is a programme which, whilst aimed at technical information and interpersonal information, is in fact mediated by different moral climates, being somewhat in advance of traditional views and somewhat behind the thinking of non-traditional adolescents. We note, in passing, that the research itself operates with a ‘sexuality script’, the four roles identified all been heterosexual. A finer identification of sexual preferences would probably unearth even more divergent views about such programmes. The general moral for evaluation, however, is that programmes will always carry social assumptions which will meet with acceptance or resistance. The decision, for instance, about whether road safety lessons should give advice on making difficult crossings or recommend avoiding them altogether is not just a technical matter for the practitioner. It will actually feed into the success or otherwise of the programme, according to whether the child comes from a highly protective environment or one that stresses independence.

Before leaving ‘gender scripts’, their almost ubiquitous influence on the programmes under review should be noted. For instance, Perry et al. (1998) discover of a Minnesota version of the 5-a-day fruit-and-veg initiative:

‘Girls appeared to be more receptive than boys to increasing their vegetable consumption. Girls have been shown to be more receptive to other health education programmes concerning eating patterns and physical activity. Since dieting is far more prevalent among female than male adolescents, perhaps the intervention, even without any low calorie message about vegetables, heightened interest in dieting - and thereby vegetables - even in our preadolescent population’.

This section highlights one of the key tasks of evaluation research. New knowledge should be incorporated into later programmes and the more difficult targets attacked progressively. The ‘safety’ field always faces loaded dice in terms of ‘which children are at risk’ and ‘when are they at risk’. Concerning drowning for instance, Towner et al. (1996) note, ‘The under-5 age group is the most vulnerable … Boys over 5 years are much more likely to die as a result of drowning than girls and the [gender] gradient is even steeper for older children .... Older children (5-14) and particularly boys are most at risk during warm weather when exposure increases, this is reflected in the large difference in mortality figures in different years’. These are merely the starting points. Programmes only start to reload the dice when the risks within the risk groups are identified.

Lesson 10. Before any intervention or programme commences there should be intelligence on the persons and situations at risk. And after any evaluation such understanding should be refined considerably. This information is best gathered by a close analysis of subgroup response to the initiative.

3.11 Age groups

‘By age three, approximately one-third of US children could correctly identify ‘Old Joe’, a cartoon camel, and link him with cigarette use’, Hall and Zigler (1997)

We turn to the most examined of all ‘sub-groups’ in the research under review. Discovering whether programmes met the needs and capacities of children of different age groups was an obvious preoccupation amongst many, many of the evaluations we encountered. The basic push here is the notion of ‘developmentally appropriate practice’. In this instance, there is more than a hotchpotch of theory, practitioner-wisdom and common-sense to guide the programme builder and evaluator, since there is a whole ‘discipline’ of child development devoted to the question, ‘at what age can we expect the children to ….?’

For once, we begin at the end, for there is an extensive literature applying developmental psychology to children’s road safety education, most notably Thomson et al. (1996) study for the (then) Department of Transport. The researchers spell out their ambitions for the developmental perspective as follows:

‘To identify a hierarchical, age-graded progression of the skills involved in traffic behaviour, from simple to complex. Any educational programme should fit into an overall plan whose aim should be to develop road skills in a systematic and hierarchical manner, taking into account the age and previous experience of the child. Each programme should thus evolve out of the achievements of previous programmes and lay foundations for the next’.

The authors acknowledge that no comprehensive plan of this type exists at the moment but do make some important steps by way of showing, for instance, that in the early years practical training is more effective than learning to recite a code:

‘Skiing or swimming, driving or learning to ride a bike all require practical experience: no one ever learned any of these things by just sitting at a desk, yet this is precisely how we expect young people to learn how to cross the road. Meanwhile it seems that the practical experience that does turn children into skilled pedestrians is picked up on real roads and independently. .... The trouble with simply telling children to ‘look around’ before crossing is that, when these instructions are divorced from the reality to which they refer, they end up going through a ritual whose function they scarcely understand. However, children trained in timing judgements are much more likely to look before crossing because they can’t make timing judgements without doing so’.

Researchers in other areas have shared the goal of devising the age-graded progression of programmes, with what looks to our eyes as some rather topsy-turvy results. For instance, McConnell (1996) reports of a children’s fire-safety programme:
‘Perhaps the most surprising outcome of this study was the powerful influence of the Kid Safe program on the 3-year old children. During the conceptualisation of this project, the possibility was raised of excluding three year olds because there was scepticism about both their ability to profit from the lessons and their capacity for attention during the long periods of knowledge testing. Fortunately, these children were retained, and those exposed to the Kid Safe program raised their knowledge level concerning fire safety from 48.46 on the pre-test to 79.14 on the post-test. This gain of more than 30 points represents a 62% improvement. In fact, contrary to our expectations, 3-year olds showed stronger effects than either 4- or 5-year old children. As an indicator of how successful the program was with these children, it will be recalled that the Kid Safe program boosted the knowledge level of 3-year old children to a level above that shown by untrained children two years older’. By contrast, Buller et al. (1996) of a Sunny Days, Healthy Ways programme noted ‘Sixth-graders attitudes and self-reported preventive behavior were somewhat resistant to the curriculum .... [they] may be ready for more advanced content, but the curriculum is probably too advanced for kindergarten through third grade’.

Loescher et al., (1995) evaluation of another sun-safety programme provides a further degree of subtlety with the finding, in respect of four and five year olds, that ‘the curriculum affected knowledge and comprehension significantly, testing of the application component did not reveal significant improvement’. In other words, a knowledge gain was not transferred into practice, since ‘It is reasonable to assume 4- to 5-year old children need parental help in applying their new-found knowledge, given their developmental stage and the limited control they have in this area of their life (e.g., they are unable to buy sunscreen for themselves)’.

Detectable in these uneven patterns is what development psychologists refer to as ‘domain specificity’, what realists refer to as ‘contextual constraint’ and what sociologists refer to as ‘culture’. In other words, the stage at which children can take on board safety lessons is not solely determined by age (defined in terms of ‘age of their bones’ or ‘stage of cognitive development’). The utility of lessons also depends on a child’s access to hazards and to prevention strategies, which in turn might be associated with an age-profile somewhat different from its ‘development stage’. Children face the perils of traffic, smoking, drugs, poor diet, sex abuse, excessive sun at quite different ages (and sometimes never). Road-crossing practice is free to any child who is able to partake, but not so sunscreen, smoke-detectors, healthy diets, bicycle-helmets etc. The ‘age-appropriateness’ of programmes is thus also shaped by culture and whilst we would warmly endorse Thomson et al. (1996) consideration that ‘specifying a general theoretical framework is an indispensable requirement to develop effective and coherent educational countermeasures’, it is clear that this theory needs to incorporate elements of sociology as well as developmental psychology.

Before we leave the question of age, its influence on the research process itself must be given consideration. As we have suggested above, the evidence seems to suggest that programme effectiveness appears to be very sensitive to age. That evidence, however, has to be culled from investigations with children and they are, perhaps, not the easiest of research subjects. The issue of the validity and reliability of evidence coming from the mouths of children must be broached (see Hood et al., 1996). One daunting possibility is that the findings on age variability in programme outcomes are an artefact of the research act itself, with the data collection instruments registering differences because of what they are rather than what they ask. We need to pose again the familiar question of ‘at what age can we expect the children to ......?’ and complete as the final clause, ‘be research subjects’, ‘answer questionnaires’, ‘take part in focus groups’, ‘give accurate self-reports of their behaviour’, ‘provide attitudinal information’ etc.

We begin with an instant qualification to what has been just said. In certain, very important respects we would deny the premise just developed. Much of the vital apparatus of evaluation research has nothing to do with the point of contact between researcher and child. Getting the right basic strategy is all important in evaluation and issues such as whether one goes for ‘RCTs’ or ‘formative methods’ or ‘participatory methods’ or ‘realist methods’ or ‘theory-driven methods’ are not governed by the presence of children as subjects. Even some rather technical issues in research, such as choosing a sample, creating comparison groups, coding and analysing data and so on, are issues on which children need neither to be seen nor heard.

Nevertheless, there is a point in the research act where the voice of children becomes crucial. Programmes seek to influence the way children act. In measuring such outcomes it is indeed necessary for children to be heard and, even more preferably, seen. Our evidence from the hundreds of reports that we read is that evaluation research is still getting to grips with matching its methods precisely to children of different age groups. Once again it can be said that ‘no comprehensive plan of this type’ exists at the moment. We come to broadly the same conclusion on the basis of examining the general methodological literature on researching children. Dispute is the norm here with some arguing that naturalistic, qualitative, and even play-centred methods are necessary to enter to the world of childhood, whilst other camps claim that there is no social science method that cannot be adapted for children. A somewhat ‘balanced’ view is represented by the following:

‘Researching children’s lives is at an exploratory stage.... Some methods such as interviews and self-complete instruments are more suitable for older children. Participant observation and small group discussion are particularly suitable with five to six year olds. Small groups are ideal with children of the same age (peer groups) rather than different ages (sibling groups). Small group discussions allow children to set their own agenda and the research topic to be woven into children’s talk about their daily
lives and social worlds. The ‘draw and write technique’ and structured activities can be effectively integrated into discussion and interviews in order to provide a focus for children especially where the research topic is abstract and not immediately salient in children’s lives.’ (Mauthner, 1997).

Mauthner, however, is quick to point out that such age-related summary can never be absolute since the appropriateness of a technique is also determined by gender and ethnicity, as well the topic and the setting of the research. Old research hands, no doubt, would apply the same qualification to adults as well. We conclude that knowledge of the ‘age-appropriateness’ of research instruments is still at the trial-and-error stage - as it should be.

The key lesson on children as research subjects lies in the adaptation of instruments rather than foreclosing judgements about their suitability. The trick, basically, is to be much more ‘user-friendly’ and to present questions in frameworks and formats with which children are already familiar. We have already seen Warden at al’s (1997) adaptation of the familiar traffic-light motif in their usage of red, yellow and green discs to stand for different safety judgements. Attitude measurement, more generally, can be transformed from mundane old ‘Likert scales’ into ‘chuff charts’, with scale points of ‘agreement’ and ‘disagreement’ being represented as ☺ and ☹ etc. Similarly, in Williams et al. (1989) research on healthy-eating, ‘recall methods’ of food consumed on the previous day were much improved with the children drawing and writing about their intake in school lesson mode, rather than just trying to remember what they had eaten.

Basically, the motto of childhood inquiry seems to be, ‘make do and mend’ and we refer readers to the new generation of research methods text-books on childhood (Grieg and Taylor, 1999) for further practical examples.

Occasionally, children are rather more comfortable with a data collection technique than adults. Focus groups and computer-assisted-personal-interviews (CAPI) are two favoured and fashionable techniques in the current research tool-box. Many authors we consulted salute the fact that focus groups are a natural extension of everyday talk in small group work in the primary classroom. Adults can be considerably shyer with their opinions in round-table discussion. In the same way, the task of answering a question by moving-mouse-and-clicking-box may seem monstros-strange to older cohorts, but second-nature to the IT generation.

Mauthner’s (1997) paper reproduces a series of passages from her group interviews with five and six-year olds which demonstrate some advantages of the focus group for such subjects. The topic was conceptions of health and the general wisdom is that young children can be somewhat guarded on such matters in individual interviews. Mauthner (MM) notes that group interviews, however, tend to spring into spontaneity. For example:

Suraya: ‘Do you want to see my head? There it’s like this’ (shows scars from a fire).

Loretta and Nina: ‘Let’s see.’

MM: ‘Oh Dear.’

Loretta and Nina’s curiosity is captured and they are prompted to work through their own experience of accidents, hospitals, needles etc. This is not to say that all of the standard problems of the focus group strategy evaporate in these circumstances, of course. Balance is frequently lost because of the dominance of a particular group member and Suraya demonstrates that children are just as capable of being the culprits in this instance:

Loretta: ‘How can you do a tape when she’s talking of her.’

MM: ‘That’s right if you talk all the time we won’t be able to hear Nina and Loretta.’

Suraya: ‘But I was talking first.’

Nina: ‘She likes talking.’

Suraya: ‘Yeah.’

The progress of computer-based programmes and their evaluation may be gauged with a brief look at Russ et al. (1998) research of EatLife, a WWW-based nutrition programme for young girls. This Virginian study used two tenth-grade classes as ‘experimental’ and ‘comparison’ groups. The former completed a pre-treatment, ‘food-frequency’ questionnaire on-line, worked though the EatLife modules, and then returned to the web for post-test. The comparison group completed a paper-and-pencil version of the food-frequency test before and after the same materials were presented in classroom-based, health-education lessons. Results showed significant improvement of the electronically-tutored over the desk-bound in terms of meeting all dietary targets (except, for unexplained reasons, the reduced ‘soda consumption’ goal).

Before we jump to conclusions on the world-wide wizardry involved here, the usual evaluation suspects need to be interrogated in terms of lack of random assignment and lack of inspection of long-term change. In particular, we need to know what is it about the programme that has worked before we can talk about transferable lessons. The EatLife authors make much play of the instant, personalised feedback available through the computer programme. Detailed graphics were presented to each logged-on name, providing individualised progress maps and strategies to meet or continue with each goal (Hi, Clarissa. Try to do better on ......). It may well be that it is this ‘target-raising potential’ of the programme rather than the ‘curriculum presentation’ which is the key electronic resource, and this research is yet to sort them out (just as it is unable to say what it was about the human delivery which was so persuasive on Pepsi). Provided the technology is in place, such initiatives are extremely portable and will no doubt multiply. The journal Computers in Human Services may be consulted for enthusiastic progress reports, with a paper by Resnick and Sherer (1994) providing an overview.
In Sections 3.15 through 3.17, we take a closer look at some of the methods that involve children in measuring actual programme outcomes via the use of ‘approval ratings’, ‘self-reported behaviour’, ‘educational gains’ and so on. We will report serious doubts about some of these approaches. The presence of children as research subjects makes none of these techniques easier but the real problem lies with the unreliability of such indicators in the first place.

**Lesson 11.** Having children as programme subjects and research subjects provides evaluation with a special concern with ‘age’. The aptness of intervention never depends simply on age *per se*, but on the child’s developmental stage and their exposure to the contributory problem and their control over prevention strategies. The key strategy for promoting children’s participation in the research process lies in the adaptation of standard instruments, presenting questions in frameworks and formats with which particular age-groups are already familiar.

### 3.12 Localities

‘A’m terrified aw they verandas an’ they stairs’
(Glaswegian, high-rise mum, quoted in Roberts et al., 1995)

We have already made it clear that the effectiveness of programmes is contingent not only on differences in the subjects to whom they are applied but also on differences in the situations in which they are applied. Children face different traffic conditions, health environments, safety surroundings, crime settings and in this final element in the *for-whom-and-in-what-circumstances* section we examine how ‘locality’, ‘community’, ‘neighbourhood’ and so on should be reckoned with in evaluation research.

We take as our illustration for this lesson just one key study - the work of Roberts et al. (1995) on children’s safety issues in Corkerhill, a deprived, high-density, high-rise housing estate to the south of Glasgow. This is not a programme evaluation as such but an effort to examine the local causes and consequences of children’s accidents. Its implications for policy, programming and evaluation are nevertheless considerable. Unlike the other studies we examine here, the Corkerhill research takes as its starting point spatial and social class inequalities of risk and accidents:

‘The social class gradient for childhood accidents deserves special mention because it is steeper than the class gradient for all other causes of death. Children aged 0-14 with fathers born into class V are twice as likely to die from any cause as their counterparts born to fathers in class I. However, such children are three times more likely to die from an accident and five times more likely to die by being hit by a car; and the gap between classes for deaths caused by fires, falls and drowning is even higher’.

The author’s explanation for these patterned outcomes lies in the unequal distribution of risk faced by working-class children. They tend to live in particularly hazardous environments of which Corkerhill is a typical case. Such an explanation predisposes the researchers to a particular policy-line, ‘Reducing inequalities in the distribution of accident risks is likely to be the most effective strategy for reducing the overall accident rate’ and the case study seeks to flesh out such a proposal for this particular estate.

The inquiry employed three methods:  
i Group interviews with parents and teenagers and local safety officials on the nature, causes and consequences of accidents - what accidents are common? which ones concern people most? where do local accidents tend to happen? what promotes safety and what puts children at risk? which areas were safe and unsafe? how to cope with anxiety? how to reduce danger and make things safer? 

ii A local census of child accidents. This high-response-rate survey provided population-based data on the accident prevalence in the community by asking about actual experiences of accidents - seriousness, location, age of children, immediate cause etc. - over a given reference period. This operated rather like *The Crime Survey* as an ‘alternative’ to official sources such as hospital injury data or figures on health service use. 

iii In-depth studies of accidents and near accidents that children have experienced and parents have dealt with in the past year. Twenty five actual accidents and near-misses were scrutinised to examine the scope parents had in anticipating or intervening in the incidents.

These inquiries produced an incomparable local data base on accidents. The survey was able to map different parts of the estate according to which accidents occurred for which age group. Not only was the frequency of accidents in Corkerhill twice the Glasgow average, but there was an unequal distribution of risk within the estate. Table 4 gives figures for the distribution of accidents at home and, as with many crime surveys, a surprising number of ‘repeat victims’ is discovered.

### Table 4 Non-random distribution of household accidents in Corkerhill

<table>
<thead>
<tr>
<th>Number of accidents per child</th>
<th>Number of children</th>
<th>Percentage</th>
<th>Expected percentage (Poisson distribution)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>258</td>
<td>70</td>
<td>43</td>
</tr>
<tr>
<td>1</td>
<td>55</td>
<td>15</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4+</td>
<td>32</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

**Source:** Roberts et al., 1995

Group interviews revealed a keen understanding of risk on the estate. The ‘parents groups’ identified:
‘factors associated with the design of the estate, such as road layout, on-street parking, the location of waste disposal and so on. They also include factors associated with the design of different housing types on the estate, such as unguarded flat-topped roofs, ungated external stone stairways, verandas with low railings and dangerous gaps. Parents expressed concerns about the internal design of their homes, such as the location of the kitchen vis-à-vis the front door and the living room: and they talked of the hazardous fixtures and fittings - live sockets, easily opened windows and so on’.

Some risks were quite peculiar to the estate with, for instance, a nagging problem of to-leave-or-not-to-leave-the-kids being presented by the ‘greens’:

‘in flatted and tenemented properties, hanging out the washing is rarely a straightforward task, since it generally requires the washing to be carried down concrete staircases to the communal green where the clothes lines are located. One fifth of the parents we spoke to never used the greens (often encountering the displeasure of the local housing factor by hanging out their clothes on their verandas). Of those needing to juggle the demands of washing and children, four out of five had to find a way to transport the washing and the children at the same time, often because they could not find a safe place to leave them. The majority achieved this holding hands or strapping them to a buggy’.

The discussion of actual accidents revealed that in nearly all cases the parents were nearby, had managed to contain the physical and psychological injury, had reasonable ideas about what caused the accidents, and in many cases took action to prevent it re-occurring. Yet certain classes of accident did recur and the authors put this down to environmental factors over which the residents had relatively little control - such as the 86% of homes whose hot water system did not have a thermostat, the metal fencing which had replaced hedges, the off-estate, across-main-road, school location, and so on.

So much for the study, but what are the implications for future safety programmes and their evaluation? Although we regard the empirical work as exemplary, we learned rather different lessons from it than did the authors. Roberts et al. (1995) argue fiercely for an ‘environment over behaviour solution’ as follows:

‘In our group interviews, we asked respondents to suggest cheap ways of preventing accidents. Those groups made up of residents in the community were far more imaginative and better informed, particularly in the detail of how this would be done, than were the group of professionals .... The professionals assumed that the community needed to know about the dangers: the community knew it has enough information on what the dangers are, and made appropriate proposals to deal with them’.

Before we leap to an environmental management solution and into the politics of how significant changes can be made to deprived areas, we need to take notice of other points, said and unsaid, by the respondents. In the survey, 90 percent of parents favoured the idea of more time being spent on safety education for children. 70 per cent of residents believe that parental attitudes are an important factor in predisposing children to accidents, with the same percentage seeing the children’s behaviour itself as an important risk factor. Residents, it seems, do not share the researcher’s confidence about their collective understanding of safety. Despite their active involvement in the project, Corkerhill residents also believed that having ‘professionals’ carrying out the research would provide it with more clout, and again we are left wondering whether they would be quite as disparaging as Roberts et al. (1995) about the potential contribution of the health visitors, health promotion officers, police officer, fire prevention officer, and road safety officer who made up the professional safety panel. Finally, we note that despite the apparent utilisation of three teenage focus groups, little is reported about children’s perceptions of risk and so the research is silent on the matter of whether the kids possess and utilise the storehouse of safety nous demonstrated by the parents.

So what should we take away from Corkerhill? First, a crucial point with massive implications for all safety research with children:

‘risks are specific to local environments, .... prevention strategies are learned in these local contexts and .... policies for safe keeping must therefore be based on local knowledge’.

Second, we need to consider how policy and programmes can incorporate and harness that knowledge. Roberts et al. (1995) clearly favour an action based agenda in which a problem identified becomes a campaign issue. Their rigorous local aetiology of accidents presents an incomparable base for a programme of environmental improvement. The Corkerhill research is thus a model of how to do formative evaluation (recall the earlier examples about how not to do it). The residents have a keen sense of priorities (verandas, thermostat control, play areas etc.) and some realistic and prudent ideas about putting solutions into practice. Third, however, we would argue that such local knowledge should also be the bedrock of those education programmes so despised by Roberts’s team. The Corkerhill research does not persuade us that the
fined-grained safety perceptions of some residents are shared by all adults. And indeed it is clear from their own testimony that the people of Corkerhill do not believe that their own children are blessed with this wisdom. The key point, therefore is that the formative research should also provide the curriculum for safety education. Rather than teaching only abstract and generalised safety lessons (using slogans, cartoons, handbooks etc.), local knowledge (about children being badly injured by slipping through gaps in the verandas, about the lorries which illegally use the car park, about the hill which obscures drivers’ view of the traffic lights, about the gang battle which is forcing some children to take a more dangerous route to school, about the vandalism that besets the pelican-crossing, and so on) should be gathered and promulgated as part of a community education approach.

Roberts et al. (1995) research reveals the vital place of ‘community’ in safety research. We prefer to think, however, that accident prevention is not a matter of individuals-versus-communities but individuals-in-communities. Knowledge of the geography and sociology of accident prevalence is vital for the construction of programmes of all types. It also is vital for testing initiatives. The Corkerhill study, of course, did not involve a programme as such but, had there been one, success would be measured in terms of growth in understanding of, and response to, the local hazards.

Lesson 12. Safety problems congregate in hazardous environments. Many of the risks children face are specific to local environments. Programmes aimed at children’s safety thus need to tap into local knowledge of the conditions which generate the problems they are set to tackle. Evaluation research needs to map differences in peoples’ everyday wisdom about their communities and measure success in terms of growth in understanding of, and response to, the local hazards.

3.13 What outcomes?

“How does one measure what has been prevented from occurring?” (Warden et al., 1997)

Our review moves on to the vital issue of how to measure the outcomes of programmes. In this and the following seven sections we attempt to learn lessons on how best to decipher the impact of initiatives. Complexity comes to haunt the evaluator in yet another guise, and we begin by noting that programmes always generate complex outcomes and that good evaluation needs to begin with a sense of what these might be. In this, our first and most general, lesson on the measurement of outcomes, we produce a simple typology covering some of the different parameters of ‘complexity’, leaving the subsequent lessons to generate some solutions and recommendations:

- Range. Human interaction is restless and resourceful and so participants take many different things away from a programme encounter. The stock-in-trade intervention in our review was educational and ‘education’ has many roles. Learning can build knowledge, confidence, practical skills, social skills, cognitive development, morality and so on, and it is necessary to observe for the match or mis-match between these diverse pay-offs and programme goals.

- Policy Ambiguity. Programme goals may have been set in terms of bringing about ‘improvements’ (in which case the evaluator has an impossibly open agenda for measurement) or of ‘shifting a performance indicator’ (in which case the researcher may have too narrow a brief to evaluate the overall costs and benefits of a programme).

- Long and Short Term Effects. Programme impacts can dissipate over time and the problem of relapses and repeats is commonplace in many of the child and youth related social problems we encountered. Evaluators rarely have the resources to study outcomes longitudinally.

- On Paper and in Practice. Many of the educational programmes under review involved teaching the children sets of rules for good practice. It is often difficult for the evaluator to see whether classroom knowledge is put into actual practice in the home and the street. More generally, the evaluator faces a task with any programme of measuring whether new knowledge translates into attitudinal transformation and into behavioural change.

- Measuring What Has Not Happened. Many programmes we examined have the task of making subjects refrain from some dangerous or undesirable behaviour. This raises the paradoxical problem raised by Warden in the epigraph above about how one measures what has not happened. The behaviour targeted by an intervention may not occur because it has been ‘prevented’ or ‘cured’, but also because subjects’ remain in ‘blissful ignorance’ or ‘lack opportunity’ or ‘move on’. There is an obvious outcome difference in terms of interventions attempting the breaking of a habit rather than preventing its uptake, but abstinence in general is hard to interpret.

- Unanticipated consequences. The are many proverbial about education which warn of its perversity outcomes: ‘a little learning is a dangerous thing’, ‘not being able to see the wood for the trees’, ‘you can take a horse to water but you can’t make it drink’. These remind us that no programme will be inoculated from unexpected outcomes and set the evaluator the perverse task of ‘anticipating the unanticipated’ when monitoring outcomes.

In the light of these difficulties, we begin with the ideal scenario, which is that the variety of processes present in a programme require a series of outcome measures to track their workings. Any single measure of outcomes runs the risk of producing a highly selective reading of how the programme has operated. Off-the-shelf measures should not be trotted out unthinkingly, nor is there any simple technical formula which the researcher can apply to assure that measures are valid and reliable. The test of a good measurement system is its ability to reproduce the whole footprint of programme outcomes.
Lesson 13. Programme effects are always diverse and evaluators should never rely on singular measures to monitor programme success (or failure). Evaluators should always use multiple measures in order to monitor the patterned outcomes associated with the diverse processes involved in constructing, implementing and experiencing initiatives.

3.14 Hierarchy of outcomes
Programmes are generally considered to reach their objectives by stages. The most typical is the sequence identified above in which an intervention delivers a knowledge gain which is expected to translate into an attitudinal shift and then onwards into behavioural change. Our initial imagery of programmes embedded across levels of the idea, individual, institution and infrastructure also suggests a ladder up which a programme must climb if it is to succeed. In certain policy fields these potential causal flows are well recognised and have been formalised and targeted as ‘hierarchies’ of outcomes or objectives.

Dannenberg and Fowler (1998) offer a well-worked example, used in the US in the field of injury prevention and control, which is presented here as Table 5.

Injury-prevention programmes have a self-evident goal, but injuries may vary from the fatal to the minor and, according to the problem targeted, injury data may have to be sought in death certificates, emergency department records and so on. Sometimes these official records are of limited use, such as when the injury is untreated and thus lost to official data, or when the programme locality does not coincide with data collection district, or when the type of injury is sufficiently rare that it takes too long for clear outcome differences to show up in official records. In these cases a ‘surrogate measure’ is used which attempts to record intermediate changes in behaviour, policy or environment which go towards the desired improvement in safety. In these instances direct observation or survey questions are the preferred methods used to estimate outcomes. Which level of the hierarchy is appropriate and which measurement technique is eventually used depends on the character of the intervention and programme goal, and the third column gives examples of programme evaluations that have used each type of measure.

A second example of hierarchy, which reveals more about the issues at stake, comes from a Community Fire Safety Task Force report on fire-safety schemes (1997). Table 6 provides a detailed listing of the targets which must fall if ‘outreach schemes’ are to result in the ‘bottom-line’ success of a programme.

Not only does the measurement of the sequence of events from outreach to end impact require different techniques and forms of data, it is also associated in the report with a ‘hierarchy of proof’. This is a claim about the relative validity and reliability of different outcome measures. The audience must understand the material and remember it. It must add to their knowledge or remind them what they know. The target audience must act on the information. Actions to improve safety of the home environment need to be done correctly and the changes maintained. The behavioural or environmental changes must have a significant impact on the types of problem that actually occur and not be overwhelmed by factors beyond control or not addressable by Community Fire Safety.

### Table 5 Hierarchy of outcomes used in the evaluation of injury interventions

<table>
<thead>
<tr>
<th>Inquiry outcome</th>
<th>Data sources</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A Actual injuries</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Fatal injury</td>
<td>Death certificates</td>
<td>Refrigerator entrapment deaths</td>
</tr>
<tr>
<td>2. Hospitalised injury</td>
<td>Hospital discharge data</td>
<td>Child clothing related burns</td>
</tr>
<tr>
<td>3. Emergency department treated injury</td>
<td>Emergency department records</td>
<td>Glass related lacerations</td>
</tr>
<tr>
<td>4. Any medically treated injury</td>
<td>Medical clinics and emergency departments</td>
<td>Softball Injuries</td>
</tr>
<tr>
<td><strong>B Surrogate indicators of injuries</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Observed behaviour</td>
<td>Observers at selected times and places</td>
<td>Child car safety seats</td>
</tr>
<tr>
<td>2. Environmental changes</td>
<td>Survey of hazards or safety devices</td>
<td>Home smoke detectors</td>
</tr>
<tr>
<td>3. Policy changes</td>
<td>Monitoring regulatory activities</td>
<td>Hot tap temperatures</td>
</tr>
<tr>
<td>4. Self reported behaviour</td>
<td>Survey sample of individuals</td>
<td>Bicycle helmet use</td>
</tr>
<tr>
<td>5. Knowledge, attitude, and beliefs</td>
<td>Survey sample of individuals</td>
<td>Knowledge of safety practices</td>
</tr>
</tbody>
</table>

Source: Adapted from Dannenberg and Fowler (1998)

### Table 6 Hierarchy of outcomes used in the evaluation of fire safety

<table>
<thead>
<tr>
<th>Outreach</th>
<th>Knowledge gain</th>
<th>Behaviour change</th>
<th>Environment change</th>
<th>End impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Getting the safety information to the target audience and reaching enough of the audience to make a difference.</td>
<td>The audience must understand the material and remember it. It must add to their knowledge or remind them what they know.</td>
<td>The target audience must act on the information.</td>
<td>Actions to improve safety of the home environment need to be done correctly and the changes maintained.</td>
<td>The behavioural or environmental changes must have a significant impact on the types of problem that actually occur and not be overwhelmed by factors beyond control or not addressable by Community Fire Safety.</td>
</tr>
</tbody>
</table>

Source: Community Fire Safety Task Force (1997)
measures. The more community fire safety initiatives are considered a success, the higher they should climb the ladder representing ‘strength of proof’ (see Table 7).

Our view of such hierarchies is that they should be writ large in organising evaluation rather than carved in stone. There are often insurmountable obstacles in reaching the peak, as in aforementioned points about the infrequency of certain outcome events and the lack of correspondence of programme locality and official data collection areas. Also important is the practical consideration that measurement of ultimate outcomes is generally much more expensive and time-consuming. Furthermore, the greater the length of the chain to final outcome, the greater the possibility that ‘confounding’ mechanisms, other than the process assumed in the hierarchy, may have been influential in accounting for the behavioural changes.

Such difficulties notwithstanding, we are on the whole persuaded by ‘bottom-line’ reasoning which argues that road safety programmes are there to prevent road accidents, that crime-reduction initiatives should prevent crime, that smoking cessation is about stopping children smoking and so what counts are actual numbers of injuries prevented, crimes reduced and smokers quitting. We are thus not convinced by the ‘goal-free’ approach (Scriven, 1991) of some constructivists that evaluators should concentrate on some of the initial steps in the hierarchy, by way of seeking agreement and approval amongst stakeholders about the way in which the programme should be implemented. In truth, this is not an ‘either-or’ argument. The key point about the hierarchy is that it represents a sequence of steps, each of which is required for bottom-line outcomes to come to fruition. So, for instance, knowledge of disparity of ‘knowledge gains’ or ‘approval ratings’ are precisely the clues we need to understand why the programme will sometimes fail and sometimes succeed. The key feature of the notion ‘hierarchy of outcomes’ thus lies in the ‘connectivity’ or ‘lack of connectivity’ of the various stages.

We conclude with another broad target of measurement in evaluation - the gradual building, over time, of an account of the stronger and weaker links in the chain of outcomes associated with each class of programme. Just occasionally, it will be possible to simply ‘assume’ that one step leads to another as in the following example from Dannenberg and Fowler (1998) - ‘The use of a surrogate measure presupposes a clear link between it and actual injuries. For example, it was assumed that increased bicycle helmet use after the passage of a mandatory helmet law would be associated with reduced injuries because prior work demonstrated the protective effect of helmets’. But, in relation to most programmes, connectivity is not established and we continue in the next sections to explore the linkages.

Lesson 14. A hierarchy of outcomes should be established for each family of programmes following through the changes in understanding, attitudes and behaviour, which the programme is seeking to affect. Outcome measurement is more secure the further it travels along the understanding → attitudes → behaviour chain. It is often infeasible to measure long-term behavioural outcomes of a programme and actual outcome measures should be chosen to optimise the balance between the ‘feasibility’ and ‘strength of proof’ of potential measures.

The majority of the studies we reviewed did not have the luxury of following initiatives right through to bottom-line effects, with practical and/or funding constraints requiring them to stop the search for outcomes at some intermediate location. We move on in the following sequence of lessons to post warnings against three of the most fragile intervening destinations making suggestions, as we go along, by way of potential solutions. The three weakest links in the outcome chain are ‘approval ratings’, ‘self-reported behaviour’ and ‘educational gains’.

3.15 Approval ratings
If there is a person-on-the street understanding of ‘evaluation’ it is probably reckoned to be something to do with ‘approval scales’. A product is consumed and the consumers are asked to rate its taste, looks, value-for-money and so on. A course is taken and students are asked to rate its presentation, content, challenge and so on. An intervention is mounted and participants are asked to rate its appeal, value, and capacity to change them. The standard methodology consists of the production of a set of attitude statements about a programme, on which the subjects rate their strength of agreement - disagreement. This is a relatively straightforward operation, indeed it is the method most often carried out by practitioners and other lay researchers, in the form of small surveys of participants at the completion of a programme.

Table 7 Strength of proof of a variety of fire-prevention outcome measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Proof of Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of fires and deaths and injuries resulting from fire</td>
<td>Strongest proof</td>
</tr>
<tr>
<td>Percentage of households with smoke detector</td>
<td></td>
</tr>
<tr>
<td>Percentage of children who know about ‘Get Out and Stay Out’ and ‘Stop, Drop and Roll’ messages</td>
<td></td>
</tr>
<tr>
<td>Percentage of children or of schools receiving safety education programme</td>
<td></td>
</tr>
<tr>
<td>Percentage of children or of schools receiving safety education materials</td>
<td></td>
</tr>
<tr>
<td>Percentage of teachers who think program materials are good and use them</td>
<td>Weakest proof</td>
</tr>
</tbody>
</table>

Source: Adapted from Community Fire Safety Task Force (1997:20)
A simple example of the kind of question we have in mind is item 1 in a questionnaire to parents used in the evaluation of a trial of the ‘Teenscape’ sex abuse prevention programme (Elliott, 1995).

1. Do you feel that your child’s response to being taught ‘Good Sense Defence’ was:

   positive [ ] negative [ ] made no difference [ ]

Such an approach has certain well-rehearsed problems which flow from it being a kind of ‘popularity exercise’. There are all kinds of endearing things about human nature which influence subjects to be largely positive about experiences such as attending a programme. This desire exists in slightly different forms, under slightly different names, in the methodological literature - 'subject effects', 'halo effects', 'Hawthorne effects' etc. We enter into no detail here other than to offer the base scenario that if subjects have invested some time on a programme, and become acquainted with the practitioners on whom they are reporting, and have some hopes to display themselves in a positive light, and have bothered to spend time on a questionnaire - then there is little spur on them to be anything other than positive. Readers are invited to guess the recorded percentages in the three tick boxes above.

There are, of course, methodological strategies to combat the tendency to the ‘managed response’. Attitude, opinion and rating questions can be set in sufficiently large number so that they contain a set of inter-linked themes, or the same question can be asked with different wording, or the questions can be posed at different points of time - and in all cases the research is able to see if there is ‘internal consistency’ to the pattern of response, with such consistency being regarded as a check on the reliability of the subject’s responses. Again we enter into no particular detail on such strategies because they constitute, in our opinion, sledgehammers to crack a none too significant nut. The information (on approval for a programme) that the researcher digs out with these supplementary reliability checks is so far removed in outcome hierarchy from actual behavioural change that undue precision at this stage is hardly warranted.

In short, the measurement of programme approval ratings is incredibly difficult to do well, a cinch to do badly and, furthermore, not at all difficult to ‘fit up’. There is a mass of material in the questionnaire design literature on approval and satisfaction ratings which show that by changing subtly the item which the subject actually rates, the researcher can shift the overall levels of approval significantly. The standard text book examples (Foddy, 1993) usually come from opinion polling on such matters as Prime-Ministerial popularity, which record quite different levels of support according to whether the question asked is about ‘doing a good job’, being ‘approved of’, ‘who would make the best leader?’ and so on.

Such ‘accidents’ or ‘tricks of the trade’ of phraseology can also crop up in do-it-yourself evaluation. Finckenhauer (1982) reports some rather ‘loose practice’ associated with the prison and lifer’s group’s own evaluation of their _Scared Straight _programme which were the basis of the

3.16 Self-report

Short of measuring actual behaviour, the next option available in the researcher’s tool-kit is to attempt to chart programme outcomes by getting the subjects to ‘self-report’ on changes in their behaviour. Once again, however, we are faced with a method well-known for placing pressure on the subject to ‘conform’. Programmes with children tend to have a high normative content, forms of ‘correct’ behaviour may well be stressed over and over again during the intervention, and the subject’s life is made somewhat easier if they recount that they do indeed avoid parked cars when crossing, have quit smoking, no longer do drugs, do apply sun-cream, do eat five portions of fruit-and-veg per day and so on. The open question, of course, is to what extent do such verbal reports correspond to actual conduct or to what extent are they contaminated by ‘faking good’, the ‘acquiescence effect’, the ‘social desirability’ effect, ‘hypothesis seeking-behaviour’, and so on.

Our review provided mixed messages on the validity of self-report. Buller _et al. _(1996) evaluation of a skin cancer prevention curriculum for elementary school aged children questioned them about the frequency of sun-bathing, usage of sun-creams and so on, and concluded, ‘These data raise some concern about the validity of children’s self reports. Sun protection at this age may depend less on decisions
made by children and more on those made by parents’. By contrast, Perry et al. (1998) study of a Minnesota version of the 5-a-day fruit and vegetable programme, in which children were asked to recall food consumption in the past 24 hours, asserts that ‘these methods have been shown to provide valid estimates of group intake’.

To get to the bottom of these contrasting views one needs to examine how the judgements on validity are made. The ideal is to have an ‘objective’ measure against which the ‘subjective’ self-report can be tested. In certain programming fields these exist. So for instance, the Buller study mentioned above measured tanning using the ‘Minolta CR-200 Chroma Meter, a valid and reliable means of assessing skin color changes that is more precise and accurate than relying on visual skin assessment’. (1996) and went on to report that, ‘surprisingly, few self-reported prevention behaviours were correlated with changes in suntanning’. In the field of smoking cessation, validity of self-report is checked out by comparing what children have said with measurements of carbon monoxide from expired breath. This apparatus has been used in many studies and a meta-analysis by Patrick et al. (1994) concludes that self-reports of smoking are generally valid. There seems to be a variation within the general picture, however, with reports about age and ethnic differences in the validity of smoking self-report. Wills and Cleary (1997) and Bauman and Ennett (1998) have both discovered differences in the validity of self-reporting between black and white children. They disagree sharply, however, on the extent of mis-reporting and this is possibly due to age differences in their samples. Probably the safest conclusion about the validity of self report lies in the latter pair of authors’ (1998) warning that ‘sweeping generalisations based on small numbers of cases can be misleading’.

But what of the validity of self report when the research does not have the back-up of an objective piece of physical apparatus to apply to the human body? Reported behaviour that has physiological effects with long-standing chemical traces is much easier to check out than behaviour that has a fleeting social presence in out of the way places. Nutritionists have probably made the greatest strides in this respect with several ‘validation studies’ being made of a standard method of self-report, which involves children logging food consumed in a 24-hour diary. The validity of these diaries is checked through the use of observers during school meals and parents’ reporting on actual consumption at home. Lytle et al. (1993) inquiry concluded that ‘the 24-hour recall method assisted by food records is a valid method for assessing dietary intake of children as young as 8 years old for the purpose of group comparison’. But they do note that ‘children have a difficult time quantifying the portion sizes of almost all foods ... for some food groups the overestimation by recall is more than 100% of that observed’. One limitation of this strategy is that the children are fully aware that they are taking part in a validation study (although they might not quite put it like this). The presence of the school observers was clear to see, trays had to be put aside in a special place for ‘left-overs inspection’, parents had to follow a rather special strategy is that the children are fully aware that they are taking part in a validation study (although they might not quite put it like this). The presence of the school observers was clear to see, trays had to be put aside in a special place for ‘left-overs inspection’, parents had to follow a rather special...
much akin to a proposed study. More confidence may be had in self-report if there is an objective, physical baseline against which to validate the subject’s statements. We note that this approach is far from fool-proof. Luepker et al. (1996) study of the CATCH programme (1996), which was aimed at general improvements in physical health, used a vast array of objective measures (body mass index, triceps skin-fold, supacarpal skin-fold, heart rate, systolic blood pressure, diastolic blood pressure, total cholesterol, HDL cholesterol, apoliprotein) and it is somewhat difficult to imagine how one would track self-report back to such a pile of symptoms. Self-report of purely social activity, nevertheless, remains harder to validate and road traffic behaviour falls squarely in that category.

Finally, we pass a brief warning against a ‘shortcut’ approach which leaves us somewhat distrustful. The strategy is to add extra conditions to self-report in order to boost its accuracy. The most famous version of this is the ‘pipeline methodology’ which was introduced into the evaluation of anti-smoking initiatives by Evans et al. (1977) in 1977. The idea is to convince subjects that the research team have an independent measure of their behaviour which will be used to check upon self-report. As we have seen, tests and specimens can indeed be taken, in certain instances, to validate self-report. The idea of the pipeline is that the anticipation of the check-up in itself is a spur to accuracy. The strategy then becomes a so-called ‘bogus pipeline’ in situations where no real follow-up is planned, or when the technology to do so does not even exist. Apart from the potential deception involved here, the pipeline is in fact no automatic helpline because the basic methodological problem remains. Just as subjects react to a programme in different ways, they respond to the challenge of the potential check-up in different ways. To investigate these conditions, Murray and Perry (1987) conducted a series of trials in which subjects were randomly assigned and asked to self-report under a number of research conditions (variation in the ‘credibility’ of the pipeline, the ‘illegality’ of the substance to be detected, assurances about subject anonymity and confidentiality). They report, ‘.... pipeline procedures significantly increase disclosure of tobacco and marijuana use when students are promised confidentiality but not anonymity. However, when anonymity was assured, disclosure of cigarettes was just as high without the pipeline; for marijuana use, disclosure was higher without the pipeline. No effects were observed for alcohol disclosure.’ In short, as with all measurement techniques we review, effectiveness depends on context.

Lesson 16. Programmes cannot be considered properly tested if outcomes are measured solely through the usage of self-report. The accuracy of self-report varies from topic to topic and from subject to subject in any programme area. Each field of programming needs to build up its own intelligence on the validity of self-report by considering the power of recall, level of stigma, degree of anonymity and capacity for deception associated with the task.

3.17 Educational gains
The old jibe about people who are ‘good on paper but unknown quantities as far as practice goes’ could have been invented for evaluation research. The problem occurs and recurs in a huge percentage of the studies we have reviewed. Educational, or knowledge, gains were observed routinely but many studies were unable go on to claim long-term changes in behaviour (for the kinds of reasons already rehearsed). A brief sample of ‘confessions’ follows:

- Gibson et al. (1998) report, following a school asthma program, significant gains in knowledge about asthma (from a very low baseline). However, asthma related ‘quality of life’ was not altered by the intervention. The authors explain, ‘To effect a change in quality of life may require a longer period of follow-up after the intervention, a more intense intervention, or the addition of medical intervention to optimise therapy and a large sample size’.

- Sundelin et al. (1996) report of a local intervention designed to galvanise the impact of a televised safety initiative that ‘The local campaign increased parents’ inclination to follow the programmes. No significant association was found, however, between the number of programmes followed and measures undertaken in the homes as a direct consequence of the programmes’.

- Liller et al. (1998) report that children in intervention schools attending the More Health poison-prevention lesson answered significantly more questions correctly than those in control schools. However, ‘It was not possible to do follow up interviews to determine whether knowledge was retained. More importantly, we were unable to study children’s behaviour to see if their knowledge translated into actions’.

- Gielen et al. (1996) report significant increases in knowledge for both parents and children and some reported behavioural changes at home after a visit to a million dollar ‘children’s safety village’, but acknowledge, ‘We were unable to observe the use of specific skills or confirm self reported practices’.

- Tudor-Smith et al. (1995) report of a mobile health education resource in Wales for children. ‘The results suggest that the programme was highly rated by the pupils, and increased their knowledge about substance use and its promotion in the media. However, there was no improvement in the pupils’ ability to choose a healthy course of action when under peer pressure to do otherwise, and no change in intentions to smoke when aged 16. Further research is advocated’.

- Fielder et al. (1996) report on a skin cancer prevention campaign that, ‘Analysis of questionnaires administered to visitors ... six weeks after the campaign showed increased awareness about sun protection measures after the campaign, but no change in behaviour’.

And so it was in every drawer of the filing cabinet. And so it is, no doubt, in road safety education, where the issue would be whether the child’s proven ability to plot safest routes across a diagram of a complex junction translated into practical steps across actual tarmac.
The issue in this instance is not about the reliability and validity of the data on knowledge gains. Educational testing is part of the fabric of social life. Techniques for measuring factual recall are tried and tested. The standard methodological qualms on this topic about ‘memory fade’ and ‘learning effects of testing’ have equally standard remedies. One can investigate deterioration in longer term recall of information by testing over time (should the inquiry have the resources). The fact that the educational test itself may give a knowledge boost can be detected in gains made by an untreated control group and avoided using ‘four-group designs’ (Cook and Campbell, 1979). We feel safe to record, therefore, that the vast majority of knowledge gains reported in our review were indeed as recorded.

Educational gains are, therefore, a potentially weak link in the evaluation chain - not so much by dint of unreliability – but, because they still represent a long jump to the next point in the hierarchy, namely behavioural change. Accordingly, we make the radical suggestion that there is little need for evaluation to record, over and over again, that children grasp well-presented information. The pressure to do so, unfortunately, is relentless, for what often grips the policy-maker and practitioner imaginations are new ways of delivering programmes (children’s villages, clubs, CD-roms, mobile safety buses etc.) and this is what tends to get funded and evaluated. Nevertheless, for us, the lesson of so many studies is the imperative to get beyond the soft objective of ‘educational gains measures’, developing them in ways that come nearer to providing pointers to future behaviour.

There are two significant ways forward. First, even if outcome data is strictly limited to a ‘change in understanding’ perspective, it is worth assuming that ‘aggregate’ positive change will invariably take place and concentrating evaluation effort on the differences within the overall picture. Thus road safety campaigns will no doubt throw up different rates of understanding across subjects in the ‘code’ and there will nearly always be differences in ability of different sub-groups of subjects to retain that knowledge. Such discrepancies will provide pointers to the potential mixed fortunes of an initiative that may provide for closer-grained understanding of behavioural outcomes, when and if such data becomes available. The Liller study from the list above did in fact come up with disparities in outcome knowledge about different types of poisons (gas, liquid, solid, snake, spider) and also marked variation in understanding across school grades. Such peaks and troughs in children’s comprehension may be ultimately echoed in hospital records of poisoning and lead to an improved targeting of the programme.

The second way to improve educational-gain measures is to move them nearer to the point of practice. Knowledge gains are often better demonstrated ‘in the field’ rather than ‘on paper’. In a road safety situation this might involve getting children to demonstrate how they would cross at a particular junction (as in a ‘pedestrian proficiency test’) rather than how they would cope on a diagram. The ‘pretend road method’ (in which children view traffic approaching on the real road but cross on an adjacent pretend one) is an example of what we have in mind (Young and Lee, 1987). Such ‘in situ’ testing still carries some artificiality in that the presence of a researcher may still encourage ‘faking-good’, but we are half a step nearer to real decisions. This principle of getting-closer-to-reality has, as we have seen in earlier sections, also been taken up by programme-builders who attempt, through a variety of live and simulated interventions, to increase the verisimilitude of the learning situation. Here, we want to stress the importance of the same idea in measuring programme outcomes. The ultimate test of a knowledge gain, of course, is to see if it is put into practice without awareness of being tested. And this is the conundrum which we go onto examine next.

Lesson 17. Any education programme worth its salt will produce educational gains and less evaluation time and effort is required to show that this is the rule. The more ‘proximate’ to reality the knowledge-gain measure the better. Evaluations without the resources to carry through investigation to behavioural outcomes should concentrate on assessing the variability in knowledge gains across different subjects and different aspects of the programme curriculum.

3.18 Simulations
In this section we reach the limit of ‘proximate’ or ‘surrogate’ measures of programme outcomes. The preceding lessons have all pointed to the hazards in assuming that a programme works on the basis of measures taken on paper rather than observed in practice. Initiatives on children’s safety are particularly vexed with the question of whether lessons learned in the secure environment of a programme are put into practice when the child is faced with real danger. One extreme solution, on the completion of a programme, is to put its efficacy to test by simulating the real and threatening conditions that could confront the child.

Seldom has such an approach been tried, due to the awesome practical and ethical difficulties involved. One noted, and perhaps notorious, attempt was the evaluation of a Need to Know personal safety course by Fryer et al. (1987), which involved the ‘staging of an actual situation in which each of the children had an opportunity to leave the school building with a stranger’. The research involved the classic OXO design - pre-test simulation to test by simulating the real and threatening conditions that could confront the child.

‘Two scenarios were constructed. In each, the would-be stranger requested the child’s assistance. In the pre-test, he asked the child to accompany him to his car to help bring treats in for his son’s birthday party. In the other simulation, the post-test, the child was
similarly asked to come to the stranger’s car to bring in puppets to be used in a puppet show: ‘Hello, I’m presenting a puppet show here at the school today. I have puppets and other neat things outside in my car. Will you come out and help me bring them inside?’ If the child agreed, he/she was told that the stranger would come for him/her later for help. After all children had been through the simulation and had returned to their classrooms, they were told that the stranger had gotten help from the school’s main office and would therefore not need their assistance’.

‘Each simulation was carried out in a way that could not be expected by the children, who were unaware that they were, in fact, being tested. They were not informed after the fact that they had had a simulated encounter. In both instances, teachers had fabricated an excuse to have the child leave the classroom and move to the designated area for the confrontation with the stranger. Collaborating office staff, classroom teachers, and school nurses took steps to ensure that no one else was present in the hallway in which the simulation occurred and that the child could have departed the building with the stranger unseen by school personnel and/or other students. Thus, the major criterion to judgement of risk to the child’s becoming victimised was his/her compliance/non-compliance with the stranger’s request and other specific behavior in response’.

The encounters were videoed and a ‘pass’/’fail’ rating was applied to each meeting. The results of the trial are recorded in Table 8.

**Table 8 Outcomes of the Need-to Know simulation**

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre-test simulation</th>
<th>Post-test simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Passed (%)</td>
<td>Failed (%)</td>
</tr>
<tr>
<td></td>
<td>10 (43.5%)</td>
<td>13 (56.5%)</td>
</tr>
<tr>
<td></td>
<td>11 (52.4%)</td>
<td>10 (47.6%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>Passed (%)</th>
<th>Failed (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18 (78.3%)</td>
<td>5 (21.7%)</td>
</tr>
<tr>
<td></td>
<td>11 (52.4%)</td>
<td>10 (47.6%)</td>
</tr>
</tbody>
</table>

Source: Fryer et al. (1987)

In one sense the experiment is a success, in that the scheme is associated with a marked increase in the ability to say ‘no’ to the (simulated) stranger approach - of 13 children who originally failed, 8 responded correctly to the safety rules stressed in the initiative, with no lapses from those who had originally passed the test. Nevertheless, a failure rate of over twenty per-cent for those who have gone though a programme is worthy of note - for safety education’s sake and from a methodological point of view, since such a high level would be considered troubling in paper-and-pencil testing. Yet more perplexing is the behaviour of the controls. In this case, there was a two-way flow in perception of the encounters - three from pass to fail and three from fail to pass.

These are dramatic results. For the umpteenth time, we see that a programme works for some subjects and not others. Fryer and colleagues make some impression on this puzzle in their discovery that the ‘improvers’ within the experimental group were superior to the two-time failures in their scores on the Harter self-esteem test. This suggests, perhaps, that gentler souls needed a somewhat different approach from the ‘concrete rules’ approach of this particular initiative. The cross-over behaviour of six of the control children is even more eye-opening. The two stranger-encounters are clearly not parallel in their minds, the attractions of the ‘birthday’ and ‘puppet show’ scenarios taking hold in different ways. This takes us back to a point made earlier about the importance for evaluation of studying how programmes work. The child’s ability to discriminate between, and apply, safety rules needs to be understood in and through a range of contexts. More’s the pity that Fryer did not posses the resources to investigate scenarios other than these two particularly compelling ‘invitations’. The significance of work such as Warden et al. (1997) and Moran et al. (1997), examined in Section 3.6, on the credibility of different stranger-encounters is once again emphasised through this research.

The Need-to-Know example reminds us of the need to have close regard for potential differences between programme-based understanding and simulated encounters and real world situations. Just as we recommended that programme content is better the more it approximates with ‘reality’, the same is undoubtedly true of the measurement apparatus used to monitor programme outcomes. So, should simulations become more commonplace in evaluation? The ethical dilemmas involved in this particular inquiry have made many in the safety field run a mile from the ‘simulation’ as a feasible outcome test. In other fields, in which the programme goal has no implications of threat or danger, simulation is a straightforward option. One example is the use of ‘confederates’ in littering research (Huffman, 1995) whose task is to seed an area with rubbish in order to provide a field-test for tidy minds. In short, we are not about to recommend that, following road safety lessons, schools be surrounded by parked cars with lorries readied to hurtle by, but it remains important to arrange for children to be confronted with some limited traffic conditions in order to get an (almost) real feel for actual lessons learned.

Lesson 18. Measuring programme outcomes by simulating the situations in which they are intended to have effects is an important, if limited, method for the evaluator. Provided ethical and practical issues can be dealt with, the more true-to-life the outcome measures the better.

3.19 Retrospective search for outcomes

Programmes generally walk a path from understanding to attitude change to behavioural change. As we have seen, it is the last step that is hardest for the programme to secure and the evaluator to monitor. In areas like child abuse prevention, it is especially difficult for the research and policy communities to know whether children are truly capable of resisting the approach of motivated older
offenders. One potential methodological solution is for research to travel the pathway in the opposite direction by conducting a survey which involves parents and children recounting their experiences of programmes, of attitude change, of knowledge gain and, if appropriate, actual victimisation.

Finkelhor and colleagues (1995a, 1995b, 1995c) conducted the largest study of this type that we encountered in our review. It consisted of a telephone-survey with 2000 children (10-16 years of age) as well as their ‘caretakers’. The survey was sectioned, asking for details about ‘school-based program exposure and content’, ‘parental instruction’, ‘knowledge about sexual abuse’, ‘victimisation experiences’, ‘self-protection strategies’ and ‘self-perceived efficacy in coping with confrontations’. The key benefit of the survey is that it allows far greater coverage of children who have actually been threatened or victimised than is the case with a standard programme trial. Finkelhor’s overall finding is that taking part in a victimisation prevention programme does not reduce the chance of being victimised, but there is much to learn from the other data. The following two paragraphs present a summary of the main findings:

‘The National Youth Victimization Prevention Study interviewed a representative sample of 2000 U.S. children and their caretakers about the children’s experience with child abuse and victimisation prevention programs. Two-thirds of the children reported being exposed to at least one program at some time, 37% within the last year. Programs that gave children a chance to practice, that promoted discussions with parents, and that included information on dealing with bullies were more likely to result in utilisation of program skills. Although satisfaction levels were generally high for all groups; girls, black children, and children from lower socio-economic status families, as well as their parents, had more positive reactions and reported more skill utilisation. Some children did report, and their parents confirmed, more worry about abuse and fear of adults. However, the children with increased fear were also the children who themselves and their parents reported the most positive feelings about the programs and the most skill utilisation. This suggests that the level of worry and fear induced by the programs was appropriate to the subject’. (1995a).

‘Exposure to a more comprehensive prevention program was not associated with reduced incidence of victimisation, injury, or upset. However, some of the exposure conditions were associated with an increased likelihood that the children would disclose victimisation, an increased likelihood that they would see themselves as having successfully protected themselves, and a decreased likelihood that they would blame themselves for the episode. Exposed children acquired some knowledge about sexual abuse and, when actually confronted by a threat, an ability to do the things they had been taught. A non-significant trend was also noted toward increased injury for exposed children during sexual assault.’ (1995b)

This presentation of results strung across two papers is actually extremely instructive for the methodologist, for it captures rather well the strengths and weakness of post-hoc evaluation using a survey approach. We follow the two listings down, giving a brief summary of pros and cons. The first item provides important basic information on the overall coverage and take-up of such programmes and a national survey is uniquely positioned to do so. Second, we have evidence on ‘knowledge gains’, ‘approval ratings’ and ‘reactions’ in relation to children’s experiences of programmes. We have seen that the encounter with a programme is always complex and that the details of that interaction are crucial to its success. Our report has already scoured the problems of ‘self-report’ and we note that a retrospective survey, of course, is entirely bound to such information as well as having the added disadvantage that the self-report is based on recollection, some of it taking place over rather long periods. So although surveys have the advantage of independence from a programme and thus avoid the problems of self-serving researchers and crowd-pleasing subjects, we are inclined to believe that the method is of restricted use in understanding the inner-workings of a programme. Whilst the National Youth Victimisation Prevention Survey came up with some useful outline demographics of the ‘reactions’ to programmes, it did not operate at a level which would inform any fine-tuning of initiatives.

The survey cashes in, however, when it comes to understanding the longer term outcomes of a programme and the findings about victimisation, its disclosure, self-blame, and the possible difference in injury-levels would be hard to trace in a conventional programme-based evaluation. Assigning these differences precisely to programme effects is, however, a problem; since contact with the programme is traced entirely retrospectively. We have seen that it is not merely attendance on a programme but the nature of the encounter which is important, and reconstructing that experience through a survey interview is most difficult. Also to be stirred into the cost-benefit pot is the problem of non-response in surveys on sensitive topics and we note that Finkelhor reports some problems on this score (1995c). The survey, in short, can get us further along the outcome trail than most methods but the causal inferences drawn back to programmes are of a distinctly broad-brush variety.

Road safety research could find room for a general survey in order to gain a broad picture of programme outcomes and (non-fatal) victim experiences. Most children will have had some experience of road safety education and some mature (if retrospective) reflection on its consequences would be of value. As ever, we are describing a method that only delivers part of the story, namely an outline overview of outcomes. Commissioning a specialist inquiry to establish such broad patterns of influence is probably wasteful of resources. These days space in large-scale surveys is up for hire and some carefully chosen, well-located questions
Programmes are human constructs and so programmes, dear things, never work out quite as planned. One of the great differences between lay and social scientific understanding is that the latter is primed to anticipate that human action will always generate unintended consequences and so the search for ‘perverse effects’ is part of the classic methodological agenda (Merton, 1968). In evaluation research this requires that researchers should include anticipation of the unanticipated consequences of an intervention as part of the research design. At one level this sounds completely barmy - how can one direct a systematic search for something one doesn’t know will happen? The solution to the riddle is that because they are so frequent, it is possible to develop an ‘eye’ for where to look for the unintended and the following paragraphs offer some standard sub-types of side-effect.

i ‘Turning-against’ or ‘overcompensating’ in respect to a programme idea.

Many programmes aimed at children are targeted at establishing norms for appropriate behaviour - and so work by distinguishing right and wrong, good and bad, proper and improper, safe and unsafe. Alas, such normative standards are hard to define and in particular the ‘dividing line’ is often conceptually slippery. One class of unintended consequences may thus occur because the subjects are too bold or too timid in relation to the normative standards assumed in the programme. Programmes thus have the potential to backfire for a substantial minority when programmes intimidate when designed to calm fears, or beguile when they are supposed to repel (e.g. road safety campaigns producing fear of traffic, sexual abuse campaigns producing fear of touching, stranger-danger campaigns producing fear of approach, scarifying straight campaigns scaring silly or crooked, anti-drugs campaigns conniving with a what’s-to-lose mentality, and anti-smoking campaigns evoking a why-worry response).

In respect of this class of programmes, most researchers are in fact attuned to look for negative side-effects. The problem of the fine dividing line between the safe and the unsafe, for instance, has long been recognised as the ‘good touch-bad touch’ issue in the child-abuse prevention literature. As we have seen, the main effects of a programme are dependent on the precise content of the programme message and the characteristics of the recipients. And so it is with unintended effects, with some studies reporting unwanted effects (Taal and Edelaar, 1997) and others (Binder et al., 1987) discovering the children-subjects to be “free of unduly negative sentiments about touching”. Taal and Edalaar’s research is a study in the subtleties of what-works-for-whom-in-what-circumstances-and-in-what-respects and their understanding of side-effects is equally discriminating, as when against expectations they discovered:

‘Older children experienced more negative feelings about physical touches in the long run, while younger ones felt more positive, but only in the short term. The impending puberty of the older elementary school pupils may sensitize them more than the younger pupils about information concerning sexual relationships. Because many innocent touches in the program’s exemplar situations turn out to be precursors to unintended sexual intimacy, this makes older children wary of the physical touch.’

ii Sleeper effects in educational programmes

Another class of unanticipated effects emanates from programmes which are prolonged, have diverse participants, and involve the sharing of experiences on many different matters. We are talking, of course, about educational programmes. As the saying goes, ‘education is for life’, and this should prime the researcher to anticipate a rich tapestry of outcomes with potential shifts occurring in knowledge, self-confidence, aptitudes, practical skills, inter-personal skills, cognitive development, beliefs, morality and so on. What makes educational outcomes particularly tough to detect is that these attributes can be ‘cashed in’ at different points in the life-course, with some effects indeed being ‘for life’, with others ‘fading’ and yet others ‘sleeping’.

The classic exemplification of the ‘dormant effect problem’ lies in evaluations of certain US pre-schooling initiatives. Programmes such as Head Start and Sesame Street, were based on the notion that educational disadvantage can be overcome by interventions at the kindergarten stage when especially profound cognitive development is taking place. Initial evaluations confirmed the picture with programme participants substantially outperforming controls in terms of IQ gains. However, a few years down the track, such gains were discovered to have dissipated totally, with the IQs of the experimental subjects and controls returning to parity. Considerable patience on the part of funders, policy-makers and researchers was required to discover that certain sleeper effects only really stirred during the teenage years, when it was observed that children in the head-start groups began to do better in terms of school drop-out, referral to special services, employment, and, most surprisingly, in terms of delinquency rates.

The discovery of these unanticipated long-term effects occurred in the much loved and loathed High Scope project (Schweinhart et al. 1993), which has tracked the lives of 123 children who lived in the neighbourhood of the Perry Elementary School in Ypsilanti, Michigan. The study has reached its 27th year and is still going strong.
Schweinhart explains the cross-over of the ‘intelligence’ and ‘social responsibility’ outcomes as follows:

‘Clear evidence of fade-out has been found only for gains in children’s intelligence-test scores. In the 1960s, the hypothesis was that even though early educational programs were found to raise young children’s test scores, subsequent educational programs would not affect them. Instead, it may be argued, a difference in intelligence test scores reflects a difference in educational settings. When children who have attended pre-school programs and children who have not attended pre-school programs come together into the same, standard elementary school classrooms, their intelligence test scores also come together.’

‘The essential process connecting early childhood experience to patterns of improved success in school and the community seemed to be the development of dispositions that allowed the child to interact positively with other people .... The pre-school program seemed to create for them a chain of events that led to their assuming greater social responsibility.’

Whilst the validity of these particular propositions has still to be fully accepted, this class of unanticipated effects shows the significance of the emergence of the right context for the benefits of certain programme mechanisms to be triggered into action. Such a ‘time-watch’ for outcomes may be a prerequisite of the evaluation of all educational initiatives, a lesson which McGuire et al. (1997) draw for the mental health field as follows:

‘Focus on one easily measured and immediate outcome, the IQ score, turned out to be misleading and could have contributed to a beneficial intervention being discontinued. The initial decision to rely on such a narrow outcome did not take into account the context of the children’s lives, the multiple agencies involved in their care, or the long term benefits, which were behavioural and emotional rather than cognitive. The lessons are clear. Before evaluating child and adolescent mental health services we must identify appropriate research designs and meaningful outcome measures while having realistic expectations about the utility of short-term results for informing policy decisions’.

iii Laws will be circumvented, technology will fail.
A further class of unintended consequences falls into the ‘sod’s law’ category. The two main recipients are legal and technological interventions. Human cussedness or stupidity will, in the last analysis, always outwit technology and the same qualities will lead eventually to the circumvention of laws. Technological loopholes are legion. We have already seen in this report the perverse effect of teenagers acting up to ‘hidden’ CCTV cameras and smoke alarm batteries finding their way into

Walkmans. In the home environment, tamper-proof containers have been shown to tempt certain children (Vernberg, in Towner et al., 1996). In the transport field, some road users can be observed trying to avoid or overcome safety devices – pedestrians avoiding guard rails, cyclists ignoring signals, motorists jumping traffic lights, etc. The widespread use of automobile airbags has saved numerous lives but, in America, has also contributed to the deaths of a small number of children (and small adults) seated in the front passenger seat.

Twisting and turning under the power of the law is also commonplace. A rather exotic example unearthed in the review was a ban on the local sale of alcohol to reduce drink-related violence on a native American reservation. This led to an increase in pedestrian and hypothermia deaths as certain desperate souls were forced to venture further from home in search of booze (Gallaher et al., 1992). In a different sphere and at the corporate level, it can be seen that cigarette advertising has transformed itself with the speed of a formula one car in the face of a series of legal restrictions.

Again, we see that - to some extent - the unanticipated may be anticipated in programmes relying only on technological and legal interventions. Researchers should always expect a certain amount of failure, side-stepping and back-sliding around such programmes. Taking up the air-bag example, or indeed seat-belt initiatives which preceded it, we see that a series of further technical refinements and publicity and education were needed to derive full benefits. Indeed, one of the most consistently successful programme types that we encountered in the safety field was the model in which legal enforcement is combined with educational promotion on the benefits of the new law. Cote et al. (1992) and Dannenberg et al. (1993) research on cycle helmet utilisation took advantage of the fact that legislation can operate at the county level in the US. They were able to organise a revealing comparison which contrasted the fortunes of three cycle-helmet programmes - one ‘education-only’, one ‘law-only’, one ‘law-plus-education’, the latter county experiencing massively greater rates of helmet-usage.

In the last analysis, of course, some unintended effects of programmes will remain unanticipated in research and only be discovered accidentally. Danneberg and Fowler (1998) give the example of a Massachusetts law, designed for its environmental benefits, which mandated a deposit on glass bottles. The introduction of the new legislation coincided with a safety review of emergency department records, which revealed a decline in the number of glass related lacerations in children after the law went into effect. Danneberg also quotes the US ‘right-turn-on-red’ law, which was promoted to save fuel and lessen driver impatience but led to an increase in pedestrian injuries. No ‘method’ exists for linking such a class of unintended ‘cross-departmental’ effects. Perhaps the only way for them to be uncovered is for researchers and policy-makers to keep an eagle-eye open for reports of legal and environmental changes ‘outside’ their own policy-spheres which might just nudge their way in.


Lesson 20. Evaluation research should always seek a multiplicity of outcomes, including those which fall outside the intentions and expectations of programme-architects. Unanticipated outcomes of programmes flow from programme complexity, subject over-reaction and bloody-mindedness, technological failure, legal loopholes and, occasionally, from nowhere.

3.21 Meta-analysis and review

Our own efforts produced one unanticipated lesson for us, which concerned the sheer volume of existing research in the area under review. We have offered a glimpse of the scale of enterprise through the use of a diversity of illustrative studies. However, for each case study we have included, there stand three or four more that were read but which did not make it into these pages because they repeated methodological lessons already registered. And for each of these unreported studies stand dozens of unread studies which have accumulated as research activity throughout the world has come to focus on the care of children.

We have already argued that researchers have to anticipate the unexpected and this flood of evaluations really could have been foretold. The lifeblood of policy-making is attempting to make progress through ‘fresh ideas’ and it is to the benefit of politicians, policy-makers, practitioners, evaluators and methodologists all that such a supply-line continues. The social organisation of evidence-led policy-development has a fixation with the new and a tendency to chase the latest initiative, a feature embodied in the appellation - ‘demonstration programmes’. Confronted with a bursting filing-cabinet of previous studies, however, one is tempted to believe that there is nothing new under the sun. Ideas turn up again and again because of the limited leverage of social interventions. Policies and programmes are aimed at bringing about individual and social change by any means legal, decent, honest, truthful. And yet policy-makers and practitioners are, in truth, able to offer relatively few ways of inducing change (our assertion here is not meant to reflect upon the lack of imagination of programme architects but on the limited nature of the incentives they may draw upon).

The line of thought above has led to calls, which we wholeheartedly support, for programme development to be much more cumulative, with initiatives being contained within, and supported by, an ongoing process of review of existing research. This makes perfect methodological sense. We have made the point repeatedly about the difficulties of ‘reading across’ from one-off evaluations. The ultimate question for the policy-maker, of course, is not - did this programme with all its particulars of delivery and personnel manage to work in this situation with all its peculiarities of locality and history? The realistic policy question is better rendered - would this sort of programme manage to work in a reasonable range of circumstances and conditions?

This challenge of measuring the success of families of programmes across their different incarnations has been taken up in the form of evaluation techniques variously described as ‘meta-evaluation’, ‘synthesis’, or ‘review’. Once again, we were surprised by the volume of such studies. In certain fields of health education, reviews are now so commonplace that reviews-of-reviews have now to be contemplated (Towner et al., 1996). The sublimities of meta-meta-evaluation will not be pursued here, however, since we witnessed some (entirely predictable) methodological blood at ground level in terms of disagreement about how best to extract evaluation’s collective lessons. Our final section is thus reserved for trying to draw out some conclusions about how the process of review itself should be conducted.

As a first and crude approximation, one can say there are polar opposite approaches to synthesising the results of previous inquiries, namely the ‘numerical’ and the ‘narrative’. Our brief remarks warn of the sterility of reviews which work entirely at the ends of this spectrum and seek to press a case for an as yet, rather underdeveloped and uncodified middle-way.

The numerical approach, or what is often just referred to as ‘meta-analysis’ is based on a comparison of outcomes across a whole family of programmes. It differentiates ‘sub-types’ of programmes within a particular family in an attempt to discover which one is associated with the greatest effect sizes. A typical set of results is illustrated in Table 9, which comes from Durlak and Wells’s (1997) meta-analysis of primary prevention mental health programmes.

<table>
<thead>
<tr>
<th>Type of program</th>
<th>n</th>
<th>Mean effect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Environment-centred</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School-based</td>
<td>15</td>
<td>0.35</td>
</tr>
<tr>
<td>Parent-training</td>
<td>10</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>Transition programs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Divorce</td>
<td>7</td>
<td>0.36</td>
</tr>
<tr>
<td>School-entry / change</td>
<td>8</td>
<td>0.39</td>
</tr>
<tr>
<td>First-time mothers</td>
<td>5</td>
<td>0.87</td>
</tr>
<tr>
<td>Medical/dental procedure</td>
<td>26</td>
<td>0.46</td>
</tr>
<tr>
<td><strong>Person-centred programs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affective education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Children 2-7</td>
<td>8</td>
<td>0.70</td>
</tr>
<tr>
<td>Children 7-11</td>
<td>28</td>
<td>0.24</td>
</tr>
<tr>
<td>Children over 11</td>
<td>10</td>
<td>0.33</td>
</tr>
<tr>
<td><strong>Interpersonal problem solving</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Children 2-7</td>
<td>6</td>
<td>0.93</td>
</tr>
<tr>
<td>Children 7-11</td>
<td>12</td>
<td>0.36</td>
</tr>
<tr>
<td>Children over 11</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td><strong>Other person-centred programmes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behavioural approach</td>
<td>26</td>
<td>0.49</td>
</tr>
<tr>
<td>Non-behavioural approach</td>
<td>16</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Source: Durlak and Wells, 1997

The essential problem with such numerical meta-analysis is betrayed in that tight-fisted term - the ‘mean effect’. This is an aggregate of aggregates. The net effect of any particular programme is made up of the balance of successes and failures of individual subjects. As we have seen, this overall outcome depends not merely upon ‘the
programme’ but on its subjects and its circumstances. At the extreme, we can still learn from a negative net effect since the application of an initiative to the wrong subjects and circumstances can leave behind vital clues about what might be the right combination. No such subtlety is available in a purely numerical meta-analysis which simply extracts the net effect from a study and combines it together with other programmes from the same stable to generate an average effect for that class of programmes. Vital explanatory information on the different ‘processes’ and ‘contexts’ associated with each initiative is squeezed out automatically in the process of aggregation.

Worse still is the wholesale and instant elimination of a whole rank of existing evaluations from the purview of formal meta-analysis. One can only include in the calculation of ‘mean effect’ studies which generate a clear ‘net effect’ and that, generally speaking, generates a hash calculation of ‘mean effect’ studies which generate a clear formal meta-analysis. One can only include in the whole rank of existing evaluations from the purview of formal meta-analysis. One can only include in the calculation of ‘mean effect’ studies which generate a clear ‘net effect’ and that, generally speaking, generates a hash pre-qualifying rule along the following lines:

‘.. only papers which described an experimental or quasi-experimental design were included in the review. All papers which did not report an evaluation with pre-and post-intervention measures and a control group were also rejected.’ (Foxcroft et al., 1997)

Foxcroft et al. (1977) review of alcohol misuse programmes for young people thus began with the pre-screening of 500 potential studies, information was abstracted from 155, with only 33 studies ‘meriting’ inclusion in the review. And on the basis of these, the authors conclude that ‘the lack of reliable evidence means that no one type of prevention programme can be recommended’ (1997). This is a preposterous example of ‘methodolatory’. Apart from arrogance of the instant dismissal of the efforts of the best part of 500 research teams, the weaknesses of RCTs, noted earlier and themselves considerable, are ignored.

What has to be resisted in meta-analysis is the tendency for making policy-decisions on the casting of an eye down a net effects column such as in Table 9. This can be illustrated by the response of two authors on the receiving end of Durlak and Wells’s review. Weissberg and Bell (1997) were responsible for three out of the twelve studies reviewed on ‘interpersonal problem solving for 7-11 year-olds’ and their efforts are thus down there in the lower, ‘disposable reaches’ of net-effects league-table. They protest that three inquiries were in fact part of a developmental sequence which saw their intervention change from one with 17 to 42 lessons, and as programme conceptualisation, curriculum, training and implementation progressed, outcome success also improved in the three trials. They also point out that programmes frequently outgrow their meta-analytic classification. Initiatives are rarely just ‘person-centred’ or ‘environment-centred’ or whatever, but inevitably, and especially in the best programmes, contain elements of several programme theories.

Whilst this exchange is presented here in a negative light for meta-analysis, it does actually show the way forward in that an explanation is developed for a small part of the overall outcome pattern. Data never speak for themselves and this maxim applies with a vengeance to highly compressed data of meta-analysis. The usual rule (Section 3.2) about outcome patterns being explained by process applies just as well to meta-analysis, the way forward being to have a sufficient grasp of the particulars of the individual studies to be able to make sense of features of the collective outcome configuration. Though we found no study in the review to exemplify it, we note that advances in meta-analysis (Cook, 1992) in certain other fields are indeed dealing with the possibility that programmes have multiple outcomes and are internally complex, and are striving to incorporate within discussion a closer understanding of what happened on the ground within each programme.

Narrative reviews, by contrast, start explicitly from a ground-level view of what happened with each programme. Their first task is to offer a summary of those interventions under scrutiny, which preserves the integrity of each original study. Beginning here confronts the approach with the polar opposite problem of how to shed some collective light on a family of initiatives. In our examples, we begin with forms of narrative review that provide the dimmest of overviews and then move onto approaches which provide more general illumination. Walker’s Injury Prevention for Young Children: A Research Guide (1996) provides an example, which despite its sub-title is little more than a bibliography or index. The format is a ‘catalogue’ of what look suspiciously like the abstracts of 370 studies. These are subdivided into nine sub-sections dealing with different forms of injury (asphyxiation to vehicle injury) with brief, paragraph-length, sectional introductions. Whilst these pocket research profiles provide enough material to spark interest in particular entries and indifference with respect to others, such a process hardly begins the job of comparing and contrasting effectiveness. Seemingly, this task is left to the reader, who is in a hopeless position to do so systematically, since the entries merely log the concerns of the original authors and thus contain utterly heterogeneous information.

A giant step on from here, within the narrative tradition, is what is sometimes called the ‘descriptive-analytical’ method. Here studies are combed to a common analytical framework, with the same template of features being applied to each study scrutinised. A model example, close to the substantive concerns of this report, is to be found in Towner et al. (1996). Appendix H in that study supplies an elegant example of a ‘data extraction form’, alas too long to reproduce here, which is completed for each study reviewed, collecting information as follows:

1 Author, year of publication, place.
2 Target group, age-range, setting.
3 Intervention aims and content.
4 Whether programme is educational, environmental or legislative.
5 Whether alliances of stakeholders were involved in programme implementation.
6 Methodology employed.

6 Methodology employed.
7 Outcome measures employed.
8 Summary of important results.
9 Rating of the ‘quality of evidence’.

Accounts of each study are then tabulated (Towner et al., 1996 appendix I), so one can literally ‘read across’ from case to case, comparing them directly on any of these features. Such a procedure provides an incomparable overview of ‘what is going on’ in evaluation and some broad brush indications of interventions for which there is reasonable evidence of success (1996, executive summary) - but not yet a fully articulated methodology to guide future research and policy thinking. In the Towner report such recommendations are appended (1996, Appendix D) as a prudent seven pages of ‘discussion’ which list some exemplary studies and items of good practice across the above nine areas. Whilst this avoids skilfully the meta-analytic tendency to pronounce winners arithmetically, the logic behind the choices of ‘honourable mentions’ is not quite clear nor codified.

To conclude, the process of review is absolutely vital to evidence-led policy-development. It embodies the key principle of building initiatives by learning from past successes and failures. It involves no battles with gatekeepers, no subject sampling, no idle control groups, no subject interviewing and, indeed, no programmes to run and maintain (and is thus remarkably cheap!). A circumspect blend of demonstration projects and review is thus recommended as the basis for programme development. It has to be acknowledged that the issue of finding the precise criterion for making meta-analytical judgements has yet to be solved. We caution here against the two extremes of the ruthless arithmetic of net success versus the intuitive extraction of exemplars from examples.

As a last word, we offer a couple pointers to methodological progress on this conundrum. This first is the need to narrow the focus of meta-analytic comparisons, which often involve gigantic overviews across whole policy fields and thus chase down quite incomensurable ideas. It is to be recalled that programmes, which are theories, and ‘families of theory’ rather than ‘policy responsibilities’, are the appropriate unit for meta-analysis. We cannot imagine learning much by comparing the relative merits of offering deposits on coke-bottles to reduce littering and taking juveniles on prison visits to reduce offending, even though they are based on similar target groups. Much more instructive is exploration of the same programme theory, say, ‘peer-pressure’ as played out in the different domains of drugs, smoking, diet, exercise, safety etc. (Milburn 1995). This leads to a second desideratum which is to adapt the meta-analytic question itself. We have already argued against asking of any single programme the question, ‘does it work?’, because research can only ever reply, ‘that depends on the subject and the circumstances in which it is applied’. Meta-analysis can never deliver on the question, ‘what type of programme works best’ for exactly the same conditionality applies. The question for the reviewer should be the familiar one - what is it about this family of programmes which works for whom in what circumstances and in what respects?

**Lesson 21.** Researchers should never evaluate a programme ab ovo, they should always inspect previous broken eggs. Policy-makers attempt to produce new responses to old problems but the chances are some very similar initiative will have been put into place sometime, somewhere. Periodic reviews of existing research are thus vital. The process of review itself, however, takes on many forms from the catalogue-of-abstracts to quantitative-meta-analysis. The most beneficial way to conduct a review is to study each piece of research in the light of its contribution on ‘what works for whom in what circumstances.’

4 Conclusion

In these closing remarks, we make no attempt to recapitulate what is already a long argument. Instead, we look forward to the ultimate objective of this exercise, which uses the evaluation apparatus suggested here in order to put to empirical test some of the new generation of road safety programmes as they enter the policy arena. Our purpose here is neither to suggest nor second-guess what these initiatives should be. Whilst policy-making might aspire to be evidence-driven, we have yet to see the case made that it should be methodology-inspired. Nevertheless, if the research strategy suggested here is correct, it does have broad implications for the typical make-up of successful programmes. Certain forms of evidence have been identified as crucial to the evaluator and it is clear that programmes will work only if they are active at those very points.

We return to the beginning in lesson one (Section 3.1) to make the point. Our very first remark was that programme success is embedded within the ideas, individuals, institutions and infrastructure that constitute initiatives. We argued that the choice between ‘bottom-up’ and ‘top-down’ policy-making was somewhat bogus in that bright, grassroot ideas only tend to seed if they are persuasive in the wider forum and, conversely, that successful legislative change needs to carry the will of individuals to whom it is directed. This elementary point about the need to envisage programmes as ‘packages’ whose ideas drive both individual and institutional change was a feature of several of the policy-fields we inspected.

Towner (1995) makes a similar point in her critique of the trifurcation of policy-making in health education into the preventative model (stressing individually-focused instruction), the radical model (stressing social and environmental change) and the empowerment model (stressing the need for ‘subjects’ to achieve control of the change process). Such text-book simplifications, alas, hold considerable sway across much academic policy-thinking. Towner, by contrast, notes that the key characteristic of successful initiatives lies in the combination of these three broad mechanisms and provides examples of their incorporation in a series of fruitful ‘community’, ‘neighbourhood’ and ‘health alliance’ programmes. She is
particularly scornful of the simplification which assumes that educational initiatives are always classroom based and are directed down into rows of seated children. Health education, she points out, often has its greatest impact in persuading health workers, policy-makers and the general public that legislative and environmental change is the order of the day.

This desideratum for programme ideas to impact on individuals, institutions and infrastructure must not be mistaken as a plea for gigantic, centralised, multi-agency, policy-interventions capable of launching a whole battery of education, environment and legislation change. Indeed this capacity to create linkages can be the characteristics of a quiet local scheme. The simplest road safety instruction is deepened when taken out onto the ‘mock’ road, toughened when tested on real roads, anchored when it involves parents supervising the school walk, broadened when it seeks to combine other goals such as improved fitness levels and air-quality, strengthened when it is supported by traffic management regulation from the local authority, invigorated when supplied with oxygen of imitation and mass-media support, and authenticated when the evaluator discovers the combination of subjects and circumstances to harness such a process.

As conscientious reviewers, we leave the final word to one of our key sources:

‘There are still many gaps in our knowledge of how best to influence policy makers, professionals, children, parents and the wider community. Very few evaluated studies pay any attention to measuring process. It is important to be able to identify successful and unsuccessful components of campaigns and understand why a programme works in specific circumstances and localities and not others. Only then can we transport those effective components to other areas and to learn from earlier mistakes. There has been a false distinction between the role of health education, environmental measures and legislation: in practice it is their inter-linked and synergistic effect that is important. The challenge is to make the educational process more effective in all the contexts in which it takes place’ (Towner, 1995)

6 References


5 Acknowledgements


Shiner M and Newburn T (1996). *Young people, drugs and peer education - an evaluation of the Youth Awareness Programme (YAP).* Home Office.


Abstract

This report describes a review of the concepts and principles underlying evaluation methodologies used in a range of child health and safety programmes. One objective was to identify any recent innovative techniques in these areas which could be transferred to the road safety education field. The conclusions are presented as twenty-one key elements of good practice that should underpin evaluation in all areas, but no fundamentally new techniques were identified to add to existing practice in road safety evaluation.

Related publications

TRL360  *Helping parents to protect pre-school children from accidents* by G Davies, J Davies, G Harland, G Murray and S Levene. 1998 (price £35, code H)

PR117  *The high risk child pedestrian: socio-economic and environmental factors in their accidents* by N Christie. 1995 (price £25, code E)

CT45.3  Child safety on the road update (1998-2001) *Current Topics in Transport: selected abstracts from TRL Library’s database* (price £20)

CT78.1  Road safety education update (1995-1998) *Current Topics in Transport: selected abstracts from TRL Library’s database* (price £20)

Prices current at August 2001

For further details of these and all other TRL publications, telephone Publication Sales on 01344 770783 or 770784, or visit TRL on the Internet at www.trl.co.uk.