ESRC UK Centre for Evidence Based Policy and Practice: Working Paper 3

Evidence Based Policy: I. In Search of a Method

Ray Pawson

ESRC UK Centre for Evidence Based Policy and Practice
Queen Mary
University of London

r.d.pawson@leeds.ac.uk

Pre-publication version: submitted to *Evaluation*

Ray Pawson is Visiting Senior Research Fellow at the ESRC UK Centre for Evidence Based Policy and Practice, Queen Mary, University of London

Abstract

Evaluation research is tortured by time constraints. The policy cycle revolves quicker than the research cycle, with the result that 'real time' evaluations often have little influence on policy making. As a result, the quest for Evidence Based Policy (EBP) has turned increasingly to systematic reviews of the results of previous inquiries in the relevant policy domain. However, this shifting of the temporal frame for evaluation is in itself no guarantee of success. Evidence, whether new or old, never speaks for itself. Accordingly, there is debate about the best strategy of marshalling bygone research results into the policy process. This paper joins the imbroglio by examining the logic of the two main strategies of systematic review, namely 'meta-analysis' and 'narrative review'. Whilst they are often presented as diametrically opposed perspectives, this paper argues that they share common limitations in their understanding of how to provide a template for impending policy decisions. This review provides the background for Part II of the paper which considers the merits of a new model for EBP, namely 'realist synthesis'

Key words: Evidence-Based Policy, Methodology, Systematic Review, Meta-analysis, Narrative Review, Realism

Evidence Based Policy: I. In Search of a Method

Introduction

My topic here is 'learning from the past' and the contribution that empirical research can make to that cause. Whether policy makers actually learn from or simply repeat past mistakes is something of a moot point. What is clear is that policy research has much to gain by following the sequence whereby social interventions are mounted and in trying, trying, and then trying again to tackle the stubborn problems that confront modern society. This is the *raison d'être* behind the current explosion of interest in evidence-based policy (EBP). Few major public initiatives these days are mounted without a sustained attempt to evaluate them. Rival policy ideas are thus run through endless trials with plenty of error and it is often difficult to know which of them really have withstood the test of time. It is arguable, therefore, that the prime function of evaluation research should be to take the longer view. By building a systematic evidence base that captures the ebb and flow of programme ideas, we might be able to adjudicate between contending policy claims and so capture a progressive understanding of 'what works'. Such 'resolutions of knowledge disputes', to use Donald Campbell's phrase (Campbell and Russo, 1999 part III), are the aim of the research strategies variously known as 'meta-analysis', 'review' and 'synthesis'.

Standing between this proud objective and its actual accomplishment lies the minefield of social science methodology. Evaluation research has come to understand that there is no one 'gold standard' method for evaluating single social programmes. My starting assumption is that exactly the same maxim will come to apply to the more global ambitions of systematic review. It seems highly likely that EBP will evolve a variety of methodological strategies (Davies, 2000). This paper seeks to extend that range by offering a critique of the existing orthodoxy, showing that the lessons currently learnt from the past are somewhat limited. These omissions are then gathered together in a second part of the paper to form the objectives of a new model of EBP, which I call 'realist synthesis'.

Section 1.1 of the present paper offers a brief recapitulation of the overwhelming case for the evaluation movement to dwell on and ponder evidence from previous research.

Sections 1.2 and 1.3 consider the main current methods for doing so, which I split into two perspectives, namely, 'numerical meta-analysis' and 'narrative review'. A detailed critique of these extant approaches is offered. The two perspectives operate with different philosophies and methods and are often presented in counterpoint. There is a heavy whiff of the 'paradigm wars' in the literature surrounding them and much mud slinging on behalf of the 'quantitative' versus the 'qualitative', 'positivism' versus 'phenomenology', 'outcomes' versus 'process' and so on. This paper eschews any interest in these old skirmishes and my critique is, in fact, aimed at a common ambition of the two approaches.

In their different ways, both aim to review a 'family of programmes' with the goal of selecting out the 'best buys' to be developed in future policy making. I want to argue that, despite the arithmetic clarity and narrative plenitude of their analysis, they do not achieve decisive results. The former approach makes no effort to understand how

programmes work and so any generalisations issued are insensitive to differences in programme theories, subjects and circumstances that will crop up in future applications of the favoured interventions. The latter approach is much more attuned to the contingent conditions that make for programme success but has no formal method of abstracting beyond these specifics of time and place, and so is weak in its ability to deliver transferable lessons. Both approaches thus struggle to offer congruent advice to the policy architect, who will always be looking to develop new twists to a body of initiatives, as well as seeking their application in fresh fields and to different populations.

1.1 'Meta' is better

Whence EBP? The case for using systematic review in policy research rests on a stunningly obvious point about the timing of research *vis-à-vis* policy – namely, that *in order to inform policy, the research must come before the policy.* To figure this out does not require ESRC-approved methodological training, nor a chair in public policy, nor years of experience in Whitehall. Yet, curiously, this proposition does not correspond to the sequence employed in the procurement of most evaluation research. I depict the running order of a standard piece of evaluation (if there be such a thing) in the upper portion of Figure 1.

Figure one: The research/policy sequence

i. Evaluation Research (standard mode)

Design Implementation Impact

ii. Meta-analysis/Synthesis/Review

Design Implementation Impact

The most significant manifestation of public policy in the modern era is the 'programme', the 'intervention', or the 'initiative'. The gestation of such schemes follows the familiar design → implementation → impact sequence, each phase being associated with a different group of stakeholders, namely programme architects, practitioners and participants. As a rule of thumb, one can say that evaluation researchers are normally invited to enter the fray in the early phases of programme implementation. Another iron law of research timing is that the researchers are usually asked to report on programme impact somewhat before the intervention has run its course. Evaluations, therefore, operate with a rather narrow bandwidth and the standard 'letting and reporting' interval is depicted in Figure 1.

Much ink and vast amounts of frustration have flowed in response to this sequencing. There is no need for me to repeat all the epithets about 'quick and dirty', 'breathless' and 'brownie point' evaluations. The key point I wish to underscore here is that, under the traditional running order, programme design is often a research-free zone. Furthermore, even if an evaluation manages to be 'painstaking and clean' under present conditions, it is still often difficult to translate the research results into policy action. The surrounding *realpolitik* means that, within the duration of an evaluation, the direction of political wind may well change, with the fundamental programme philosophy being (temporarily) discredited, and thus not deemed worthy of further research funding. Moreover, given the turnover in and the career ambitions of policy makers and practitioners, there is always a burgeoning new wave of programme ideas waiting their turn for development and evaluation. Under such a regime, we never get to the Campbellian 'resolution of knowledge disputes' because there is rarely a complete revolution of the 'policy-into-research-into-policy' cycle.

Such is the case for the prosecution. Let us now turn to a key manoeuvre in the defence of using the evidence base in programme design. The remedy often suggested for all this misplaced, misspent research effort is to put research in its appropriate station (at the end of the line) and to push many more scholars back where they belong (in the library). This strategy is illustrated in the lower portion of Figure 1. The information flow therein draws out the basic logic of systematic review, which takes as its starting point the idea that there is nothing entirely new in the world of policy making and programme architecture. In the era of global social policy, international programmes and cross-continental evaluation societies, one can find few policy initiatives that have not been tried and tried again, and researched and researched again. Thus, the argument goes, if we begin inquiry at that point where many similar programmes have run their course and the ink has well and truly dried on all of the research reports thereupon, we may then be in a better position to offer evidence-based wisdom on what works and what does not.

On this model, the key driver of research application is the feedback loop (see Figure 1) from past to present programming. To be sure, this bygone evidence might not quite correspond to any current intervention. But since policy initiatives are by nature mutable and bend according to the local circumstances of implementation, then even real-time research (as we have just noted) has trouble keeping pace.

Like all of the best ideas, the big idea here is a simple one – that research should attempt to pass on collective wisdom about the successes and failures of previous initiatives in particular policy domains. The prize is also a big one in that such an

endeavour could provide the antidote to policy making's frequent lapses into crowd pleasing, political pandering, window dressing and god-acting. I should add that the apparatus for carrying out systematic reviews is also by now a big one, the recent mushrooming of national centres and international consortia for EBP being the biggest single change on the applied research horizon for many a year[1].

Our scene is thus set. If all goes to order, EBP will provide cumulative understanding distilled in a practical form that will allow, to adapt a phrase from Pease (1998), the public policy grease to get to the social problem squeak. I turn now to an examination of the methodological means to this end, and try to discover whether, in fact, all has gone to order.

1.2 Numerical meta-analysis

The numerical strategy of EBP, often just referred to as 'meta-analysis', is based on a three-step 'classify', 'tally' and 'compare' model. The basic locus of analysis is with a particular 'family of programmes' targeted at a specific problem. Attention is thus narrowed to the initiatives developed within a particular policy domain (be it 'HIV/AIDS prevention schemes' or 'road safety interventions' or 'neighbourhood watch initiatives' or 'mental health programmes' or whatever). The analysis begins with the identification of sub-types of the family, with the classification based normally on alternative 'modes of delivery' of that programme. Since most policy making is beset with rival assertions on the best means to particular ends, meta-evaluation promises a method of sorting out these contending claims. This is accomplished by compiling a database examining existing research on programmes making up each sub-type, and scrutinising each case for a measure of its impact (net effect). The overall comparison is made by calculating the typical impact (mean effect) achieved by each of the sub-types within the overall family. This strategy thus provides a league table of effectiveness and a straightforward measure of programme 'best buy'. By following these steps, and appending sensible caveats about future cases not sharing all of the features of the work under inspection, the idea is to give policy architects some useful pointers to the more promising areas for future development.

A typical illustration of numerical strategy is provided in Table 1, which comes from Durlak and Wells' (1997) meta-analysis of 177 Primary Prevention Mental Health (PPMH) programmes for children and adolescents carried out in the USA. According to the authors, PPMH programmes 'may be defined as an intervention intentionally designed to reduce incidence of adjustment problems in currently normal populations as well as efforts directed at the promotion of mental health functioning'. This broad aim of enhancing psychological well-being by promoting the capacity to cope with potential problems can be implemented through a variety of interventions, which were classified by the researchers as in column one of Table 1. First of all there is the distinction between person-centred programmes (which use counselling, social learning and instructional approaches) and environment-centred schemes (modifying school or home conditions to prepare children for life change). Then there are more

[1] For details of, and access to, EBP players such as the Cochrane Collaboration, Campbell Collaboration, EPPI-Centre etc., see the Resources section of the ESRC UK Centre for Evidence Based Policy and Practice website at http://www.evidencenetwork.org

specific interventions targeted at 'milestones' (such as the transitions involved in parental divorce or teenage pregnancy and so on). A range of further distinctions may also be discerned in terms of programmes with specific target groups, such as those focused on specific 'developmental levels' (usually assigned by 'age ranges') or aimed at 'high risk' groups (assigned, for instance, as children from 'low-income' homes or with 'alcoholic parents').

Table 1: The 'classify', 'tally' and 'compare' model of meta-analysis

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

[a]non-significant result
Source: Durlak and Wells (1997, p129)

A standard search procedure (details in Durlak and Wells, p120) was then carried out to find research on cases falling into these categories, resulting in the tally of initiatives in column two of the table. For each study the various outcomes assessing the change in the child/adolescent subjects were unearthed and 'the effect sizes (ESs) were computed using the pooled standard deviation of the intervention and control group'. After applying corrections for the small sample sizes in play in some of the initiatives, the mean effect of each class of programmes was calculated. This brings us finally to the objective of meta-analysis – a hierarchy of effectiveness – as in column three.

Let us now move to a critique of this strategy, remembering that I am seeking out weaknesses in the logic of the whole genre and not just this example. The basic problem concerns the nature and number of the simplifications that are necessary to achieve the column of net effects. This – the key output of meta-analysis – is, has to be, and is intended to be a grand summary of summaries. Each individual intervention

is boiled down to a single measure of effectiveness, which is then drawn together in an aggregate measure for that sub-class of programmes, which is then compared with the mean effect for other intervention categories. The problem with this way of arriving at conspectus is that it squeezes out vital explanatory content about the programmes in action in a way that renders the comparisons much less rigorous than the arithmetic appears on paper. This process of compression occurs at three points:

i) the melding of programme mechanisms
ii) the oversimplification of programme outcomes
iii) the concealment of programme contexts

i) Melded mechanisms

The first difficulty follows from an absolutely taken-for-granted presumption about the appropriate locus of comparison in meta-analysis. I have referred to this as a 'programme family' above. Now on this method, it is simply assumed that the source of family resemblance is the 'policy domain'. It one sense this is not at all unreasonable. Modern society parcels up social ills by administrative domain and assembles designated institutes, practitioners, funding regimes *and* programmes to tackle each problem area. Moreover, it is a routine feature of problem solving that within each domain there will be some disagreement about the best way of tackling its trademark concerns. Should interventions be generic or targeted at sub-populations? Should they be holistic or problem-specific? Should they be aimed at prevention or cure? Should they have an individual focus, or be institution-centred, or area-based? Such distinctions feed their way onwards into the various professional specialities and rivalries that feature in any policy sector. This in turn creates the field of play for meta-evaluation, which is asked to adjudicate on whether this or that way of tackling the domain problem works best. In the case at hand, we are dealing with the activities of a professional speciality, namely 'community psychology', which probably has a greater institutional coherence in the US than the UK, but whose sub-divisions into 'therapists', 'community educators' and so on would be recognised the world over.

Now, this is the policy apparatus that generates the programme alternatives that generate the meta-analysis question. Whether it creates a commensurable set of comparisons and thus a researchable question is, however, a moot point. And this brings me to the nub of my first critique, which is to cast doubt on whether such an assemblage of policy alternatives constitutes a comparison of 'like with like'. Any classification system must face the standard methodological expectations that it be unidimensional, totally inclusive, mutually exclusive and so forth. We have seen how the different modes of programme delivery and the range of targets form the basis of Durlak and Wells' classification of PPMH programmes. But the question is, does this classification system provide us with programme variations on a common theme that can be judged by the same yardstick, or are they incommensurable interventions that should be judged on their own terms?

One can often gain a pointer to this issue by considering the reactions of authors whose studies have been grist to the meta-analysis mill. 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 barely statistically significant efforts are thus to be found down there in the lower reaches of the net effects league table. They protest that their three inquiries were in fact part of a 'developmental sequence' which

saw their intervention change from one with 17 to one with 42 modules. And, as programme conceptualisation, curriculum, training and implementation progressed, outcome success also improved in the three trials. As well as wanting 'work-in-progress' struck out of the review, they also point out that programmes frequently outgrow their meta-analytic classification. So their 'person-centred' intervention really begins to bite, they claim, when it influences whole teaching and curriculum regimes and thus becomes, so to speak, 'environment-centred'. The important point for this pair of authors is that the crux of their programme, its 'identifier', is the programme theory. In their case, the proposed mechanism for change was the simultaneous transformation of both setting and teaching method so that the children's thinking became more oriented to problem solving.

Weissberg and Bell also offer some special pleading for those programmes (not of their own) that come bottom of the PPMH meta-evaluation scale. Parenting initiatives, uniquely amongst this programme set, attempt to secure improvements in children's mental health through the mechanisms of enhancing child rearing practices and increasing the child development knowledge of parents and guardians. This constitutes a qualitatively different programme theory, which may see its pay-off in transformations in daily domestic regimes and in long-term development in the children's behaviour. This theory might also be particularly sensitive to the experience of the parent-subjects (first time parents rather than old moms and dads being readier to learn new tricks). It also suffers from problems of take-up (such initiatives clearly have to be voluntary and the most needful parents might be the hardest to reach).

Such a programme stratagem, therefore, needs rather sophisticated testing. Ideally this would involve long-term monitoring of *both* family and child; it would involve close scrutiny of parental background; and, if the intervention were introduced early in the children's lives, it would have to do without before-and-after comparisons of their attitudes and skills. When it comes to meta-analysis, no such flexibility of the evidence base is possible and Durlak and Wells' studies all had to face the single, standard test of success via measures monitoring pre-intervention/post-intervention changes in the *child's* behaviour. A potential case for the defence of parenting initiatives thus lurks, on the grounds that what is being picked up in meta-analysis in these cases is methodological heavy-handedness rather than programme failure.

Note well that my point here is *not* about 'special pleading' as such. I am *not* attempting to speak up for 'combined' programmes or 'parenting' programmes or any specific members of the PPMH family. Like the typical meta-analyst, I am not close enough to the original studies to make judgements on the rights and wrongs of these particular disputes. The point I am underlining is that programme sub-categories just do not speak for themselves. The classifications are not just convenient, neutral labels that automatically allow inspection according to a common criterion. This is especially so if the sub-types follow bureaucratic distinctions which provide for rival programme cultures, which in turn seed the framework with different programme theories.

If the classification frames do pick up different programme ideas or mechanisms, then it is important that the initiatives therein be judged by appropriate standards and this can break the back of any uniform measurement apparatus. The greater the extent that different programme mechanisms are introduced into meta-analytic comparisons, the more we bring into play differences in intervention duration and development,

practitioner experience and training, subject self-selection and staying power, and so forth, which are themselves likely to influence the outcomes. In short, the categories of meta-analysis have the potential to hide, within and between themselves, very many significant capacities for generating personal and social change and we thus need to be terribly careful about making any causal imputations to any particular 'category'.

ii) Oversimplified outcomes

Such a warning looms even more vividly when we move from cause to effect. This brings us to programme outcomes and the second problem with numerical meta-analysis, which is concealed in that rather tight fisted term – the 'mean effect'. The crucial point to recall as we cast our eyes down the outputs of meta-analysis, such as column three in Table 1, is that the figures contained therein are means of means of means of means! It is useful to travel up the chain of aggregation to examine how exactly the effect calculations are performed for each sub-category of programme. Recall that PPMH programmes carry the broad aims of increasing 'psychological well-being' and tackling 'adjustment problems'. These objectives were put to the test within each evaluation by the standard method of performing before-and-after calculations on indicators of the said concepts.

Mental health interventions have a long pedigree and so the original evaluations were able to select indicators of change from a whole battery of 'attitude measures', 'psychological tests', 'self-reports', 'behavioural observations', 'academic performance records', 'peer approval ratings', 'problem solving vignettes', 'physiological assessments of stress' and so on. As well as varying in kind, the measurement apparatus for each intervention also had a potentially different time dimension. Thus 'post-test' measures will have had different proximity to the actual intervention and in some cases, but not others, were applied in the form of third and subsequent 'follow-ups'. Further, hidden diversity in outcome measures follows from the possibility that certain indicators (cynics may guess which!) may have been used but have gone unreported in the original studies, given limitations on journal space and pressures to report successful outcomes.

These then are the incredibly assorted raw materials through which meta-analysis traces programme effects. Now, it is normal to observe some variation in programme efficacy across the diverse aspects of personal change sought in an initiative (with it being easier to shift, say, 'self-reported attitudes' than 'anti-social behaviour' than 'academic achievement'). It is also normal to see programme effects change over time (with many studies showing that early education gains may soon dissipate but that interpersonal-skills gains are more robust – c.f. McGuire et al, 1997). It is also normal in multi-goal initiatives to find internal variation in success across the different programme objectives (with the school-based programmes having potentially more leverage on classroom-based measures about 'discipline referrals' and 'personal competence' than on indicators shaped more from home and neighbourhood, such as 'absenteeism' and 'drop-out rates'). In short, programmes always generate multiple outcomes and much is to be learnt about how they work by comparing their diverse impacts within and between programmes and over time.

There is little opportunity for such flexibility in meta-analysis, however, because any one study becomes precisely that, namely 'study x' of, for instance, the 15 school-based studies. Outcome measurement is a tourniquet of compression. It begins life by

observing how each individual changes on a particular variable and then brings these together as the mean effect for the programme subjects as a whole. Initiatives normally have multiple effects and so its various outcomes, variously measured, are also averaged as the 'pooled effect' for that particular intervention. This, in turn is melded together with the mean effect from 'study y' from the same sub-set, even though it may have used different indicators of change. The conflation process continues until we gather in all the studies within the sub-type, even though by then the aggregation process may have fetched in an even wider permutation of outcome measures. Only then do comparisons begin as we eyeball the mean, mean, mean effects from other sub-categories of programmes. Meta-analysis, in short, will always generate its two-decimal-place mean effects but since it squeezes out much ground level variation in the outcomes, it remains open to the charge of spurious precision.

iii) Concealed contexts

My third critique puts the 'like with like?' test to a further element of all social programmes, namely the subjects who, and situations which, are on the receiving end of the initiatives. No individual-level intervention works for everyone. No institution-level intervention works everywhere. The net effect of any particular programme is thus made up of the balance of successes and failures of individual subjects and locations. Thus any 'programme outcome' – single, pooled or mean – depends not merely upon 'the programme' but also on its subjects and its circumstances. These contextual variations are yet another feature that is squeezed out of the picture in the aggregation process of meta-analysis.

Some PPMH programmes, as we have seen, tackle long-standing issues of the day like educational achievement. Transforming children's progress in this respect has proved difficult, not for want of programmes but because the educational system as a whole is locked into a wider range of social and cultural inequalities. The gains made under a specific initiative are thus always limited by such matters as the class and racial composition of the programme subjects and, beyond that, by the presence or absence of further educational and job opportunities. The same programme may thus succeed or fail according to how advantageous is its school and community setting. At the aggregate level, this proposition generates a rival hypothesis about success in the 'league tables'. Unless we collect information about the prevailing contextual circumstances of the programme, it might be that the accumulation of 'soft targets' rather than successful programmes *per se* could produce a winning formula.

The same point also applies even to the seemingly 'targeted' initiatives. For instance, the 'milestone' programmes in Durlak and Wells' analysis make use of a 'modelling' mechanism. The idea is that passing on the accounts of children who have survived some trauma can be beneficial to those about to face the 'same' transition. The success of that hypothesis depends, of course, on how authentic and how evocative are the 'models'. And this may be expected to vary according to the background of the child facing the transition. Different children may feel that the predicaments faced in the models are not quite the same as their own, that the backgrounds of the previous 'victims' are not quite parallel to their own, that the programme accounts do not square with other advice they are experiencing, and so on. It is quite possible (indeed quite routine) to have a programme that works well for one class of subjects but works against another, and whose differential effects will be missed in the aggregation of outputs. This points to the importance of the careful targeting of programmes as a key

lesson to be grasped in accumulating knowledge about programme efficacy. But this is rather unlikely in meta-analysis, which works to the rival philosophy that the best programmes are the ones that work most widely.

Again, I must emphasise that the hypotheses expressed in the above two paragraphs are necessarily speculative. What they point to is the need for a careful look at subject and contextual differences in terms of who succeeds and who fails within any programme. And the point, of course, is that this is denied to a meta-analysis, which needs unequivocal and uniform base-line indicators for each case. At the extreme, we can still learn from a negative net effect since the application of an initiative to the wrong subjects and in the wrong circumstances can leave behind vital clues about what might be the right combination. No such subtlety is available in an approach that simply extracts the net effect from one study and combines it together with other programmes from the same stable to generate an average effect for that sub-class of programmes. Vital explanatory information is thus once again squeezed out automatically in the process of aggregation.

Let me now try to draw together the lessons of the three critiques of meta-analysis. What I have tried to do is cast doubt on the exactitude of the mean effect calculations. Arithmetically speaking, the method always 'works' in that it will produce mechanically a spread of net effects of the sort we see in Table 1. The all-important question, of course, is whether the league table of efficacy produced in this manner should act as an effective guide to future policy making. Should the PPMH meta-analyst advise, for instance, on cutting parenting programmes and increasing affective education and interpersonal problem solving schemes, especially for the under-sevens? My answer would be an indubitable 'no'. The net effects we see in column three do not follow passively from application of that sub-class of programmes but are generated by unmonitored *differences* in the underlying programme mechanisms, in the contexts in which the programme is applied and in the measures which are used to tap the outcome data.

Let me make it abundantly clear at this point that my case is not just against this particular example. There are many less sophisticated expositions of meta-analysis. And there are more sophisticated models which attempt technical solutions to some of the problems presented here, by way of narrowing the bandwidth of studies, testing for heterogeneity of outcomes, and so on (Cook et al, 1992). So, note well, what I am taking issue with is the basic 'logic' of meta-evaluation. The central objective for evidence-policy linkage on this strategy is to go in search of the stubborn empirical generalisation. To put this rather more precisely, the goal is the replication of positive results for a class of programmes in a range of different situations.

This is rather like the misidentification of 'universal laws' and 'empirical generalisations' in natural science (Kaplan, 1964). The discovery of the former does not imply or require the existence of the latter. To arrive at a universal law we do not require that X is always followed by Y. Rather, the expectation is that 'Y' always follows 'X' *in certain prescribed situations* and that we need a programme of theory and research to delimit those situations. Thus gunpowder does not always explode if a match is applied – that consequence depends (on other things) on it being dry; the pressure of a gas does not always increase linearly with temperature – unless there is a fixed mass; falling bodies like cannon balls and feathers don't follow the laws of

motion – unless we take into account friction, air resistance and so on. The meta-analytic search for 'heterogeneous replication' (Shadish et al 1991) is rather akin to the search for the brute empirical generalisation in that it seeks programme success without sufficient knowledge of the conditions of success.

The message, I trust, is clear: 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 1. The contexts, mechanisms and outcomes that constitute each set of programmes are so diverse that it is improbable that like gets compared with like. So whilst it might seem objective and prudent to make policy by numbers, the results may be quite arbitrary. This brings me to a final remark on Durlak and Wells and to some good news (and some bad). The PPMH meta-analysis is not in fact used to promote a case for or against any particular sub-set of programmes. So, with great sense and caution, the authors avoid presenting their conclusions as a case for thumbs up for 'interpersonal problem solving for the very young', thumbs down for 'parenting programmes', and so on. Indeed they view the results in Table 1 as 'across the board' success and perceive that the figures support the extension of programmes and further research in PPMH as a whole.

Their reasoning is not so clear when it comes to step two in the argument. This research was done against the backdrop of the US Institute of Medicine's 1994 decision to exclude mental health promotion from its official definitions of preventative programmes, with a consequent decline in their professional status. Durlak and Wells protest that their finding (ESs of 0.24 to 0.93) compare favourably with the effect sizes reported routinely in psychological, educational and behavioural treatments (they report that one overview of 156 such meta-analyses came up with a mean of means of means of means of means of 0.47). And what is more, 'the majority of mean effects for many successful medical treatments such as by-pass surgery for coronary heart disease, chemotherapy to treat certain cancers…also fall below 0.5.' If one compares for just a second the nature, tractability and seriousness of the assorted problems and the colossal differences in programme ideas, subjects, circumstances and outcome measures involved in this little lot, one concludes that we are being persuaded, after all, that chalk can be compared to cheese.

1.3. Narrative reviews

We move to the second broad perspective of EBP, which comes in a variety of shapes and sizes and goes by an assortment of names. In an attempt at a catch-all expression, I shall refer to them as 'narrative reviews'. Again, I will seize upon a couple of examples, but once again note that my real objective is to capture and criticise the underlying 'logic' of the strategy.

The overall aim of the narrative approach is, in fact, not a million miles from the numerical strategy in that a family of programmes is examined in the hope of finding those particular approaches that are most successful. Broadly speaking, one can say that there is a slightly more pronounced explanatory agenda within the strategy, with the reviews often having something to say on why the favoured interventions were in fact successful. In methodological terms, however, the significant difference between the two core approaches lies in respect of the following (related) questions:

- What information should be extracted from the original studies?
- How should the comparison between different types of initiatives be achieved?

On the first issue, the contrast is often drawn between the numerical approach, which is said only to have eyes for outcomes, whereas narrative reviews are pledged to preserve a 'ground-level view' of what happened with each programme. The meaning of such a commitment is not always clear and this accounts for some variations in the way it is put into practice. One detectable strategy at work is the 'tell it like it is' philosophy of attempting to preserve the 'integrity' or 'wholeness' of the original studies. Another way of interpreting the desideratum is that the review should extract enough of the 'process' of each programme covered so that its 'outcome' is rendered intelligible to the reader of the review.

The interesting thing about these goals is that they are very much the instincts of 'phenomenological' or 'case study' or even 'constructivist' approaches to applied social research. The great difficulty facing narrative review is to preserve these ambitions over and over again as the exercise trawls through scores or indeed hundreds of interventions. The same problem, incidentally, haunts comparative-historical inquiry as it attempts to preserve 'small n' historical detail in the face of 'large N' international comparisons (Ragin, 1987).

The result is that data extraction in narrative reviews is always something of a compromise. The difficulty can be highlighted by pausing briefly on a 'worse case scenario' from the genre. Walker's *Injury prevention for young children: a research guide* (1996) provides a particularly problematic example for, despite its sub-title, it is little more than an annotated bibliography. The format is indeed an index of inquiries, taking the shape of what look suspiciously like the paraphrased abstracts of 370 studies. The working assumption, one supposes, is that abstracts are supposed to provide the 'essence' of the programme and the 'gist' of the research findings, and these are the vital raw ingredients of review. The summaries are then divided into nine sub-sections dealing with different forms of injury (asphyxiation to vehicle injury). Each sub-section begins with brief, paragraph-length introductions which go no further than describing the broad aims of the programmes therein.

Now, 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. Note that I am not slaying a straw man in the name of a whole strategy here, merely pointing out an acute form of underlying tension between the goals of 'revealing the essence of each case study' and 'effecting a comparison of all case studies'.

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, from the same field of childhood accident prevention, is to be found in the work of Towner et al (1996). Appendix H in that study supplies an 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
7. Outcome measures employed
8. Summary of important results
9. Rating of the 'quality of evidence'

This represents a move from trying to capture the essence of the original studies via an 'abstract/summary' to attempting to locate their key aspects on a 'data matrix'. The significant point, of course, is that, being a narrative approach, the cell entries in these tabulations are composed mainly of text. This text can be as full or as brief as research time and inclination allows. In its raw form, the information on any single study can thus easily run up to two or three pages. This may include the extraction of a key quotation from the original authors, plus some simple tick-box information on, say, the age range of the target group. The record of findings can range from effect sizes to the original authors' thoughts on the policy implications of their study. And furthermore, the review may also stretch to include the reactions of the reviewer (e.g. 'very good evaluation – no reservations about the judgements made'). In short, the raw data of narrative review can take the form of quite a mélange of different types of information.

Normally such reviews will also provide appendices with the entire matrix on view in a condensed form, so one can literally 'read across' from case to case, comparing them directly on any of the chosen features. A tiny extract from Towner's summary of 56 road safety initiatives is reproduced in Table 2 in order to provide a glimpse of the massive information matrix.

Table 2: The 'summary matrix' in narrative review

Road Safety Education – Experimental Programmes

Author & date of publication	Country of study	Injury target group (age in years)	Study type	Type of Intervention	Healthy Alliance	Outcome Measure(s)	Outcome(s)
Yeaton & Bailey 1978	USA	5-9	Before and after study	One to one real life demonstrations to teach 6 street crossing skills	Crossing patrol, schools	Observed behaviour	Skills improved from 48% - 97% and 21% - 96 %. Maintained at one year
Nishioka et.al. 1991	Japan	4-6	Before and after study with 2 comparison groups	Group training aimed at dashing out behaviour	No details	Reported Behaviour	Improvements in behaviour dependent on level of training. 40% of children shown to be unsafe AFTER training

Plus 7 more experimental programmes

Table 2: The 'summary matrix' in narrative review (cont.)

Road Safety Education – Operational Programmes

Author & date of publication	Country of study	Injury target group (age in years)	Study type	Type of Intervention	Healthy Alliance	Outcome Measure(s)	Outcome(s)
Schioldborg 1976	Norway	Pre school	Before and after study with control group	Traffic club	Parents	Injury rates, observed behaviour	No effect on traffic behaviour. Reported 20% reduction in casualty rates and 40% in Oslo
Anataki et. al. 1986	UK	5	Before and after study with control group	Road safety education using Tufty materials	Schools, road safety officers	Knowledge	All children improved on test score over the 6 month period, however children exposed to the Tufty measures performed no better than non intervention group

Plus 4 more experimental programmes

Road Safety Education – Area-Wide Engineering Programmes
10 programmes
Road Safety Education – Cycle helmet studies
10 programmes
Road Safety Education – Enforcement Legislation
3 programmes
Road Safety Education – Child Restraint Loan Schemes
9 programmes
Road Safety Education – Seat Belt Use
9 programmes

Source: Towner et al (1996)

There is little doubt that such a procedure provides an incomparable overview of 'what is going on in' in a family of evaluation studies. There is one clear advantage over the numerical approach in that it admits a rather more sophisticated understanding of how programmes work. Meta-analysis is somewhat trapped by its 'successionist' understanding of causality. The working assumption is that it is the programme 'x' that causes the outcome 'y', and the task is to see which sub-type of initiative (x_1, x_2, x_3 etc.) has the most significant impact.

In narrative review, the programme is not seen as a disembodied feature with its own causal powers. Rather programmes are successful when (imagine reading across the data matrix/extraction form) the right intervention type, with clear and pertinent objectives, comes into contact with an appropriate target group, and is administered and researched by an effective stakeholder alliance, working to common goals, in a conducive setting, and so on. This logic utilises a 'configurational' approach to causality, in which outcomes are considered to follow from the alignment, within a case, of a specific combination of attributes. Such an approach to causation is common amongst narrative inquiry of all kinds and features strongly, once again, in the methodology of comparative historical research (Ragin, 1987, p125). There is no need

to labour the point here but one can see a clear parallel, for instance, in the logic used for explaining programme efficacy and that applied in charting the success of a social movement. Explanations of why England experienced an early industrial revolution turn on identifying the *combination* of crucial conditions such as 'technological innovation', 'weak aristocracy', 'commercialised agriculture', 'displaced peasantry', 'exploitable empire' and so on (Moore, 1966, Ch. 1).

However, the capacity of the narrative approach for teasing out such a 'holistic' understanding of programme success by no means completes the trick of accomplishing a review. The next and crucial question concerns how comparisons are brought to bear upon such configurational data. How does one draw transferable lessons when one set of attributes might have been responsible for success in programme 'A' and a different combination might account for achievements of programme 'B'? My own impression is that the 'logic of comparison' has not been fully articulated in the narrative tradition, and so there is considerable variation in practice on this matter. Simplifying, I think there are three discernible 'analytic strategies' in use, as discussed in the following paragraphs. All three are often found in combination, though there is a good case for arguing that approach 'b' is the most characteristic:

a) Let the reviews speak for themselves

No one should forget that systematic review is a long and laborious business, which itself occupies the full research cycle. It all begins with difficult decisions on the scope of the programmes to be brought under review. Then the sample of designated cases is located by all manner of means from key word searches to word-of-mouth traces. Then comes the wrestling match to find a common template on which to code the multifarious output, be it learned, grey, or popular. Then comes the matter of data extraction itself, with its mixture of mechanical coding and guesswork, as the researchers struggle to translate the prose styles of the world into a common framework. Then, if the business is conducted properly, comes the reliability check, with its occasionally shocking revelations about how research team members can read the same passage to opposite effects. Then, finally, comes the quart-into-pint-pot task of presenting the mass of data as an intelligible set of summary matrices and tables.

Little wonder then, that there is a tendency in some systematic reviews to draw breath and regard the completion of this sequence as the 'job done'. On this view, EBP is seen largely as a process of condensation, a reproduction in miniature of the various incarnations of a policy idea. The evidence has been culled painstakingly to enable a feedback loop into fresh policy making, but the latter is a task to be completed by others.

However sympathetic one might be to the toils of information science, this 'underlabourer' notion of EBP must be given very short shrift. Evidence, new or old, numerical or narrative, diffuse or condensed, never speaks for itself. The analysis and usage of data is a sense-making exercise and not a mechanical one. Social interventions are descriptively inexhaustible. Of the infinite number of events, actions and thoughts that make up a programme, rather few get set down in research reports and rather fewer get recorded in research reviews. It is highly unlikely, for instance, that the evidence base will contain data on the astrological signs of project participants or the square footage of the project headquarters. It is relatively unlikely that the

political allegiance of the programme architects or size of the intervention funding will be systematically reviewed. In other words, certain explanations for programme success or failure are favoured or disqualified automatically according to the particular information selected for extraction.

Since it cannot help but contain the seeds of explanation, systematic review has to acknowledge and foster an interpretative agenda. If, contrariwise, the evidence base is somehow regarded as 'raw data' and the interpretative process is left incomplete and unhinged, this will simply reproduce the existing division of labour in which policy-makers do the thinking and other stakeholders bear the consequences. EBP has to be somewhat bolder than this or it will provide a mere decorative backwash to policy making, a modern day version of the old ivory tower casuistry, 'here are some assorted lessons of world history, Prime Minister – now go rule the country'.

b) Pick out exemplary programmes for an 'honourable mention'

However modestly, most narrative reviews do in fact strive to make the extra step into policy recommendations and we turn next to the main method of drawing out the implications of the evidence base. This is the narrative version of the 'best buy' approach and it takes the form of identifying 'exemplary programmes'. In practice this often involves using the 'recommendations for action' or 'executive summary' section of the review to re-emphasise the winning qualities of the chosen cases. The basis for such a selection has been covered already; successful programmes follow from the combination and compatibility of a range of attributes captured in the database. A *de facto* comparison is involved in such manoeuvres in that less successful programmes obviously do not posses such configurations. These are entirely legitimate ambitions for a narrative review. But the question remains – how is the policy community able to cash in on the merits of exemplary programmes? What are the lessons for future programming and practice? What theory of 'generalisation' underlies this approach?

The answer to the latter question is captured in the phrase 'proximal similarity' (Shadish et al, 1991). Basically, the idea is to learn from review by following the successful programmes. What makes them successful is the juxtaposition of many attributes and so the goal for future programme design is to imitate the programme as a whole or at least try to gather in as many similarities as possible. In the varied world of narrative reviews this, I believe, is the closest we get to a dominant guiding principle. (Note that it is almost the opposite to the idea of heterogeneous replication – it is a sort of 'homogeneous replication'.)

There are several problems with using the proximal similarity principle as the basis for drawing transferable policy lessons. The first is that it is impossible to put into practice (quite a drawback!). As soon as one shifts a programme to another venue, there are inevitably differences in infrastructure, institutions, practitioners and subjects. Implementation details (the stuff of narrative reviews) are thereby subtly transformed, making it almost impossible to plan for or to create exact similarities. And, if one creates only 'partial similarities' in new versions of the programmes, one doesn't know whether the vital causal configurations on which success depends are broken. Such a dismal sequence of events, alas, is very well known in the evaluation community. Successful 'demonstration projects' have often proved the devil to replicate, precisely because programme outcomes are the sum of their assorted, wayward, belligerent parts (Tilley, 1996).

A second problem with a configurational understanding of programme efficacy stems from its own subtlety. There is no assumption that the making of successful programmes follows only the single 'recipe'. Interventions may get to the same objectives by different means. Yet again, the parallel with our example from comparative historical method rears its head. Having identified the configuration of conditions producing the first industrial revolution does not mean, of course, that the second and subsequent industrial nations follow the same pathway (Tilly, 1975, Ch. 2). What counts in history and public policy is having the right idea for the apposite time and place.

Wise as this counsel is, it is a terribly difficult prospect to identify the potential synergy of an intervention and its circumstances on the basis of the materials identified in a narrative review. As we have seen, programme descriptions in narrative reviews are highly selective. Programme descriptions are also 'ontologically flat' (Sayer, 2000). This is quite literally the case as one follows the various elements of each intervention across the data matrix. What Sayer and other realists actually have in mind with the use of the phrase, however, is that the various 'properties' extracted in a review are just that, a standard set of observable features of programmes. Now, change actually takes place in a 'ontologically deep' social world in which the dormant capacities of individuals are awakened by new ideas, from which new ways of acting emerge.

However painstaking a research review, such 'causal powers' and 'emergent properties' of interventions are not the stuff of the data extraction form. In the standard mode of presenting narrative reviews (recall Table 2), one gets a picture of the particulars of each intervention without getting a feel for the way that the programmes actually work. For instance, several reviews of road safety education, including the aforementioned work by Towner, have indicated that children can better recall rules for crossing busy roads if they are learned using roadside 'mock-ups' rather than through diagrams and pictures. Why this might be the case is presumably to do with the benefits of learning practical skills 'in situ' rather than 'on paper'. In other words, we are not persuaded that one particular programme configuration works better than does another because of the mere juxtaposition of its constituent properties. Rather what convinces is our ability to draw upon an implicit, much used and widely useful theory.

Note that my criticism here is not to do with missing evidence or wrong headed interpretation. Rather I am arguing that the extraction of exemplary cases in narrative overview is not simply a case of 'reviewing the evidence' but depends, to a considerable degree, on the tacit testing of submerged theories. Why this state of affairs is so little acknowledged is something of a mystery. Perhaps it is the old empiricist fear of the 'speculative' and the 'supra-sensible'. Be that as it may, such an omission sits very oddly with the design of interventions which is all about the quest for improvement in 'programme theory'. No doubt the quest for verisimilitude in road safety training came from the bright idea that learning from actual practice (if not through mistakes!) was the way ahead. The moral of the tale, however, is clear. The review process needs to acknowledge this vital characteristic of all programmes and thus appreciate that the evidence base is theory-laden.

c) Meld together some common denominators of programme success

As indicated above, the broad church of narrative review permits several means to forward policy ends. So as well as the piling high of descriptive data and the extraction of exemplars, there is also a range of hybrid strategies which owe something to the method of meta-analysis. I have space to mention just one variant on this theme here. The more extensive the review, the more likely is it that a range of good practice has cropped up across the various sub-families of initiatives contained therein. Instead of handling these 'one at a time' for flattery by means of imitation, there are also attempts to extract some 'collective lessons' in narrative form.

Whilst the cell entries take the form of text rather than numbers (Table 2 again), the basic database of a narrative review is still in the form of a matrix. This allows one to pose questions about which manifestations of the various programme properties tend to be associated with programme success. Given the narrative form of the data, this cannot be put as a multivariate question about the respective correlates of programme success. But what is possible is to pick upon each programme property in turn and try to discover what the exemplars share in common in respect of that particular 'attribute'. Such an exercise tends to arrive in the 'summary and conclusions' section as a series of recommendations about good practice in programme design and implementation. This is, of course, very much in keeping with the emphasis on intervention processes that pervades the qualitative evaluation tradition.

Thus, of programme development, the message of such a review will often be something like, 'inter-agency collaboration is essential to develop different elements of a local campaign' (Towner, 1996). Of the programme targets, the shared quality of successful interventions might be 'initiatives must be sensitive to individual differences' (ibid). In other words, the theory of generalisation in this narrative sub-strategy has shifted to what might be called the search for the 'common denominators of implementation success'. This is a more confusing and confused project because it cross-cuts the 'heterogeneous replication' and the 'proximal similarity' principles. I find it very hard to fault the collected wisdom of recommendations such as the above. But, because they combine the disembodied, variable-based method of meta-analysis with a narrative emphasis on process, they often end up looking rather trite. Often we are presented with bureaucratic truisms, with certain of the recommendations for programme effectiveness applying, no doubt, to organising the bus trip to Southend.

Conclusion

What conclusions should be drawn from this brief review of the main methods of systematic review? Let me make it clear what I have attempted to say and tried not to say. I have disavowed what I think is the usual account, a preference for the numerical or the narrative. Neither, though I might be accused of doing so, have I declared a plague on both houses. The key point, I think, is this. There are different ways of explaining why a particular programme has been a success or failure. Any particular evaluation will thus capture only a partial account of the efficacy of an intervention. When it comes to the collective evaluation of whole families of programmes, the lessons learned become even more selective.

Accordingly, any particular model of EBP will be highly truncated in its explanation of what has worked and what has not. In this respect, I hope to have shown that meta-

analysis ends up with de-contextualised lessons and that narrative review concludes with over-contextualised recommendations. My main ambition has thus been to demonstrate that there is plenty of room left right there in the middle for an approach that is sensitive to the local conditions of programme efficacy but then renders such observations into transferable lessons. And what might such an approach look like? All the basic informational mechanics have to be there, of course, as does the eye for outcome variation of meta-analysis, as does the sensitivity to process variation of narrative review.

Putting all of these together is a rather tough ask. The intended synthesis will not follow from some mechanical blending of existing methods (we caught a glimpse of such an uneasy compromise in my final example). What is required, above all, is a clear logic of how research review is to underpin policy prognostication. The attentive reader will have already read my thoughts on the basic ingredients of that logic. I have mentioned how systematic review often squeezes out attention to programme 'mechanisms', 'contexts' and 'outcome patterns'. I have already mentioned how 'middle-range theories' lurk tacitly in the selection of best buys. These concepts are the staples of realist explanation (Pawson and Tilley, 1997), which leads me to suppose there might be promise in a method of 'realist synthesis'.

Bibliography

Campbell, D; Russo, M J (editor) (1999) *Social experimentation* Sage: Thousand Oaks, CA. 405pp

Cook, T D et al (1992) *Meta-analysis for explanation: a casebook* Russell Sage Foundation: New York. 378pp

Davies, P (2000) The relevance of systematic review to educational policy and practice *Oxford Review of Education* Sep/Dec 26(3/4) pp365-78

Durlak, J A; Wells, A M (1997) Primary prevention mental health programs for children and adolescents: a meta-analytic review *American Journal of Community Psychology* Apr 25(2) pp115-52

Kaplan, A (1964) *The conduct of inquiry: methodology for behavioral science* Chandler Publishing Co: San Francisco, CA. 428pp

McGuire, J B; Stein, A; Rosenburg, W (1997) Evidence-based medicine and child mental health services: a broad approach to evaluation is needed *Children and Society* 11(2) pp89-96

Moore, B (1966) *Social origins of dictatorship and democracy: lord and peasant in the making of the modern world* Beacon Press: Boston, Mass. 559pp

Pawson, R; Tilley, N (1997) *Realistic evaluation* Sage: London. 235pp

Pease, K (1998) *Repeat victimisation: taking stock* Home Office Police Research Group: London. 47pp (Crime Detection and Prevention Paper 90) Available via: <www.homeoffice.gov.uk/rds/policerspubls1.html>

Ragin, C (1987) *The comparative method: moving beyond qualitative and quantitative strategies* University of California Press: Berkeley, CA. 185pp

Sayer, A (2000) *Realism and social science* Sage: London. 211pp

Shadish, W R; Cook, T D; Leviton, L C (1991) *Foundations of program evaluation: theories of practice* Sage: Newbury Park, CA. 529pp

Tilley, N (1996) Demonstration, exemplification, duplication and replication in evaluation research *Evaluation* 2(1) pp35-50

Tilly, C (1984) *Big structures, large processes, huge comparisons* Russell Sage Foundation: New York. 176pp

Towner, E; Dowswell, T; Jarvis, S (1996) *Reducing childhood accidents - the effectiveness of health promotion interventions: a literature review* Health Education Authority: London. 85pp

Walker, B (1996) *Injury prevention for young children* Greenwood Press: Westport, CT. 182pp

Weissberg, R; Bell, D (1997) A meta-analytic review of primary prevention programs for children and adolescents: contributions and caveats *American Journal of Community Psychology* Apr 25(2) pp207-14