ESRC Research Methods Programme
Working Paper No 1

Assessing the quality of evidence in evidence-based policy: why, how and when?

Ray Pawson

University of Leeds

Not to be cited without permission

Contact: Ray Pawson
Phone 01132 334 419
Email r.d.pawson@leeds.ac.uk

Abstract

Interest in the issue of 'research quality' is at an all time high. Undoubtedly, one of the key spurs to the quest for higher standards in social research is the evidence-based policy movement. The chosen instrument for figuring out best-possible, future interventions in a particular policy domain is the systematic review of all first-rate, bygone evidence from previous studies in that realm. In trying to piece together the evidence that should carry weight in policy formation, a key step in the logic is to provide an 'inclusion criterion' as a means identifying those existing studies upon which most reliance should be placed. This paper examines some recent yardsticks used to sort the evidential sheep from the research goats by questioning why, how and when such research standards should be brought to bear. It concludes that the drive to cast standards as formal checklists of quality indicators is premature, and that appraising quality is not and cannot be a technical preliminary to research synthesis. Open and critical debate on the interpretation of research findings remains the surest way to establish and maintain investigatory standards.

Key words: research quality, research standards, hierarchy of evidence, evidence-based policy, systematic review, research synthesis.

UNEDITED FIRST DRAFT - COMMENTS WELCOME

r.d.pawson@leeds.ac.uk

Introduction: a summary of the argument

There are four parts to the paper. The notion of 'research standards' is dissected by considering, in turn: i) why, ii) how and, iii) when they become established. Close inspection of the 'why' and 'how' questions reveals a weakness in the quality standards envisioned in the current models of systematic review. In the process of this critique an alternative model of research quality is developed based on the 'when' question and the contribution that an inquiry makes to explanatory synthesis. The paper concludes: iv) with a demonstration of the new model in assessing the merits of the evidence on the efficacy of Megan's Law.

I. On the first issue (why?), the paper concurs in the utmost with the quest for high-quality research and thus has no quibble with the implication that there are forms of knowledge that can be privileged by dint of methodological rigour. The critique mounted here is thus not a piece of (self-) defeatist postmodernism maintaining that research standards lie in the eye of the beholder. There *are* mechanisms for sorting science from non-science, for distinguishing between social science and common sense, for differentiating rigorous from slipshod inquiry. But such criteria are *not* to be mistaken for technical competence, for the strategies and techniques of social and evaluative research are manifold and antagonistic, and ever growing in their complexity and diversity. To ring-fence and kite-mark only a portion of this technical capacity would be to blunt the scientific imagination.

Instead of looking for high quality in the *practice* of research, it follows that establishing standards is a matter for the *process* of inquiry. What counts is the capacity of a piece of research to marshal evidence in order to test and refine a theory under investigation. Research progresses only insofar as each investigation contributes to the adjudication of a better set of explanatory propositions. Evidence cumulates in the process of debate and counter-debate on the veracity of those explanations. The key lesson of post-empiricist philosophy of science is that methodological standards are long-term, emergent processes of inquiry. They are the medium and outcome of investigation. It is this view of research quality that is championed in this paper. It is this view that has to be incorporated to sustain the excellence of the evidence in evidence-based policy.

II. On the second issue (how?), the paper takes the form of a critical examination of some recent attempts from the systematic review community to capture standards in the form of prescribed schedules of quality indicators. It begins by identifying the template for research quality established in evidence-based medicine. Here, there is a clear, consensual and easily-recognised 'gold-standard' for research synonymous with the usage of randomised controlled trials. Widespread agreement that such a design is indeed the touchstone of research integrity allows the application of a 'quality filter' *directly* and *early* in the review. Inferior (non-RCT) studies are discarded (often in large numbers) leaving the review to pool together the findings of authoritative investigations. I argue that the (apparent) success of such a strategy depends upon a disconcertingly narrow interpretation of what it is necessary to know in order to declare that an intervention is fit for use in future policy-making and practice. Reviews are whittled down to the single explanatory quest (is the treatment effective?), leading to a single quality filter (is the trial effective?) producing a single explanatory outcome (a measure of net efficacy). Such a logic, and such a quality check, becomes quite ineffectual if the review question is 'why does a programme work?' and confronting and assessing the evidence in respect of this explanatory quest is crucial in the development of evidence-based policy.

The paper then inspects recent attempts to broaden the notion of quality standards as appropriate to a wider vision of evidence-based policy. Social programmes are more complex than medical interventions and the apposite evaluative question is more likely to turn on 'why interventions have differential effects for different subjects and circumstances?'. Quite properly, it is assumed that the policy response to these various issues needs to be buttressed by evidence of all shapes and sizes and colours and countenances. Quite understandably, fresh standards are being developed to cover 'qualitative research', 'evaluation

research', 'action research', 'emancipatory research' and so on. The expectation is that they will be able to imitate the quality appraisal function in evidence based medicine. The 'new standards', however, are complex, abstract, fragmented and, in some cases, contradictory. Accordingly, the paper agues they can have no role as an inclusion hurdle in systematic review. They are far too extensive in range, too ambiguous in tone, and too subtle in application for them to act as preliminary quality filters prior to research synthesis. They are repositories of generalised research wisdom rather than sharp decision points capable of expediting uncluttered reviews.

III. On the third issue (when?), the paper locates its positive suggestions for a standards regime appropriate to the complexity of the social policy evidence-base. The starting point is the principle raised in the first section, namely that the veracity of investigation is a matter of process rather than practice. Science is not a collection of durable empirical uniformities. Findings do not speak for themselves. What is always at issue is their interpretation. Accordingly, what qualifies a study as being of 'good quality' is not its technical competence as such, but whether its technical infrastructure will bear the weight of the inferences to which it lays claim. The acid test of research quality is whether a study provides good explanation and this involves examination of how it jockeys for position amongst competing explanations. Inquiries are judged to be competent only when they secure a place in a developing network of explanations. Research quality is confirmed only when synthesis is achieved.

Interestingly, this same proposition features strongly in claims made about the veracity of a wide variety of non-positivist social research. Strategies such as 'pattern explanation', 'analytic induction', 'middle-range theory building', 'triangulation' and so on, all stress that the process in which qualitative evidence becomes warranted is its confederation within a larger system of explanation. Curiously, these notions, which are indubitably about validity, have been overlooked by the 'new standards' compilers, who have preferred to look for broad-based technical competence. The implication for evidence-based policy, however, is clear. Systematic review aspires to a high-speed condensation of the normal sequence of scientific discovery and should follow the same logic. Investigatory standards cannot be judged in one fell swoop; they are decided in the full sweep of hypothesis generation, analysis, critique and theory-building. Research quality is confirmed only when synthesis is achieved.

IV. The final section of the paper provides an illustration of the standards-as-synthesis thesis. The demonstration is made in terms of three studies that contributed to a systematic review of the effectiveness of Megan's Law, conducted previously by the author (Pawson 2002). One is a trial, one is a qualitative study, and one is a prospective simulation based on available criminal justice statistics. It is highly likely that the first two studies would have been judged as flawed or even scratched under an orthodox quality filter. The former is only a 'matched' rather than a 'randomised' trial, the second shows some blatant favouritism to its research subjects. And, as far as I am aware, there is no standards framework available at all to accredit the third type of study. Despite these dubious technical qualifications, the inquiries combine to provide a plausible and technically justifiable account of why the public disclosure of the identities of sex offenders is limited in its ability to reduce re-offence. Each study strengthens the inferences made by the other. What is warranted in the act of synthesis is not the study-as-whole but the veracity of a segment of its explanatory propositions. This explanatory ensemble is further strengthened with the triangulation of other studies into the developing explanation. Research quality is confirmed only when synthesis is achieved.

I. Why should research quality be assessed?

The very idea of evidence-based policy rests on pair of remarkably brave claims. The first, which is *not* the subject matter of this paper, is the proposition that evidence can be heard amidst the political clamour of modern policy-making. The second, which is my topic, is that evidence should have a privileged voice in policy formation because there are objective methods available to judge and justify the quality of the advice provided by social research. The answer to my 'why?' question, and thus starting point for my

assessment of quality assessment in evidence-based policy, lies with this assertion that there are demonstrable truths and untruths about the effectiveness of policy initiatives and, crucially, that there are objective means of detecting the difference.

Sorting truths from falsehoods and the methods by which we tell the difference is the prerogative of the philosophy of science. Much ink has been spilled on the grand 'demarcation debate' in which the criterion is sought for distinguishing between the authentic and the bogus in terms of scientific practice, and I begin by collecting some starting pointers from that literature. Whilst it is hardly obsessed with the lofty ambition of qualifying for the inner sanctum of 'science', the very idea of evidence-based policy rests on the matter of differentiating its efforts from 'common sense', 'intuition', 'experience', 'value judgement' and so on. Given these objectives, it is clear that are many miniature 'demarcation debates' to be confronted by way of sifting the good and the bad in the world of evaluative research. This raises the issue of whether the standards movement in policy research has absorbed the lessons of the grand demarcation debate. So first let us see how the philosophy of science has come to identify its subject matter, and I commence with three tales of caution from the attempts to spot the difference between science and non-science.

i) Do not mistake science for technical proficiency.

Science begets scores of disciplines (from astronomy to zymurgy), each with hundreds of procedures and techniques and instruments. Given the diversity of such tools it is unlikely that scientific status and the standards thereof should be associated with any particular procedures or techniques or instruments (Harré, 1972). The possibility of the direct manipulation of its subject matter varies widely across the astronomical, biological, chemical and human sciences and yet they all claim 'scientific' status. As we cross the disciplines, data collection varies from count taking to complex instrumentation to manufactured conversation and yet all are claimed as sources of 'measurement'. In the course of analysis mathematicians perform thought experiments, physicists build laboratory systems, and medical trialists randomise patients, all done in the name of 'experimentation'.

The general upshot of all this is that we should *not* seek to identify science via its use of particular instruments and specific technical manoeuvres. The fact that physicists and the hard sciences in general find precious little need for the random allocation of cases to experimental and control conditions does not suddenly relagate such comparisons from the fold. Conversely, the fact that evaluators lay great store by such comparisons does not automatically render their work scientific – if, for instance, they are used make conclusions about the external validity of the observed outcomes. What matters in all cases is the nature of the *inferences* drawn from the technical apparatus. Scientific knowledge resides in its propositions and what qualifies a proposition as 'scientific' is the not the source of empirical support it receives, but the whether its claims are consonant with that empirical support. The crucial message for quality regimes here is that we need to make absolutely sure that it is deductions that are checked for quality rather than data.

ii) Do not mistake science for the collection of hard facts. Science involves much more than the making of observations, the gathering of facts, the amassing of empirical generalisations. It isn't good data as such that makes science honest. It is the ceaseless scientific interrogation of its significance that makes data honest. It is now commonly accepted that we have no direct observational access to reality and that all observation is 'theory-laden'. The measurement apparatus of natural science is always complex and relies on instrumentation developed and perfected over many years. When your doctor suspects you have the 'flu and pops a thermometer in your mouth, she is not observing temperature directly. She is utilising a theory confirmed a century previously about the linear expansion of mercury. A theory about your illness is tested against a different, well-established theory. Several other theory-saturated instruments might well be consulted before the diagnosis is confirmed.

Drawing out from the microcosm to the larger picture, Lakatos gives a more compelling account of how science interrogates theory with evidence as follows:

> The clash is not 'between theories and facts' but between two high level theories: between an interpretative theory to provide the facts and an explanatory theory to explain them: and the interpretative theory may be on quite as high a level as the explanatory one. Accordingly, it is not that we propose a theory and Nature may shout NO; rather we propose an image of theories and Nature may shout INCONSISTENT. (Lakatos, 1970: 129)

Accordingly, the process of testing and refining theories with the evidence does not yield to simple uncontested conclusions. Patently, judgement on the success of such a process depends not only on the quality of the supporting evidence but also on the quality of the emerging theory. Often a range of theories is 'consonant with the data'. In such circumstances, there are many alternative reasons why we might light on a preferred explanation, over and above its 'fit with the evidence'. In this respect, Newton-Smith (1981) has provided a taxonomy of eight characteristics of good theory as follows: 'observational nesting'; 'explanatory fertility'; 'track record'; 'avoidance of the ad hoc'; 'internal consistency'; 'compatibility with well-grounded metaphysical beliefs'; 'simplicity'. Some of these ideas are, perhaps, self-evident and some of them may be discernible from their names. However, my purpose here is not to get involved in their exposition, but simply to insinuate another key doubt – how many of these can be handled in terms of research quality checklists?

iii) Do not mistake science for procedural uniformity. Science is not all control and calculation, checking and double-checking. It also proceeds though insight and imagination, speculative hunches and bold conjectures. Polanyi's (1966) felicitous phrase, 'we can know more than we can tell' was responsible for a foundational rethink of the philosophy of science. His point is that any particular scientific inquiry has thousands of decision points and that their resolution often relies on the experience, judgement and tacit wisdom of the researcher. This is especially so when it comes to the matter of hypothesis generation and inference making.

It follows that if one tries to create a set of standards for science in the name of procedural uniformity, and with the objective of following them faithfully, then one would actually block its progress. It would be a veritable hindrance if it were imagined that there was an existing and transparent research protocol to guide all eventualities. Progress depends, sometimes, on inspired guesswork and, occasionally, on going against prevailing wisdom. Whilst such breakthroughs always need to be consolidated with further rounds of inquiries and layers of evidence, the point remains that the path to progress is uncharted. There is a further and rather sober lesson in these moments of inspiration, which I leave to Hammersley (2002): 'if it [tacit knowledge] plays a key role in science, we can be fairly sure that it would play a key role in reviewing scientific evidence as well.'

Let me now draw together the main consequence of these three injunctions. Science is more than excellence of execution, more than durability of data. Neither is it simply a product of following a pre-determined investigatory pathway or its course would be mapped out already. The stark-staring implication for this study is that there will never be a simple litmus-test for science; there are no instantaneous warrants for declaring certain procedures as valid science. Important as they are, it is crucial not to regard particular activities such as experimentation and measurement as the touchstones, for they are embedded in a wider process of interpretation, inspection, correction and verification. Science is better conceived as a complex, self-auditing regime, a process nicely summarised by Williams (2000) as follows, 'science is a heterogeneous system of methodological checks and balances involving [empirical] testing, theory choice, and logical and mathematical reasoning'.

Accordingly, all modern attempts to capture the character of science have switched unit of analysis. Post-empiricist philosophy no longer dwells on technical definitions of science but seeks its identification in such matters as its propositional, critical and normative structure. What matter is not research practice but research programmes (Lakotos, 1970). What counts is the *logic* of scientific discovery (Popper, 1959). It

seems that science is science because it balances theory and method, concepts and evidence, guesswork and surveillance. I make no claim here that any particular branch of philosophy has captured the precise and definitive admixture of processes that constitute science. All that I am recognising, with Haack (1993), is that findings are justified when they accord with the evidence (foundationalism) *and* are consonant with other theories (coherentism).

The basic argument (foundhertism!) is thus that standards are forged in the cumulation of inquiry. Experimental research is always carried out *in tandem*. A puzzle is created which yields to a variety of explanations. Experiments are conducted with the purpose of trying to tease out the superior explanation. Provisional explanations flow from these experiments, which may still contradict. Further experiments are the carried out which are designed to adjudicate (Pawson, 1989: 116-125) between the contending explanations. Such a sequence may allow debate to settle in respect of certain aspects of the original problem, but still leave other threads untouched, … and so the process continues. The point here is that progress is *not* a matter of using an obviously superior research strategy or a patently more accurate data collection procedure that will always 'win the day'. Indeed, the experiments always rely on common instruments, take on board a shared set of well-established theories, and try to build on points of agreement between previous designs. The study that 'scoops the pool', provisionally anyway, is the one which produces the most convincing explanatory ensemble. In short, standards are an emergent property of inquiry. Standards are the medium and outcome of inquiry. Standards are always capable of development and revision. Science is the process of building on and revising what went on before and that includes notions of research quality.

It is a curiosity, then, that 'systematic reviews of evidence' are all about the cumulation of inquiry and yet the cumulative process of scientific investigation is not its model for research quality. Meta-analysis has quite a different vision of the growth of scientific knowledge. Quality assurance and research synthesis are seen as separate endeavours. The vision is of a whole range of investigations occurring *in parallel*, with each enquiry being regarded as a separate atom, resulting in a 'hopeless jumble of dissimilar findings' (Hunt, 1997:13). And on this view, science takes stock by first of all reviewing the quality of each independent study in order to filter out the 'unscientific', and only then moves on to an orderly synthesis of the evidence by pooling periodically the results of those investigations deemed 'scientific'.

We will come to the precise mechanics of this process in the next section, but I first want to affirm the actuality of this cart-before-the-horse structure by looking at the basic rationale of methodological appraisal from the point of view of one of its key exponents:

> Preparing a review entails many judgements. The focus of the review must be decided. Studies that are relevant to the focus of the review must be identified, selected for inclusion and critically appraised. Information must be collected and synthesised from the relevant studies, and conclusions must be drawn. Checklists can help in this process. (Oxman, 1995)

Here we light upon one of the great organising metaphors of systematic review, the 'preparatory checklist'. Oxman compares it to flight preparation before take off. Airlines do not rely of the brilliance of the pilots, or their hunches that the plane is flying OK, or their memory that everything has been serviced. The expectation is that they work (precisely, rigorously, laboriously and transparently) though an instrument checklist of vital components. It is also rather crucial that this happens *before* take-off. And so it should be, he argues, with systematic review. The narrative review is liable to harbour preferences, take short cuts and thus rely on faulty sources. The systematic review, however, does not leave the runway without a green light on the quality of all of its component studies.

The question is, of course, whether this procedural uniformity is quite the thing to lead us on a voyage of discovery rather than the trip to Majorca. It is beyond the scope of this paper to do so, but I want to question strongly Hunt's assertion that the invention of meta-analysis is a 'breakthrough applicable to all sciences' (1997: 12). The curiosity, surely, is the absence of these checklists and kite-marks across the mainstream of science. Where are the published inclusion criteria for *'Assessing Research Quality in*

Particulate Physics"? Where are the quality checklists for the *'Assessment of Mathematical Proofs*'? Where are the collaborations devoted the production of *'A Hierarchy of Evidence of Methods of Genetic Mapping"*?

I leave these questions as rhetorical challenges and from this section take forward the message that standards are important, indeed all-important for science and thus crucial for the development of its latest offspring, namely evidence-based policy. Otherwise, we would have to conclude that, for policy development purposes, every opinion of every Tom, Dick and Harriet counts equally. By the same token, I have tried to illustrate that, in broad terms, there are no instant measures of sound science, no quality quick-fixes, no appraisal ready-reckoners. By the lights of post-empiricist philosophy of science it is quite, quite implausible that research standards will ever be identified via technical capacity or data quality or procedural protocols. This is not a particularly glum conclusion, however, for as I have sought to show, the questions of 'why we need to assess research quality?' and 'why does science cumulate?' turn out to be one and the same thing. Quality is synonymous with successful cumulation. Meta-analysis, it seems, has taken a somewhat different path and it is to this cul-de-sac we now turn.

II. How should research quality be assessed?

This section of the paper takes the form of a selective and critical examination of some recent attempts from the systematic review community to capture standards in the form of prescribed schedules of quality indicators. This is, in short, an assessment of assessment systems. My method is basically to compare ends and means - what are the objectives of the standards invoked and how is the judgement made as to whether they are met. My thesis is that it is possible to insinuate preliminary quality checklists into reviews of research if, and only if, the research objective is descriptive, narrow and one-dimensional. If, however, one is seeking to justify that evidence is of sufficient quality to sustain wide explanatory objectives, then the assessment must be made alongside explanation building, rather than as a preliminary to it. The question that hangs fire in all this is, of course, the requisite explanatory scope of evidence based policy.

Standards for Evidence-based Medicine

I begin by identifying the template for research quality established in evidence-based medicine. Here, there is a narrow, consensual and easily recognised 'gold-standard' for research synonymous with the usage of randomised controlled trials. This point of agreement allows for a model, in fact *the* model, of systematic review in which existing research is synthesised by following the (somewhat simplified) sequence specified in Figure one. Fuller expositions using rather more elaborate checklists may be found in Davies et al (2000) and NHS Centre for Reviews and Dissemination (2001).

Figure one: the classic model of systematic review

Clarifying the question for review about the effectiveness of a particular treatment

Searching for primary studies that address this question

Appraising the quality of these studies in terms of their ability to answer the efficacy question

Extracting the data from each proficient study on the outcomes of the treatment in that particular trial

Synthesising the data by aggregating the results of all competent trials

Disseminating the findings about the overall efficacy of the treatment.

What is crucial to the logic here is the close affinity between steps one and three. If it is assumed that there is only one research design that permits authoritative statements to be made on treatment efficacy, then it

follows that appraisal of research quality can be made *directly* and *early* in the review process. And indeed this is exactly the function of quality standards articulated by Oxman above and now well established in evidence-based medicine. Meta-analysis seeks to establish a causal linkage between a particular treatment and a specific outcome and it is deemed that RCTs are the only permissible design for making such inferences.

The awesome logic of experimental control is brought to bear. If subjects are randomly allocated into experimental and control conditions, and the treatment is applied to the former but not to the latter, then any subsequent differences between the two groups must be the result of the only matter on which they differ – namely, the application of the treatment. The research question about treatment efficacy is thus deemed to respond uniquely to this particular research design. In these conditions, study appraisal may act as a *filtering process*, sanctioning only the selected studies that should survive as part of the subsequent data extraction. This filter, incidentally, also sustains step 5 of the process. If synthesis is essentially a process of pooling the 'mean effects' obtained in separate trials into an aggregate 'net effect', then the quality filter ensures the purity of each any every datum that is fed into this statistical exercise. On dissemination of the results, the checklist becomes complete and the overall sequence produces a tight bond between the *explanatory quest* (is the treatment effective?), the *quality filter* (is the trial effective?) and the *explanatory outcome* (an efficacy coefficient).

My observations here are not about affirming the superiority of randomised controlled designs but to show how they have come to assume a pivotal place in the dominant model of systematic review. The vital point is that the application of preliminary quality filters in the initial phase of a review is appropriate only under a very specific condition. Applying an inclusion criterion at step three presumes that quality appraisal captures all necessary and sufficient information to identify the bona fide answers to the question posed at step one.

The crucial issue is what happens when a research question comes along to question the quality of the quality criteria? Is the RCT design robust enough to sustain quality in respect of all the questions we might ask about treatment efficacy? There is a sense in which evidence-based medicine is already braced for this. All recent study inclusion criteria have come to appreciate that there are RCTs and RCTs. Experiments performed on and by human subjects do not conform exactly to the iron logic described above. Patients change their minds about co-operation with trials, and both patients and clinicians suffer from wishful thinking, which may itself generate positive outcomes. The simple treatment-on/treatment-off distinction is not as watertight as it first appears. Accordingly, the best RCTs use designs which also involve placebos, blinding and double-blinding of patient and care provider, forms of analysis which stipulate an 'intention to treat' demarcation of the experimental and control conditions, and much else besides (NHS CRD, 2001:9). There is nothing much here to trouble the logic of the 'quality filter', the central ratchet is merely tightened, we must be aware of the difference between mere experiments and pukka experiments.

A rather more severe threat to the old one, two, three, four, five, six in systematic reviews arises from a rather different expansion of the explanatory agenda. What is sometimes called *stage-two meta-analysis* has come to appreciate the utility of going beyond the single 'does it work?' question. That is to say, treatments obviously work better in some circumstances than others (stage of disease, coexistence with other ailments and treatments; age, sex, social class, race, genetic make-up of patient, etc. etc.). In such circumstances it is deemed that the appropriate task for systematic review is to begin to tackle the rather more subtle 'under what conditions does it work?' question.

What difference does this make to the inclusion criterion for the original studies? In one respect confounding factors such as the above are handled by random allocation. This procedure should ensure a balance of confounding conditions between experimental and control conditions, and so in this respect the orthodox quality hurdle also holds good. But trials are performed on the group of patients pragmatically assembled under particular institutional conditions rather than on the population of all sufferers. In such circumstances,

it is reasonable to assume that any particular trial, as a whole, may pull in subjects who may be rather more or less susceptible to the treatment. This eventuality is handled in meta-analysis by using tests of 'heterogeneity' and by the inclusion of 'moderators' and 'mediators' in the analysis, the latter addressing the more subtle, 'under-what-conditions' analysis.

Putting aside the question of the effectiveness of such statistical modelling, I concentrate again on the issue of the quality filter. Under this revision, meta-analysis is used to arrive at a synthesis on the issue of whether treatments work better or less well for different subjects and in different circumstances. Clearly, such explanatory eventualities could and should also have featured in the primary studies. Moreover it is a reasonable expectation that such a methodological challenge will have been handled more or less successfully from original study to original study. The primary trials may be expected to differ in respect of such design and technical decisions as follows: i) have they collected and reproduced data on the trialists' backgrounds and circumstances; ii) how many and which of these contextual feature have been inspected; iii) how are these features deemed to condition the treatment effect; and iv) whether any analysis have been performed on sub-groups of patients and on treatment variations. All of these features are clearly crucial to the contribution a study can make to the more subtle and conditional review question. They are matters that ought to feature in selection for the review. But do they become formally incorporated into the quality checklists?

My view is that they do not, that pragmatism prevails, and that preliminary quality appraisal pretty much ends with the orthodox questions about the grip of experimental control. The reason for this is instructive. The question of how well a study has performed on the matters raised in my brief 'supplementary checklist' above is clearly a matter of judgement. What is more, they are often issues that cannot be decided in advance of analysis. We cannot know with any certainty, before the synthesis is performed, which contextual conditions will turn out to exert a powerful influence.

The consequence is that decisions on which studies to include in the secondary analysis are made piecemeal. Inclusion is likely to be coloured according to whether a sufficient sample can be mustered to perform the meta-analysis, or whether enough studies have data on the same mediators and moderators to provide valid estimates. When such analytic demands increase the review is likely to 'weaken' it inclusion criterion. When, for instance, the treatment under consideration is a public health initiative and the potential scope for mediation and moderation effects multiplies, the quality hurdle is usually lowered, so that even modest before-and-studies may suddenly find favour and 'should be considered' (NHS CRD, 2001:8).

What is more, 'running repairs' are sometimes performed on included studies that have turned out to be wayward *after selection*. Studies with so-called 'missing data' can be included in a meta-analysis by using 'estimates' of how an unmeasured mediator typically performs. Studies which meet inclusion criteria but which produce outlier effect sizes are sometimes 'Windsorised' (i.e. reset within typical values). In short, and despite the checklist rhetoric, tacit and post-hoc judgements creep routinely into the process of quality appraisal. Once again, I pass no judgement on the legitimacy of these operations. I simply note that, when the research question becomes more complex, it seems that the systematic review aeroplane does in fact take off with quite a few nuts and bolts untightened.

Standards for Evidence-based Policy

Having uncovered a glimmer of tacit judgement in quality appraisal in the most formal corner of systematic review, I now want to turn to the core issue of the paper, namely the conduct of quality appraisal in evidence-based policy. Social interventions are more complex than medical interventions: their implementation is never uniform, their efficacy depends on the interpretations of the intended subjects, and the footprint of their outcomes often varies markedly across subjects and their circumstances. Given this programme anatomy, it is widely accepted that evidence-based policy will rest on a broader pedestal of research

strategies than does evidence-base medicine, and it is conceded that the business of research quality will, accordingly, be more elusive and difficult to pin down.

Nevertheless, moves are afoot to formalise standards appropriate to the constituent research strategies that comprise such a pluralistic research base. In this vein, there are now available a significant number of published research criteria for appraising 'qualitative research', 'evaluation research', 'action research', 'emancipatory research' and so on. This spurt of alternative quality indicators brings me to the crux of this paper. What is the nature of these newly minted standards? How are they meant to operate? Are they stamped by consensus? Can they act as pre-determined quality filters? Are they safe in the hands of the research staff who will apply them? In short, and in deference to the demands of the conventional model of systematic review, are they pivotal, *step-three* standards?

Let me now sketch what is (and what is not) on offer in the new standards regimes. My portrait will follow a good-news-and-then-the-bad format. I commence with a description of the principles that have informed the quest for the new standards, noting their sterling intentions. I end with a tale of unintended consequences. The end product of these efforts concerns research quality to be sure, but the typical outcome is not a 'quality filter' but a 'quality charter'. Rather significant implications flow from this difference for the conduct of systematic review.

My observations are drawn from: *'The Programme Evaluation Standards'* of the US Joint Committee on Standards for Educational Evaluation (1994); *'Action Research: A Systematic Review and Guidance for Assessment'* produced as part of the UK National Health Service, Health Technology Assessment Programme (Waterman et al 2001); and *'Assessing Quality in Qualitative Evaluation'* produced for the UK Cabinet Office (Spencer et al, 2003). I also draw on my own experience as part of a team conducting an appraisal of the potential for standards in the realm of social care knowledge for the UK Social Care Institute for Excellence (Pawson *et al, 2003??)*.

Let me begin by highlighting three dominant characteristics of the new standards movement:

I.Fit-for-purpose: The first dictum to fall is the 'hierarchy of evidence' in which one approach is deemed to sit atop of the research pecking order with the remaining strategies tailing off in terms of their utility for secondary analysis and review. In the new regime all methods are deemed to have a place under the revised first principle of 'fitness for purpose'. Most exercises thus begin with a modest justification of the place of the particular paradigm within policy inquiry. The Health Technology Assessment review is typical in this respect in a section on 'popular misconceptions and criticisms', which eschews the paradigm wars in the public health field and argues that action research, with its developmental and participatory goals, should be judged 'according to its own terms' (Waterman *et al*, 2001:3).

The door is then open for tailor-made standards, prepared and ready to meet the diverse objectives of inquiry. Fitness-for-purpose becomes defined in terms of the requirements of the broad families of social enquiry. There are some generic standards such as the requirement that 'the aims and objectives of research should be clearly stated', which should hold irrespective of mode of inquiry. But beneath these are horses-for-courses standards: if the original research deems itself 'qualitative' then reviewers should be on the look out for standards a, b and c; if the research claims to be 'action-oriented' research then they should have a regard for d, e and f; and so on. This rationale has been stretched to its limit in the SCIE project in which the knowledge of social care practitioners and users is added to the brew on the grounds that their tacit wisdom and experience cannot be discounted as part of the evidence base.

II. Full kit inspections: The next reform is to inculcate standards for a much greater range of investigatory activities. In the classic meta-analytic reviews the quality assessment focus was on *design issues* (is it a double-blinded RCT?) as the telling feature of the original studies. The new standards establishment aims to cover all phases of the research cycle. Accordingly quality criteria span inception to dissemination (e.g. problem formation, design, sampling and case selection, data collection, data analysis and interpretation,

presentation and policy implications) as well as further rules about the interconnection of these parts (e.g. do the findings follow from hypotheses and analysis?).

The Cabinet Office report is instructive in this regard. It is undoubtedly, the most comprehensive study of quality in qualitative research. It is itself a synthesis of 29 existing quality frameworks in just one corner of research. The evaluative tool that emerges parses down the perspective into 18 appraisal themes and identifies these using no less than 85 quality indicators (Spencer *et al*, 2003: 83-89). The Joint Committee's report on programme evaluation is equally extensive. There are thirty standards, categorised into four groups, corresponding to the Committee's viewpoint on the key requirements of an evaluation (1994, appendix1). The four desiderata are that the inquiry should have: i) *utility* - it should serve the information needs of intended users; ii) f*easibility* – it should be realistic, prudent, diplomatic and frugal; iii) *propriety* – it should be conducted legally, ethically and with due care towards all participants and users; iv) a*ccuracy* – it should be technically adequate to determine to merit and worth of a program. The working motto of the new movement is clearly, 'mind the quality *and* feel the width'.

III. Standards enunciation: Another common regime change is to provide a more varied menu for conveying the standards. I am not referring here to the content of the standards (discussed under the previous heading) but the means by which they are explained to potential users. In the orthodox quality regimes, standards are expressed as 'should' statements – in order to qualify as bona fide research the investigation *should* follow such-and-such a research design. These imperatives might also stretch to include operational details such the ones noted above about how to include drop-outs in experimental and control comparisons. The standard format, in short, is the key directive backed up with operational clarification.

By comparison the recent quality regimes are prolix. Because the new-fangled standards are many, varied and, above all, subtle a whole package of aide-mémoires is brought to their exposition. The goal is not simply operational clarity but stretches to include: the principles under scrutiny, scope conditions about what is and what is not covered under any given criterion, operational definitions and markers about how to judge good quality in practical terms, as well as examples and vignettes providing illustrations of and familiarisation with best practice. The Joint Committee (1994:7). provides the most compelling template of presentational forms to capture and impart the meaning and significance of each of its standards:

a. *'descriptive title'* delimiting its subject area, followed by

b. a proposition expressing the standard as a *'should statement'*, followed by

c. *'guidelines'* on the activities needed to meet the standard, followed by

d. a list of *'common errors'* that the standard is designed to avoid, followed by

e. *'illustrative cases'* to give a picture of how the standard might be applied in practice.

The importance of such a show-and-tell approach to quality assessment cannot be over-emphasised for it reveals a change in motive behind the production of standards. Standards must exist in an ongoing process of communication if they are to exist at all. Accordingly, the Joint Committee targets its efforts (1994:4) at 'people who commission evaluations, conduct evaluations, and/or who use the results of evaluations.' The HTA guidelines (Waterman *et al*, 2001:60) conclude on a similar note, 'The provision of a method to evaluate action research is essential to the future development of the action research process. The guidance provided as part of this review is seen as a starting point for the development of this process'. The goal of the new regimes, it seems, is very much about the inculcation and development of research standards rather than policing quality thresholds

A veritable standards 'industry' is gathering in pace, creating many other models along these lines (a more detailed examination of a greater range of quality packages may be found in Pawson (2003)). Developing standards for the entire toolkit of social research is a daunting task and we should be left in no doubt that these labours constitute a significant turn in expectations about what research can deliver. However, the new apparatus is also gathering in paradoxes, most especially in terms of the eventual and all-important issue of the conduct of systematic review. I now turn to concerns about the practicality of these new instruments. Let me mention just six difficulties:

I. Boundless Standards. Broadening the domain of quality standards has created quality registers that are dinosaurian in proportion. The documentation begins to look less and less like convenient checklists and more and more like textbook glossaries. The range and scope of issues raised by way of appraisal has multiplied. Previously, the design dimension was all-crucial in appraisal. But now, and quite properly, questions may be asked about: the credibility of findings, the originality and reach of findings, the match between research purpose and outcomes, the scope for generalisation of findings, the composition of the sample, the appropriateness of the analysis, the attention to ethics, the richness of the data, the coherence of the reporting, the adequacy of the conclusions, and so on. Note that this rather typical line-up merely raises some pertinent quality dimensions and not the operational criteria that then have to be used to judge whether each implicit standard is met. Each dimension tends to throw up a range of plausible indicators with, as noted, one recent instrument chalking up more than four score measures of quality. Now, size really is a problem in this respect. It is quite, quite implausible that systematic review could operate with this level scrutiny of baseline studies. The house would be endlessly measured up, but never built.

II. Abstract Standards. Broadening the domain of standards results in the usage of 'essentially contested concepts' to describe the requisite rules. Amongst the great weasel words that find their way into review criteria are 'rigour', 'clarity', 'sensitivity', 'credibility', 'consonance', 'insight' and so on. Readers are invited to spot how routinely these abstractions crop up in the spelling out of standards. Now, it is fairly easy to decipher whether a study has or has not utilised an RCT and thus deliver a pass/fail verdict on its quality. But a concern for rigour, clarity, sensitivity and so forth generates far tougher calls. For instance, the Joint Committee's directive that 'the conclusions reached in an evaluation *should* be explicitly justified, so that stakeholders can assess them' is hard to pin down. In one sense, it is blindingly obvious - teachers from primary school onwards also implore us to 'make your conclusions clear!'. In another sense, it is tautological - conclusions *are* justified propositions. And, in other sense, it is perfectly empty – so much depending on the nature of the particular conclusions and their import.

III. Conflicting Standards. Broadening the methodological base of standards increases the chances that they may gainsay. This is a particular threat to the production of standards for the social care evidence base, which, as I have mentioned, seeks to find a place for practice wisdom alongside formal research. On the one hand, this ambition is an invitation to an 'experimentalist' standards regime, which insists that subjects should be randomly allocated, regardless of their preferences, to treatment and control condition. Moreover, research quality is thought to increase if placebos are used to 'blind' this allocation, thus wringing out every last vestige of human intentionality from the test of whether the treatment works. Another hand goes out to the 'emancipatory' perspective, which requires user-involvement at every stage of the research process. In its strong form, this paradigm envisions 'survivors' of social care as possessing special insight and any research done in its name has to pass the quality-standard of user-control (Wilson and Beresford, 2000). Clearly, such a conjunction produces a contradiction of the squarest kind. More generally, I would argue that as quality markers increase in their number and coverage and philosophical stance, the tensions and ructions and wrangles between them also multiply.

IV. Permissive standards. With the above problems in mind, most recent standards compendia are prefaced and concluded with astute remarks stressing that their usage requires 'judgement'. The following from the Cabinet Office report is a prime example

> We recognise that there will be debate and disagreement about the decisions we have made in shaping to the structure, focus and content of the framework. The importance of judgement and discretion in the assessment of quality is strongly emphasised by authors of other frameworks, and it was underlined by participants in our interviews and workshops. We think it is critical that the framework is applied flexibly, and not rigidly or prescriptively: judgement will remain at the heart of assessments of quality.' (Spencer *et al* 2003: 91).

One would be hard put to find a more sensible statement on the sensitivity needed in quality appraisal. But, for some, such a twist heralds the arrival into research synthesis of a decidedly oxymoronic character, the 'permissive standard'. Whichever view one holds, it is clear that the application of quality standards is moving further and further away from being an efficient and unproblematic preliminary to systematic review.

V. Composite standards. The use of multiple quality criteria also raises novel questions about their balance. Should 'careful exposition of a research hypotheses' be prized more than 'discussion of fieldwork setting'? And how do these weigh up alongside 'clarity and coherence of reporting'? The permutations, of course, increase exponentially when one is faced with over eighty quality indicators. By and large the new standards regime has resisted formulating an algorithm to balance the different contributions (Spencer *et al* 2003: 82), once again maintaining that the overall worth of a study is a matter of 'judgement'. In this dilemma, I am reminded of the trials and tribulations of another form of assessment system. Marking schemes have come to dominate in school examination assessment, the contribution of the diverse points a candidate might make being rendered down into formulaic, pre-determined, specimen answers. As far as I am aware, such regularisation has been resisted in PhD examination, the reasoning being that it is impossible to pre-specify what constitutes an 'original contribution to knowledge'. Given that the entire exercise of evidence-based policy is about best practice in research, might it not be that the doctoral examination produces as well as its key benchmark are somehow appropriate? Such a thought, however, has rather dispiriting consequences in respect of the amount of labour and skill required in the operation of a quality filter – not to mention the fact that PhD examiners often disagree.

VI. Pick-and-mix standards. The growth of multiple quality criteria also raises another extremely awkward question about the 'unit of analysis' for quality assessment. The classic inclusion criterion produces a pass/fail verdict on the *whole study* – well designed studies produce serviceable findings. By contrast, diverse quality indicators will undoubtedly identify curates' eggs. There is no reason to believe that the report of any piece of research will cover, let alone deliver upon, a checklist comprising scores of quality indicators. This raises a key dilemma, indeed a potentially devastating one in terms of research synthesis, about whether *evidential fragments* rather than entire studies should be the subject of quality assessment. Is it not possible that an otherwise mediocre study can produce a pearl of wisdom? In the experimental domain best research practice is summoned together, focusing on achieving a valid outcome measure. But in other perspectives, the end result does not take the form of the single, safeguarded proposition. If the research medium is 'thick description', then there is no way of passing forward all findings into research synthesis. And if synthesis has to be selective, then so too must be our judgements on the worth of a study.

In short, the new standards are prone to complexity, abstraction, contradiction, imbalance and fragmentation. For all of these reasons I argue that they cannot act as quality filters in systematic review. They cannot qualify and disqualify studies as a preliminary to the main business of research synthesis. The overall verdict must be that they do not pass muster as 'stage three' standards. And yet this appraisal function is precisely the expectation for them on the part of many commissioners and designers. The documentation of standards often concludes with an overall tabulation of key principles and quality indicators, with a column left empty and designed for an appraiser's notes 'on the study being appraised' (Cabinet Office, 2003: 83-89)). Much of the development of the new indicators has occurred within and on the borders of the Campbell and Cochrane Collaborations. In these quarters, the conventional model for review (Figure

1) is still regarded as the holy grail and the new standards are intended to simply as refinements to 'bolt-on' to the existing apparatus.

Let me make the point of my argument crystal clear at this point. None of the above makes a case against using qualitative data, user-inspired evidence, developmental evaluations, administrative records and so forth in systematic review. Indeed, in my view, evidence-based policy will sink without trace unless it incorporates a more comprehensive evidence base. What is more, none of the above denies that there are standards appropriate to qualitative, action, evaluation research and so on. The Herculean labours described above are proof positive that in all modes of inquiry, wheat can be sorted from chaff. Accordingly, the critique of quality appraisal forwarded here is not an evaluation of their content but a thesis that the appraisal instruments actually serve a function other than that stated and intended.

In order to support this little heresy, let me rehearse again the nature of the new standards in order to make clear how (and how not) they may function. I have just argued that the new, all-purpose standards are not handy, ready-made quality filters. So what are they? As can be seen from their coverage, they operate on a much higher plane. They are desiderata rather than rules. They are ambitions rather than performance indicators. They are paradigm gazetteers rather than decision points. They are quality frameworks rather than analytical tools. They are tribal nostrums rather than critical questions. Their true character is perhaps revealed best in a moment of self-reflection by the Joint Committee (1994:2): 'A standard is a principle mutually agreed by people engaged in professional practice, that, if met, will enhance the quality and fairness of that professional practice.' The new quality frameworks are, in short, 'methodological charters'.

The most profound sign that we are dealing with the creation of mutually-agreed professional ordinances lies in the mechanics of compiling the new frameworks. From whence do they spring? Basically, they are distillations of the experience and expectations of research practitioners. The are repositories of the craft wisdom of the denizens of particular research and evaluative trades. A crucial step in the formulation of the quality indicators is consultation with 'experts' in the respective methodologies. These play a part in both the formulation of standards and in the verification of the end product. An appendix in the Joint Committees' monograph lists the membership of the support groups who were responsible for the their standards (1994: 191-202). The list goes on for twelve pages, members being composed of 'project staff', 'validation panel', 'chair', 'members', 'panel of writers, 'consultants', 'national review panel', 'international review panel', 'participants in field tests', 'participants in national hearings', 'student assistants' and 'clerical assistants'. It is a similar, if rather more modest, tale with respect to the Cabinet Office Study. Consultation and the measure of agreement it afforded was a feature of their new standards for qualitative research The roll call in this case can be thought of as a list of the good and great in UK qualitative methodology (Spencer *et al* 2003: 3). The expansive result is the same in both cases and attests to the old joke about the committee, dispatched to design a mouse, and turning up with an elephant.

With this rather jocular observation, I reach the sober conclusion of this section. The gradual slide from assessment tool to paradigm charter has an unintended and untoward consequence for the conduct of a systematic review. Not only does it replace a sharp decision point with a flabby set of debating issues, it also breaks the crucial link between the question addressed in a review and the quality criterion that should be applied. Recall that the backbone of systematic review in evidence-based medicine was the rock solid connection between the original hypothesis (does treatment X have rehabilitative outcome Y?) and the obligatory design (random allocation of subjects to treatment [X] and control [not-X] conditions, with pre and post-treatment observations on Y). In this classic model, the purpose of the review coincides exactly with the quality criterion. Such a direct stage-one to stage-three link disappears completely with the advent of the new comprehensive quality regimes. Boundless standards do not make for expedient inclusion criteria.

In the next section, I move on to contemplate how research question and quality criterion might be rejoined. But as a final comment here, I reprise the key principle of 'fitness-for-purpose', which launched most of

the recent discussion about quality appraisal in policy research. An undoubted blessing of the new quality regimes is the widespread appreciation that there is a 'trade-off' in the utility of different research procedures: RCTs are hot on internal validity but silent on external validity; unstructured interviews allow elbow room for the subject's viewpoints but make their comparison difficult; focus groups animate research subjects but are liable to be hogged; etc. etc.

This manoeuvre might seem to offer the new standards regime a clean bill of health in terms of the very first stricture of this paper about the overemphasis on *particular* methods as kite-marks of quality. I remain unconvinced, however, in as much as the 'purpose' pursued in the new quality regimes is still very much obsessed with the technical palate of research rather than its wider explanatory landscape. As I have tried to show, the new standards are largely method-inspired guidelines and spread their wisdom around the entire research cycle and, as such, are not tailored to a specific view of how findings are to be pulled together. They remain standards-in-search-of-a-question and standards-in-search-of-strategy-of-synthesis. Insofar as they are 'fit-for-purpose', the new standards are 'true-to-paradigm' rather than 'matched-to-task'. What is sorely needed is a way of harnessing the new thinking on standards within a more appropriate machinery for conducting research synthesis.

III. When should research standards be assessed?

I now move to the third concern of this paper, the issue of 'when' the appraisal of research quality should be performed. This turns out to be a choice of real and surprising significance for the entire function of research synthesis and not just matter of locating a convenient timetable slot for appraisal within a review. My contention here (recall Part I) is that research standards have a different source from that promoted in the orthodox model of systematic review, and it is this taproot that should be drawn upon. Scientific standards are forged in the hurly-burly of interpretation and counter interpretation of the significance of evidence and it is this process that should be imitated as we review and reappraise previous inquiries.

Doing systematic review always confronts us with a range of explanations and the sorting and sifting of explanations, whether we like it or not, is key to the job. The implication for evidence-based policy is that that the worth of a study is never a pre-given, but a matter that can only be determined as its inferences are granted a place in the developing explanation. The thesis for this section, and the major claim of the paper, is that quality is assured only when the evidence is put to synthesis. Sticking to the numbering in the traditional review sequence of Figure one (which I will revise presently) we can say that quality appraisal is not a separate *stage three* task but in fact is only realised at *stage five*. Synthesis and quality-appraisal are one and the same thing.

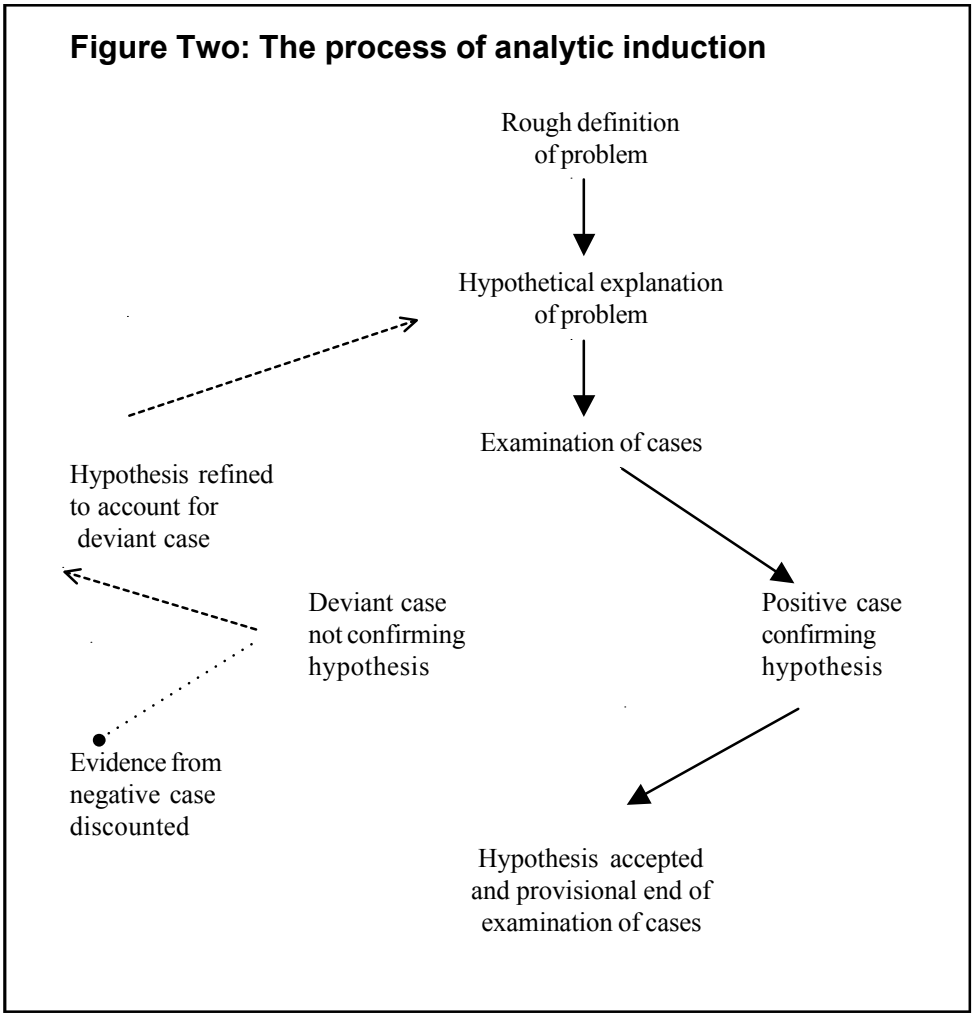
It is of particular interest to note that this sentiment is not some blushing debutante, making her first public appearance on these pages. I have tried to demonstrate in Part I how, in post-empiricist accounts of science, standards are understood as the medium and outcome of inquiry. More curiously, one notes that the same notion is the mother-of-all-ideas about objectivity in the ethnographic and interpretative corners of social sciences. Before the emotionalists and the post-modernists got their hands on qualitative method (Gubrium and Holstein, 1997), this style of research used to make claims, albeit modest ones, for the validity of its findings. It was, and in some quarters still is, assumed that observational methods can arrive at privileged accounts of events in the field by dint of a particular analytic process. Various nomenclatures are used to describe these forms of analysis, so I will skim the surface here in proffering just a couple of characteristic strategies used to justifying qualitative accounts. The first sees the ethnographer as a 'pattern builder':

> For the pattern model, objectivity consists essentially in this, that the pattern can be indefinitely filled in and extended: as we obtain more and more knowledge it continues to fall into place in this pattern, and the pattern itself has a place in a larger whole. (Kaplan, 1964: 335).

The phraseology here - the drawing in of 'more and more knowledge' - is redolent of the themes of systematic review, and it is a curiosity that those bursting to include the qualitative dimension within evidence-based policy have not pounced upon it. What is signalled in this quotation is the idea that qualitative explanation is holistic, that the worth of an individual datum is secured by its place in an unfolding sequence of actions, reactions and counteractions. That one group of stakeholders behaves in one way makes sense of the response of others and this in turn clarifies the predicament confronting a further group. Such is the stock-in-trade of qualitative explanation and process evaluation, and this pattern-building anatomy might be expected to figure more centrally in the attempt to synthesise such knowledge.

Stranger still is the neglect of 'analytic induction', which is perhaps the one qualitative analytic strategy that aspires to universal explanations of phenomenon. I have no tidy, single definition to quote, but the general idea will be familiar and goes back to Lindesmith (1947). The research begins with the rough definition and a hypothetical explanation of a problem. This prompts the examination of empirical cases to test out the worth of the hypothesis. Some of these will satisfy the tyro explanation but some will not. Confrontation with wayward cases prompts one of two moves. Either they are deemed to fall outside the scope of the explanation, which is correspondingly narrowed, *or* the hypothetical explanation is remodelled to fit both existing cases and the apparently awkward customer. More and more cases are examined until the researcher has a robust theory that has the capacity to deal with a multiplicity of cases. Figure two is a representation of this process, which is adapted from Bryman (2001: 389).

Figure Two: The process of analytic induction



The most distinctive suggestion here, of course, is the notion that ethnographers might, somewhat uncharacteristically, stride from case study to case study. Analytic induction sees this as a process of one researcher picking up the conceptual baton from another, so that explanations are refined in suites of research, rather than ethnography forever producing one-off 'tales from the field'. Once again there are strong echoes of systematic review here. Not too great a stretch of the methodological imagination is required to perceive that the prospective research design in Figure two could be applied retrospectively in explanatory autopsies of completed studies.

Most important for the argument here, however, is the role of evidence as conceived in analytic induction. It is *not* assumed that research pursues the single hypothesis. Thus it is *not* assumed that research evidence may be stockpiled as positive or negative, allowing the researcher to conclude yea or nay on that thesis. Rather, the examination of the empirical evidence from a particular case results in a decision point. One of the crucial outputs here is that the researcher may choose to pursue a revision of the theory in the light of negative evidence. The reworking of theory forces a reconsideration of the utility of the evidence. Findings that are damaging to the original theory may prove supportive to a revised one. The crucial point here is that each iteration of the process 'necessitate[s] a reanalysis and reorganisation of the data'. (Bryman 2001: 390). In short, the collision with data forces perpetual revisions to explanatory scope. And here lies the crucial difference with the view from systematic review. Rather than being cast in stone (otherwise known as a search protocol) as the first item in a systematic review, the question under investigation is identified and then revised and then revised again in cycles of analysis. Judgements on the pertinence and thus the quality of evidence are never made statically.

My plan at this point is to draw together some of the themes of post-empiricist philosophy of science with pattern explanation and analytic induction into a working model of research quality in systematic review. Although these perspectives might already seem a rather motley crew, I should mention that there are several other cousins and second cousins of these ideas that could have been put to use. This is not an exercise in the modern history of ideas, so let me rest content with a couple of quotations that reveal a distinct family resemblance to the ideas pursued here:

> The worth of studies, in their view and ours, is determined in the process of achieving synthesis. (Noblit and Hare, 1988:16)

> Exploring results that contradict or fail to confirm previous findings provides justifications to continue inquiry and to determine under what contexts and conditions the previous results are supported or not supported… These are systems of building checks and balances that prevent qualitative inquiry from producing fictitious and fickle results and provide us with confidence that far exceeds Cochrane's category of "mere opinion". (Morse, 2001:203)

The first proposition is from the school of 'meta-ethnography' (it also makes passing reference to a couple of third and fourth cousins), which is concerned with the attempt to construct an 'inductive and interpretative form of knowledge synthesis'. The second quotation is the base from which Morse builds further forms of review that she refers to as 'qualitative outcome analysis' and 'meta-synthesis'. This reference is a fraternal nod in their direction but also keeps a little distance because I am not interested here in the construction of a qualitative alternative to quantitative meta-analysis. Knowledge synthesis in social policy requires the rooting out of evidence from all sources (quantitative, qualitative, comparative, historical, legal, administrative, tacit, etc.). Research synthesis should occur between paradigms, not within them. Hence, I look back a generation for inspiration: to pattern building for its notion of the confederation of findings, and to analytic induction for its idea of the remodelling of theories. Beyond that, I summon the spirit of Lakatos and others to inspire the idea of explanation building as the route to research quality.

Enough of the family history, how might the above themes be used to capture the idea of research quality in explanatory synthesis? My main thesis is to insist that the link between the substantive goal of a review and its quality criterion must be re-forged. Although I have criticised the attempts to transplant the logic of

evidence-based medicine to evidence-based policy, remember that the success of the former resided in its one-review-question-matched-to-one-methodological-standard approach to research synthesis. The point thus holds that research standards should, above all, be pertinent to the fundamental questions asked in the review. Systematic review should not concern itself whether the constituent studies have met some generalised aspirations of qualitative research or have been faithful to the broad modus operandi of action research and so on. The overall quality question is whether the various studies under scrutiny are robust enough to make a rigorous contribution to evaluating the policy issue under test.

So what are the evaluative questions that underpin modern policy research? Without a doubt, prevailing opinion is that the simple 'does it work?' question, asked of a single intervention, has been superseded. Not unreasonably, evaluators are also charged with supplying answers to the following questions: 'for whom does it work?', 'in what circumstances does it work?', 'does it work fairly?', 'does it work efficiently and parsimoniously?', 'how is it best delivered?', 'how can it be made to work better?', 'will it continue to work?', 'what are the weaknesses in the implementation chain?', 'does it square with the broader policy agenda?', 'how does it fare against rival interventions?', 'does it complement or contradict other initiatives?' etc. etc.

In summary, the review agenda is shifting from trying to establish enduring empirical generalisations to one of providing explanations. It would be unimaginable to conduct a review that addressed all of the questions in the previous paragraph, but it is imperative to consider the demands on systematic review assuming that the underlying quest has drawn outwards from 'does it work?' to 'why does it work?'. How would the quality issue then be addressed?

Such an assessment cannot be made in one fell swoop; it requires the staged incorporation of quality questions as the synthesis develops. This section will conclude with a summary of the overall process through which explanatory synthesis and quality appraisal intertwine. But first let me isolate some of the crucial steps in synthesis and the corresponding standards. There are three of them and they can be thought of as an 'ingredients assessment', a 'pattern assessment' and a 'scope assessment'. Evidence never speaks for itself and the mark of quality of a piece of research is that appropriate inferences are drawn from the evidence available. Accordingly, all of these quality criteria are about assessing and supporting inference making.

Assessment #1 - the appropriateness of the explanatory ingredients.

The synthesis has an explanatory quest. Minimally, this will involve an understanding of how programme processes lead to programme outcomes. Frequently, it will be more complex than this and may involve, for instance, a theory of how a range of intermediate outputs have to be achieved before the output goal is reached. Alternatively, the crucial explanation might be about achieving the correct legal, constitutional and institutional framework for a programme process to operate in an optimal manner. Then again, programme subjects might be considered a key moderator of effectiveness and the reviewer might have to search for primary evaluation that considers different sub-groups of subjects. The first and broadest quality question is whether the various studies under scrutiny possess the appropriate data to test the specific element of the hypothesis under test. Do they deliver the appropriate explanatory ingredients? The task for quality appraisal here is to match the study to each of the explanatory feature required in the synthesis and to ensure that that study has a basic design capable of delivering that type of datum.

In short, an initial 'fitness-for-explanatory-purpose' test is required. The synthesis attempts to weld together information on a number of explanatory ingredients, and studies are scrutinised in respect of which element they can support. This first hurdle, then, takes the form of a question around the type of inferences made in the primary studies. What is the nature of the inferences drawn, and to what extent are those inferences

justified? The yardstick applied here goes no further than an initial plausibility check on the design utilised. It is a 'does-it-do-what-it-says-on-the-tin?' standard. For instance, if one element of the synthesis requires information about the inner workings of programme implementation, then process inquiries are obviously what we seek; if the legal status of a programme sanction is at issue, then constitutional studies are de rigueur. In short, a rather modest quality check is needed at this preliminary stage. It is a matter of finding and aligning 'appropriate' studies.

Assessment #2 - the contribution of a study to the emerging pattern.

As we move further into the review the quality issue transforms. The question now posed, by way of synthesis, is: 'do the inferences made in a study gel with those from other studies?' Ideally, evaluations should always be built with an eye on checking out, adding to and refining what went on in previous studies. Alas, they often have a none-too-splendid habit of being conducted in isolation. The synthesis, therefore, has the task of reconstructing the inferential links between the studies. And, accordingly, the utility of a study in providing these explanatory bridges should be major part of the assessment of its quality.

At the risk of tedious repetition, remember that explanatory synthesis is not the pooling of outcome scores. It is a matter of the connectivity of inferences. For instance, information on some crucial aspect of programme process from study A can be vital in understanding programme outcomes observed in study B. What counts is the contribution to pattern-building. What counts is the fertility of a study in making connections to, and sense of, the inferences made in other studies. The acid test of research quality is whether it provides good explanation and this involves examination of how it gels with other studies in the developing network. To echo Kaplan: as we review more and more studies, they continue to fall into place in a pattern, and the pattern itself has a place in a larger whole.

This explanatory fertility of study is not just a given; it is not just a matter of lighting upon it and watching it perform its explicatory weave. What is required here is a close appraisal at the level of detailed findings. This takes the reviewer beyond the general issue of the appropriateness of the design and gets down to specific concerns about the nature of data collection and analysis. Most evaluative studies fetch up with a range of inferences and conclusions and what is at issue here is the credibility of specific propositions. Is the appropriate research capacity in place to justify that particular claim?

Naturally, reviewers may choose to accept or reject any particular inference on the basis of the quality of the study. This second level of appraisal does have rather more teeth for, as is well known, evaluations frequently over-claim and it may well be necessary to judge that only parts of the original conclusions may be of value. For example, process evaluations often stray into inferences about programme impact and impact evaluations frequently make untested assumptions about the nature of programme processes. The study appraisal process at this stage makes adjudications about which are, and which are not, the valid inferences in a particular study.

Of course, some rather orthodox quality appraisal tools will be brought to bear at this point and rightly so. But note, too, the differences. The first is that this is not a 'whole study' appraisal – it is accepted that a fragment of a study's findings may stand, whilst others may not. Fitness-for-purpose is being evoked again, but what is at issue is the warrant for very specific explanatory propositions. The other novel assumption is that appraisal is performed as a collective act and not just as a standard schema applied independently to each primary study. Studies are being reviewed and being assessed for their ability to deliver on their part of the explanatory jigsaw. Quality appraisal is performed against a gathering explanation.

Assessment #3 - the contribution of a study to the scope of the synthesis.

There is a further outcome of study appraisal that is entirely absent in the conventional models. Although I have adjusted their focus somewhat, the decisions examined to this point have the standard consequences of accepting or rejecting a body of findings. Another possibility is that the reviewer will happen upon a study, assess it as methodologically sound and thus suppose its conclusions are valid, but then recognise that it does not fit well with the developing explanation. Such a study might stand in outright contradiction to other results or it might simply develop an untypical explanation for a rather typical set of findings. Such consequences, which are actually quite commonplace in the hurly-burly of applied research, will force a re-examination of the scope of a review. The review question has to be questioned.

I will return to the mechanics of this in the next paragraph. But first note that such conceptual re-jigging (whilst the flight is underway) is anathema to the conventional model of systematic review. It is shunned at all costs, being avoided by paring down the initial question posed in the review and then paring it down again and then sticking to it. Orthodox review protocols thus demand the simplest of review topics, ideally 'does X impact on Y', with 'X' and 'Y' and 'impact' being defined as tightly as possible.

I have already rejected this departure point, since it not a sensible question to ask of complex social interventions. Such interventions have diverse outcomes and demand reviews that undertake a more explanatory mission. Programme complexity, however, has the obvious consequence that one cannot anticipate, prior to review, the nature of every mechanism and context that might condition programme outcomes. One cannot foresee the explanatory content of every potential contributory study. So what I am arguing here is that the discard pile of 'negative cases', 'outliers', and 'unanticipated consequences' and so on has to be put to positive use.

Reviewers will often come across an original investigation, deemed to be methodologically sound, which indicates the need for an adjustment to the review question. Following the logic of analytic induction, the working hypotheses driving the review need to be adjusted for scope. It might be that the negative case describes a context in which the normal operation of the programme mechanism becomes dysfunctional. It might be that the outlier discovers an unintended and hitherto uninvestigated programme mechanism that appears to have a profound influence on the outcome. But before these new configurations are accepted as having explanatory significance, they too need to be checked against the network of previously accepted results. The test bed still lies in the degree of absorption into the overall pattern.

Absorption, in this case, is clearly not just a matter of observing 'consistency' with the previous cycle of sound studies. These wayward studies are hardly plain old positive replications. Accordingly, what is adjusted is the scope of the explanation. Studies A, B, C, D, E provide a provisional explanation for a certain class of issues X. Study F comes along and appears to show that A-E's explanation holds for only part of the problem (namely X_1), and that a revised theory is required to account for other features of X (namely X_2). Some synthesis of A, B, C, D, E and F thus accounts in a more comprehensive and selective way for all features of X. Achieving synthesis in this way involves returning to the original studies in order to confirm that their evidence is in fact consistent with the more limited domain X_1. The acid test of research quality remains whether a study is able to provide sound explanation, and this involves examination of how it jockeys for position within explanatory patterns in the manner just described.

A couple of final points of clarification are warranted on the nature of the quality assessment themes I have presented here. Without doubt, they involve considerable judgement and methodological acumen on the part of the reviewer. I make no apology for this. Modern quality checklists have steadily expanded to the gargantuan and are strewn with judgement calls. Permissive standards were introduced somewhat surreptitiously, but active discernment is now acknowledged to play a part in the application of every

quality criterion. As we have seen, there are even judgements to be made about which of the multiple, available, quality benchmarks to invoke. The approach suggested here brings a rationale and a focus to this choice. It is an absurdity to wade through a massive checklist of research competencies; the appropriate criterion and the respective judgement should be invoked only when they are needed. What matters is the contribution of method to inference making and it is best to judge quality only when these inferences have been isolated. Research quality and research synthesis are inseparable. But, thankfully, judgement on both of them remains. Indeed, it would be deplorable if the act of research synthesis were not every bit as skilled as the conduct of the component inquires that are its subject matter.

Finally, we come to the paramount issue of how these three principles can be packaged with a method of systematic review. This fusion is attempted in the model shown in Figure three which is intended to act as counterpoise to the classic review checklist (reproduced in Figure one). It constitutes a theory-driven approach to systematic review. The overall idea of achieving explanatory synthesis is reproduced as a seven-stage model, with a repeat loop at point eight. The issue of 'research standards' crops up at several points, but note the special emphasis within step 6, in which quality issues become clarified:

Figure three: explanatory synthesis as applied in systematic review

1. Elicit an approximate theory of how the intervention is supposed to work. This may often be found in policy proposals and the administrative and legal background to an intervention, as well as in the social science literature discussing the key concepts that lie behind an intervention. Build a preliminary explanatory model of how programme processes might lead to programme outcomes.

2. Search for evidence on these various linkages speculated upon in the initial model in the expectation that one will encounter a wide range of studies, research types, and data forms that might contribute to an understanding of the programme workings. Anticipate a many-to-one relationship between the studies and the tyro explanation; each study is likely to contribute only a link to a developing chain.

3. Identify the vital evidence from these studies remembering that the key datum is the original authors' interpretations and explanations of their findings and not just the 'results' from the outcome measures, process observations, focus groups, document analysis, personal interviews and so forth.

4. Make a preliminary inspection of each study of how its explanatory claims are substantiated by the evidence, remembering that the claims have to be judged by the lights of the paradigm within which they are created. Such an overview will provide an initial mapping of explanatory claims surrounding the programme: their quality will be expected to vary, with some of them being deemed more plausible than others.

5. Synthesise the evidence by discerning the pattern behind the explanatory claims. Identify processes in study A that lend weight to outcomes demonstrated in study B. Show how the behaviour of actors in study B explains the reaction of stakeholders in study C, and so on. Combine the evidence from studies A, B and C in order to produce an explanatory whole that is greater than the sum of their parts.

6. The process of absorbing the various explanatory claims of each study into a provisional synthesis involves making a judgement on research quality. This may involve one of three upshots: provisionally accepting some studies and their interpretations, provisionally rejecting certain findings and their interpretations, or, most probably, partially accepting and thus provisionally selecting from the findings and their interpretations. Each decision must be made and justified individually, in the light of the contribution the findings make to the overall pattern. The worth of studies is determined in the process of achieving synthesis.

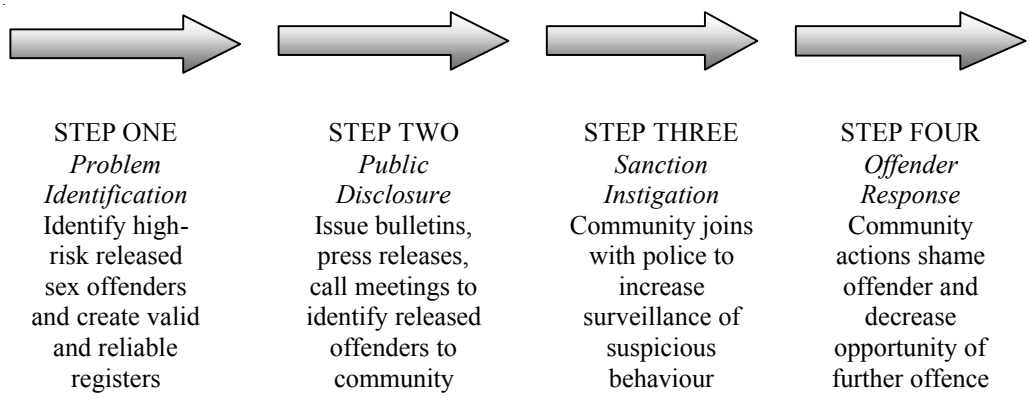
7. Repeat the process of explanatory revision, and thus the perpetual process of assessing the quality of evidence, in the light of new information from a wider range of original studies. New studies may lend weight to or lead to finer adjustments in the explanation under development. If the latter, some earlier assessments of research quality may have to be revised.

IV. Megan's Law: Building the theory, assessing the evidence

In this final Part, my aim is to provide a miniature demonstration of the 'quality-appraisal-is-evidence-synthesis' thesis. In particular, I want to put flesh on the bones of the relentless abstractions of the previous discussion by way of a worked example. The strategy arrived at in Figure three was constructed from first principles but, as ever, the proof of the methodological pudding lies in the practice. Here, I want to exemplify the above model by way an examination of the quality assessment issues that cropped up in a prototype review I have recently completed on the efficacy of Megan's Law (Pawson, 2002). Systematic reviews, of course, often call on hundreds of studies and, perhaps, thousands of fragments of evidence and, clearly, I have no space to reproduce or even summarise the full exercise here. Instead, I want to call upon the contribution of just three studies to that review, for they raise the full gamut of prescriptions and proscriptions on quality assessment argued over the previous sections.

Megan's Law is a somewhat controversial US public policy programme based on the notion of reducing repeat sex offences using interventions designed to notify communities of the presence of released sex offenders living in their midst. The programme theory is actually rather complex and is summarised in Figure four:

Figure Four: The intended process of Megan's Law



review sought to test each of these constituent theories and this involved the perusal of a great many studies, varying markedly in their research strategies and, needless to say, in research quality. Let me now examine the contribution of three aforementioned studies to the review. In each case I offer a brief review of methods and findings, followed by a reflection on the quality of the research.

I. A quasi-experimental study. The obvious starting point of a review in this domain is the evidence on re-offence. There are only two studies that have attempted to track the effect of the introduction of Megan's Law on the rate of repeat offences. Both come to similar and disappointingly inconclusive results and I concentrate here on the quasi-experimental study by Schram and Milloy (1995). Clearly, this intervention is one in which the random application of subjects to experimental and control groups is impracticable and ethically dubious. Once Megan's Law is adopted, it is impossible to sample a group of high-risk offenders about to be released and subject some of them to community notification and others to an unpublicised control condition.

What the researchers actually apply, therefore, is a quasi-experimental design, which compares the recidivism rates of members of the first group of sex offenders released under Washington's notification regime (1990-1993) to those of a 'matched' sample selected from offenders released prior to the enactment of the new law. The matching was performed by ensuring that each group had the same overall spread of single and multiple offences and the same array of victim types. Recidivism rates were calculated by tracking each offender from release, the key comparison being performed in terms of the percentage re-arrested from each group for a sexual offence within four and a half years of release. The headline results from the study are as follows (Schram and Milloy 1995: 3):

- At the end of the 54 months at risk in the community, the notification group had a slightly lower estimate rate of sexual recidivism (19%) than the comparison group (22%). Given the small numbers involved, this difference was not found to be statistically significant.

- Although there were no significant differences in overall levels of recidivism, the timing of re-arrest was significantly different for the notification and comparison groups. Ex-convicts subjected to community notification were arrested for new sex crimes much more quickly than those who were released without notification.

And what of the issue of methodological quality? Quasi-experimentation is criticised in Cochrane and Campbellian circles precisely because it is 'quasi'. It lurks somewhat down the pecking order in the conventional hierarchies of evidence because allocation to the comparison groups is not random. The argument, in this instance, is that the null result might be due to some unrecognised difference between the groups that was not picked up in the matching. Although the two groups had demonstrably similar victim profiles, it might be that a balance of other unexplored characteristics washed out the effect of the programme. Certainly, some considerable time elapsed between formation of the groups and that in itself might encourage the insinuation of confounding factors. On the other hand, the subtitle of the report refers to it as 'a study of offender characteristics' and Schram and Milloy go to unusual lengths in describing the composition and estimating the risk potential of the experimental and comparison groups. But the overall conclusion, by the lights conventional quality appraisal, would be that internal validity is not guaranteed by matching, and the study would be reckoned to have marginal utility. Judgement, which as we have seen is allowed to operate around the margins, might be somewhat kinder given the unusual circumstances of the programme.

But, from the perspective of this paper, such assessments completely and utterly miss the boat. The study actually has an entirely different weakness. It is a 'black box' analysis (Pawson and Tilley, 1997: 30) and so provides no information on the different regimes experienced by the two groups, other than the fact that they are pre- and post-notification. Now, the one thing that is quite clear from Figure four is that Megan's Law is not a singular condition or treatment. It is a typically complex process, comprising registration, notification, action and reaction, passing through the hands of prisons, police, probation and the public before being focused back on the offender. Some of its elements may have the capacity to prevent recidivism, others may fail to touch offenders, whilst other may lead to displacement, and still others may to lead offenders to snap. In the same manner the pre-notification group does not experience an 'absence of treatment' but the pre-existing concoction of corrections strategies, which also will have had their strengths and weaknesses. In short, what is lacking in Schram and Milloy's study is any implementation, legislative and contextual information on the programme. Knowledge of the balance of processes operating in any particular application of Megan's Law is crucial to understanding its effects (or in this case, its apparent non-effect and unintended effect).

This criticism, by the way, does not mean that Schram and Milloy's research is fit only for the discard pile. Basically, the mechanics of the legislation only allowed them to conduct a before-and after comparison as a means of evaluating effects, and their efforts should be judged against this limitation. Moreover, it is impossible for any study to furnish the total range of information required to evaluate such a complex programme. The whole point of carrying out a systematic review is that the requisite information can be

retrieved from other studies. And it is only when this is to hand (and explanatory synthesis begins) that we can make a judgement on the verisimilitude of the no-change-in-recidivism-but-quicker-arrest conclusion. So for me, quality appraisal must wait. It is quite easy to pick holes in this particular study, from quite different perspectives, but it is still capable of providing some inferential nuggets yet to be forged into gold standard evidence.

II. A prospective simulation. Next up for consideration is a study by Petrosino and Petrosino (1999), which offers a peek inside the black box. The research takes on the difficult task of trying to estimate the difference Megan's Law makes to the capacity of the public to defend itself against predatory attacks. Community notification was created largely in response to stranger-predatory crimes, which are relatively rare and obviously difficult to predict. This research thus attempts to answer the question 'in what percentage of sex attacks will notification give the victim (or their family or community) a prior chance to observe the threat and thus to avoid or avert it?'

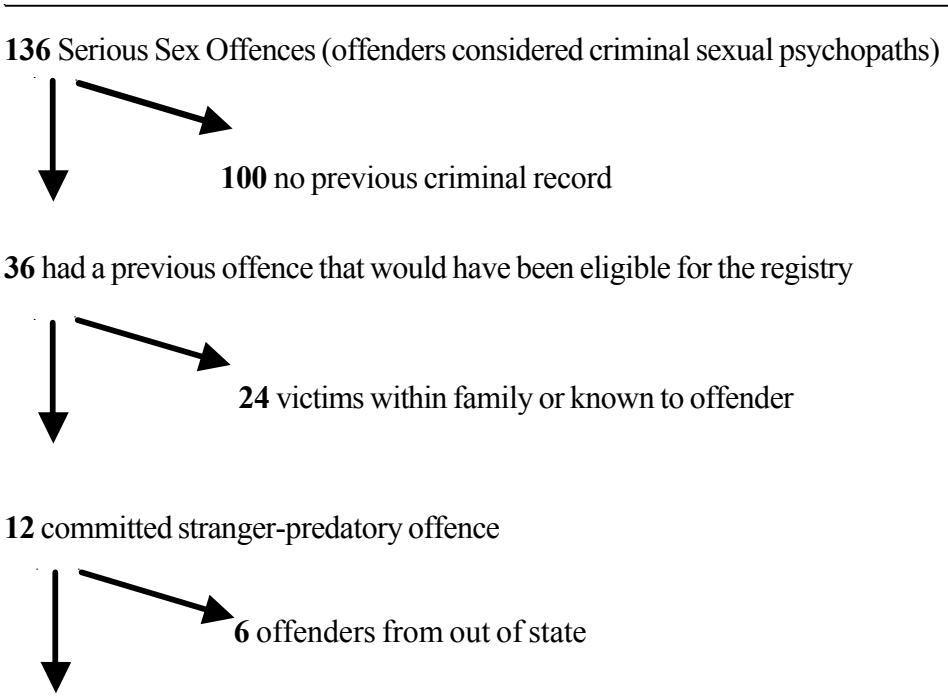
Evaluating 'preventative measures' provides research with one of its toughest tasks, for it amounts to trying to put a figure on what-might-have-happened-but-did-not. The inquiry, moreover, was conducted in Massachusetts, the last of all states to bring Megan's Law to the statute books. The researchers thus had no current registrations upon which to work. Given all these difficulties, Petrosino and Petrosino confronted the task ingeniously by working forwards from actual offences, seeking to discover how many current offenders *would have been* under surveillance, *if* the law had been in place. Their estimate is as follows (1999: 140):

> Using secondary data on 136 criminal sexual psychopaths, the authors found that 27 percent of the sample had a prior conviction that met the requirement of the Massachusetts Registry Law before their most recent sex crime. Of these 36 offenders who would have been eligible for the registry, 12 committed a stranger-predatory offence: 24 offended against family, friends or co-workers.

It is assumed here that first offenders or (more accurately) those without a record are untouched by the notification process. Clearly, Megan's Law cannot work if the offender is not registered. It is supposed, furthermore, that notification has little protective effect on the offender's 'associates', who will in all likelihood already know of the previous convictions. Clearly, these people do not need Megan's Law to pass on information that is all too familiar.

Petrosino and Petrosino's next step was to examine the details of the twelve stranger-predatory offenders (who would in future become registrants) in order to estimate the likelihood of aggressive, proactive, localised warnings getting to potential victims, and the victims being able to defend themselves. In half a dozen of these cases, it was thought very unlikely that the victim could have been forewarned or forearmed by notification because these six offenders were from out-of-state. The simulated notification chain thus ends with six victims who might have had a realistic chance of responding to warnings. It is useful to represent these findings (Figure 5) as a full theories-of-change sequence, for they tell an important tale about the interdependence of the steps of the programme.

Figure Five: The diminishing target of Megan's Law in Massachusetts



And what of the issue of methodological quality? In this case, we enter virgin territory. Although this strategy of 'what if?' analysis has begun to play a part in the evaluation toolbox, I am not aware of any attempts to formalise the approach, let alone match it to a set of quality standards. One can only suppose, once again, that it would be relegated to the list of also-rans under a conventional assessment.

It is quite possible, of course, to erect some post hoc judgements to assess the evidence. The study claims to show that Megan's Law surfaces only a limited proportion of sex offenders from the total population. But does it provide a complete and typical picture of target loss? The fact that six cases remain that could have responded to community notification does not mean, of course, that protection would have followed. Petrosino and Petrosino's closer inspection of the modus operandi of these cases actually backs this up. Two cases involved 'kidnap from a public place' and it may be that such offenders would be reckless enough to behave in the same way *under* the law. Other undiscernibles, however, remain. The fact that Massachusetts was the last state to enact the Law, plus the potential anonymity afforded by urban Boston might have exaggerated the inflow of out-of-state offenders. On the other hand, the sequence above assumes that offenders comply with registration. Other studies show that this is not the case and we do not know how many of the 136 would have escaped from community surveillance by 'dissolving' from the records.

In short, the precise profile of target reduction is an open question and since this is a 'simulation' anyway, the quality of the study remains open to doubt. The significant point, however, is that there are no published standards with which to judge the study and that *ad hoc* judgements must prevail. Given this predicament (and it is a common one) it is clear that such judgements are better made *post hoc* in the light of other evidence and thus *during* syntheses. Thus by my reckoning, this has the makings of an excellent study because it offers an important building block in pattern explanation. Inferences are the key in judging research quality. And if this study is only approximately correct in its slim estimate of the proportion of

potential offenders who actually come under community surveillance, then a great deal of sense is made of the null results from the re-offence studies noted above. The fact that the introduction of the law appears to have little impact on repeat offending may simply be due to the fact that community surveillance is a weak weapon against a rare offence whose perpetrators may still remain well hidden.

III A survey (with open-ended questions). My final case study is chosen because it represents a further selection of methodological strategies and hence a potentially different set of research standards. Zevitz and Farkas's investigation (2000) concentrated on probation and parole personnel in Wisconsin with day-to-day responsibilities for supervising sex offenders. Particular reference was made to their response to Megan's Law, specifically in terms of how Special Bulletin Notification cases (SBNs) added to and modified their caseloads. A survey was the chosen instrument, though perhaps most important for a quality assessment was the use of both fixed-choice and open questions. The report includes much quantitative information on the constraints on staffing levels, training and workloads. However, the authors chose not to rely on 'statistical measures alone' (2000:15) and the bulk of the report uses the standard, qualitative analytic method of the 'illustrative quotation' to register practitioner's viewpoints, and it is this qualitative evidence that is of interest here.

Making valid inference from reported utterances is one of the toughest arts of qualitative research and it is, of course, the critical factor in assessing it. In order to examine this I turn first to Zevitz and Farkas's conclusions (2000: 18).

> Findings indicated that although the law's primary goal of community protection is being served, there is a high cost for corrections in terms of personnel and budgetary resources. Supervision, home visits, collateral contacts with landlords and employers and escort of sex offenders consume a large portion of the agent's work week. Probation/parole agents also bear the onus of locating housing in the community for sex offenders who have undergone extended community notification. This task has proven time consuming and frequently frustrating. Consistent with a containment approach to the management of sex offenders, a major focus of corrections has been to put into place the external controls of enhanced supervision and community surveillance. Agents find themselves heavily involved in community notification meetings for SBN sex offenders. Furthermore, caseloads are high, given the inordinate amount of time required in sex-offence supervision. Yet, despite these heavy demands, agents and unit supervisors were found to be well trained and strongly motivated to do the job. The quality of supervision is high and the public is being well served by these professionals. No better evidence of this can be found in the very low recidivism rates for SBN cases.

Are these inference warranted by the data? This is a tough call and, as foretold by the Cabinet Office study mentioned earlier, considerable judgement has to be brought to bear in making it. My judgement is mixed and is follows. Levels of overwork and the frustration that followed from the 'inordinate amount of time required in sex-offence supervision' are rather vividly captured. Megan's Law, as the practitioners emphasised, is an 'unfunded mandate' and this new obligation is shown, by all manner of methodological means, to pivot the balance of daily activities towards SBN cases. Significantly, this change in case loads follows from the expectation of the programme theory about the 'co-production' of community and law enforcement response. The probation agents bear the brunt of this as follows, 'I don't think management understands the huge number of collateral contacts necessary for a sex offender caseload – family of defendant, victim's family, D.A., clinician, employer and so on' (2000: 18).

Perhaps the most significant comments (in respect of both magnitude and feelings) on caseload refer to the problems of having to manage the community's reactions. Minor harassment is an everyday occurrence, housing problems are shown to be commonplace, death threats occasional. All of this is rather nicely summarised in the sardonic observation of one probation officer: 'there is more pressure to spend greater amounts of time (baby-sit) with SBN cases, simply because they are SBN cases' (2000: 16)

In short, I am bound to say (reviewer's judgement coming up) that the authors do a fine job in demonstrating the practitioner's frustration. A notable case in point is their ethnographer's ear for a phrase like 'baby-sitting' being applied to a group like sex offenders. But in more general terms, I have doubts about the study in respect of the appraisal question 'how well can the route to the conclusions be seen?' (Cabinet Office, 2003: 87). For instance, I do not see any specific case being made for 'strong motivation', 'high quality supervision', and a 'well served public' (see above). In particular, I can see no supportive evidence at all in respect of the authors' final flourish about 'low recidivism rates'. The study was simply not equipped to pronounce on this.

So should judgement on the inquiry teeter towards the negative? One reason for the rather probation-parole-friendly conclusions might be the age-old tendency of 'going native'. Zevitz and Farkas acknowledge the vetting and pre-testing of the questions by probation agents and the launch of the instrument during a state-wide corrections conference (2000:12). But, then again, is such a rumination edging from 'judgement' to 'persecution'? Should not the assessor be concentrating on the fact that this is a 'mixed-method' investigation and is not the balance between qualitative and quantitative analysis the key matter on which they should be appraised? But then again, are there any standards published for assessing the quality of pluralist methods? I am entering into some elaborate (and rather mock) prevarications here in order to make a decisive point. It is simply impossible to come to an overall judgement on the quality of a study that forwards propositions by the dozen and supports them in an extemporised, narrative flow of descriptions, quotations, comparisons, paraphrases and pie-charts.

From the erspective of this study, of course, this is not a gloomy conclusion. The merits of Zevitz and Farkas's work should really show up in synthesis. And indeed they do. There is a rather stunning clue in Schram and Milloy's quasi-experiment in Washington that foreshadows the significance of the survey results in Wisconsin. Recall that the former study demonstrated no change in re-offence rates but a considerable quickening in arrests following the introduction of the Law. Schram and Milloy do not have a great deal to say by way of interpretation of this result. One suspects (reviewer's judgement again) that they were wringing their hands in disappointment at a null re-offence result and found a crumb of comfort in the arrest figures. They ponder briefly, 'sex offenders who are subject to a level III notification [may be] watched rather more closely after the law' and suggest that a 'qualitative study of changes in law enforcement and community behaviour' might supply the evidence in this respect (1995: 19).

Of course, we now have such studies. They begin to suggest a promising explanatory pattern about an unintended outcome of the introduction of the intervention. The Wisconsin study suggests that notified offenders leave behind a rather distinctive trail of family bust-ups, housing turmoil and spasmodic employment and, moreover, that it is someone's job to track it. And although these tribulations demonstrate that an element of the community's attention is being caught by notification, the Massachusetts study indicates that this awareness is unlikely to occur at the point of offence. Stranger predatory offences are so rare, and more likely to be committed by the non-notified, that prevention by community surveillance is unlikely. Synthesising the evidence from three flawed studies begins to make a plausible case that Megan's Law enhances detection rather than prevention.

With this triptych of evidence, I do not, of course, claim to have 'reviewed' the workings of Megan's Law. Readers interested in how the evidence accumulates further should examine the complete review. My conclusions are methodological:

1. *It is futile to try and make decontextualised assessments of the worth of entire studies.* All that would produce in studies one, two and three above are mixed verdicts.

2. *Quality judgements should be made at the level of the inference and not the study. Studies that supply multiple inferences on the basis of research strategies of diverse quality may still be useful.* Study three warrants inferences about surveillance but not reoffence.

3. *The worth of an inference is further established by its coherence with potentially justifiable inferences from other studies. Pattern-making is the core method of analysis in systematic review.* Study one signals a letdown on prevention but a compensatory effect in terms of arrest. Study two partially explains the prevention failure. Study three partially explains the arrest success.

4. *Pattern making involves attention to negative cases and thus the possibility of reconsideration of the worth of studies. Analytic induction thus plays a part in assessing study quality.* The sensitivity of the re-offence outcome detected in study one may be challenged on the grounds of its quasi-experimental design. The subsequent discovery of increased surveillance in study three may lead to a more positive assessment of the former study in respect of a slightly different question: namely, its capacity to detect differences between outcomes.

In short, the worth of a study is determined in synthesis.

References

Bryman A (2001) Social Research Methods, Oxford, Oxford University Press.

Davies H, Nutley S and Smith P (2000) What Works? Bristol, Polity Press.

Gubrium J and Holstein J (1997) The New Language of Qualitative Method, New York, Oxford University Press.

Haack S (1993) Evidence and Inquiry, Oxford, Blackwell.

Hammersley M (2002) Systematic or Unsystematic, is that the Question? Some reflections on the science, art and politics of reviewing research evidence. Health Development Agency Paper October 2002 (better reference?)

Harré R (1972) The Philosophies of Science, Oxford, OUP.

Hunt M (1997) How Science Takes Stock, New York, Russell Sage Foundation.

Joint Committee on Standards for Educational Evaluation (1994: 2nd edition) The Program Evaluation Standards, Thousand Oaks, Sage.

Kaplan A (1964) The Conduct of Inquiry, San Francisco, Chandler.

Lakatos I (1970) 'Falsification and the Methodology of Scientific Research Programmes' in I Lakatos and A Musgrave (eds.) Criticism and the Growth of Knowledge, Cambridge, CUP.

Lindesmith A (1947) Opiate Addiction, Bloomington, Principia Press.

Morse J (2001) 'Qualitative Verification' in J Morse, J Swanson and A Kuzel (eds.) The Nature of Qualitative Evidence, Thousand Oaks, Sage.

Newton-Smith W (1981) The Rationality of Science, London, Routledge.

NHS Centre for Reviews and Dissemination (2001) Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guide to those Carrying Out or Commissioning Reviews, Report no 4 (2nd edition), York Publishing Service.

Noblit G and Hare R (1988) Meta-ethnography, Thousand Oaks, Sage.

Oxman A (1995) 'Checklists for review articles' in I Chalmers and D Altman (eds.) Systematic Reviews, London, BMJ Publishing Group.

Pawson R (1989) A Measure for Measures, London, Routledge.

Pawson R (2002) 'Does Megan's Law Work: A theory-driven systematic review' ESRC UK Centre for Evidence Based Policy and Practice, Working Paper No 8 (available at www.evidencenetwork.org).

Pawson R (2003) 'Shifting Standards: The quest for quality appraisal in social care knowledge' ESRC UK Centre for Evidence Based Policy and Practice, Working Paper No ?? forthcoming – (available at www.evidencenetwork.org).

Pawson R and Tilley N (1997) Realistic Evaluation, London, Sage.

Pawson R, Boaz A, Long, A, Grayson, L and Barnes, C (2003 forthcoming) Report to SCIE

Petrosino A and Petrosino C (1999) 'The public safety potential of Megan's Law in Massachusetts: an assessment from a sample of criminal sexual psychopaths' Crime & Delinquency, 45:140-158.

Polanyi M (1966) The Tacit Dimension, New York, Doubleday.

Popper K (1959) The Logic of Scientific Discovery, London, Hutchinson.

Schram D and Milloy C (1995) Community Notification: A study of offender characteristics and recidivism, Washington State Institute for Public Policy: Seattle. 30pp.

Spencer L, Ritchie J, Lewis J and Dillon, L (2003) Assessing Quality in Qualitative Evaluation, Strategy Unit, Cabinet Office.

Waterman H, Tillen D, Dickson R and de Konig K (2001) Action Research: A Systematic Review and Guidance for Assessment National Health Service R&D, Health Technology Assessment Programme.

Williams M (2000) Science and Social Science, London, Routledge.

Wilson, A and Beresford, P (2000) Anti-oppressive practice: emancipation or appropriation? British Journal of Social Work 30: 533-74.

Zevitz R and Farkas M (2000) 'The Impact of Sex-Offender Community Notification on Probation/ Parole in Wisconsin' International Journal of Offender Therapy and Comparative Criminology 44:8-21.