

Round table discussion of the quality of evidence and its assessment
Tuesday May 20th, 10.45

Abstracts

Martyn Hammersley, Open University

A first point to be made is that assessing the quality of research evidence is far from being a straightforward matter at the present time. There are considerable differences in viewpoint, not just between qualitative and quantitative researchers, but also among qualitative researchers, even about whether there can be criteria of assessment, and if so what form these criteria could take. Another area of disagreement is, of course, the sorts of consideration that ought to be taken into account in assessing evidence (for example, whether these should be restricted to, or should even include, representational validity, as against the role of pragmatic, political, ethical, and aesthetic considerations). Moreover, even for those qualitative researchers who, like me, regard representational validity as the key criterion, there are problems with some aspects of the current interest in assessing the quality of research evidence.

There is sometimes a temptation to think that research produces an evidence-base whose quality can be assessed in general terms. This is especially likely in the context of discussions about the role of social science research in serving evidence-informed policymaking and practice. In my view it is important to remember that ‘evidence’ is a functional or relational concept: evidence is always, and only, evidence in relation to specific knowledge claims that are designed to answer particular questions in particular contexts. As a result, judgments of the quality of research findings as evidence in general terms are always abstractions, and therefore potentially misleading. This is because what is relevant and sufficiently good evidence depends, ultimately, on what question is being addressed; and, also, for what audience.

Furthermore, even where the immediate audience is fellow researchers, it is important to distinguish among the various kinds of question that can be addressed, since this has implications for the sorts of evidence required. What evidence is needed to support a descriptive claim about, say, attitudes towards a particular issue among a group of people will differ from what is required to explain their attitudes. And the evidence necessary to assess a theory about the formation or change of attitudes on that issue will be different again; as will the evidence required to support an evaluation of those attitudes.

Another point I would want to make is that even within the specific context of answering a particular research question, there are usually trade-offs between the kinds of evidence that different data collection and analytic methods can provide. The implication of this is that there is no gold standard: neither the randomised controlled trial nor the in-depth interview that ‘gives voice’ to the informant, neither the sample survey nor the participant observation study. And perhaps it is also important to say that the trade-offs involved cannot always be entirely remedied even through the combination of different techniques. Triangulation is a useful metaphor, but it is a metaphor: assessing the implications of different sorts of evidence is much more complicated and uncertain in outcome than a navigator employing bearings on two different landmarks in order to determine position.

In summary, then, there are some serious obstacles to achieving a consensus about assessing the quality of research evidence at the present time. In particular, fundamental disagreements are involved. Furthermore, while I accept that we can identify guidelines by which to assess evidence coming from different sources, relating to the likelihood of various threats to validity, assessing the quality of evidence must always be in the context of seeking to answer particular questions for particular audiences. And it is always a matter of judgment: there is no set of general rules whose application will do the work of assessment for us, thereby making the process 'transparent'.

Harvey Goldstein, Institute of Education

Conveying uncertainty and contextualising evidence in the face of complexity.

I will use the case of school league tables to reflect on the problems of conveying the idea that complexities inherent in social systems need appropriate modelling tools and a willingness on the part of users to engage with the resulting complexities of interpretation. The complexity arises from two principal sources. One is that social systems are multidimensional whose underlying regularities (if any) can be understood only by models (quantitative or qualitative) whose structures themselves have an appropriate level of complexity. In particular these models need properly to contextualise their data to avoid misleading inferences; for example in the case of performance indicators to adjust for selection effects. The second source arises from the stochastic nature of social systems (real or residual) which also requires to be modelled. In quantitative models this is achieved formally via a probability model whose summary results need to capture the stochastic (uncertainty) element through interval as well as point estimates or, if appropriate, through hypothesis testing. School league tables, and other performance indicators in crime and health for example, are typically expected to satisfy political demands for 'simplicity', 'relevance' and 'influence'. The last requirement, that they influence behaviour or perception, is predicated on the first two. Yet these are problematical and the talk will argue that a major task for social science is to find ways of making this point effectively.

The talk will also look at some general issues concerning contextualisation and the need for replication over time and space.

Ray Pawson, University of Leeds

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

(A full version of this paper is available)

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.

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

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 these 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 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, repeat 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. Methodological standards are thus the long-term, emergent process 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 in the quest for 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. The question under review is tapered down to the single issue of 'does it work?', which privileges the RCT as the only permissible design standards to measure 'effect', which in turn begets the pooling together of a 'net effects' as the appropriate mode of research synthesis.

The paper then inspects recent attempt to broaden the notion of quality standards as appropriate to evidence-based policy. Social programmes are more complex than medical interventions and the evaluative question is more likely to turn on 'why interventions have differential effects for different subjects and circumstances?'. Quite properly, it is assumed that policy formulation along these lines 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 argues they can have no role as an inclusion hurdle in systematic review. They are far too extensive in range and in the subtlety required to apply them to act as preliminary quality filters prior to research synthesis. They are repositories of generalised research wisdom rather than decision points to expedite efficient reviews.

III. On the third issue (when?), the paper bases its positive suggestions for a standards regime appropriate to a plurality of baseline evidence. The starting point is the principle raised in the first section, namely that the veracity of investigation is a matter of the 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 the 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 it provides good explanation and this involves examination of how it jockeys for position amongst competing explanations. Studies are thought to be competent when and 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 integrity of a wide variety of non-positivist social research strategies. Strategies such as 'pattern explanation', 'analytic induction', 'middle-range theory building', 'triangulation' and so on, all stress that 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 ignored 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. Research quality is confirmed only when synthesis is achieved. Investigatory standards cannot be judged in one fell swoop; they are decided in the full sweep of hypothesis generation, analysis, critique and theory-building.

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 qualitative study, and one is prospective simulation of 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 humble technical qualifications, the papers combine to provide a plausible and technically justifiable account of how particular processes unleashed via the public disclosure of sex offenders limits their capacity of Megan's Law to reduce re-offence. Each study strengthens the inferences made by the other. What is exonerated 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. Real research quality comes in inferential nuggets rather than gold standard studies.

Simon Burgess, University of Bristol
Assessing the Quality of Evidence

1. Different types of evidence for different tasks

Economists and others particularly interested in evidence:

- to assess a theory.
- to evaluate a policy.

Therefore necessarily about "causality"

Evidence requirements to determine causality are quite tough.

2. Causality - why problematic?

- Models of human behaviour and interaction
- Equilibrium
- Time series contexts
- Cross section contexts
- Selection

3. Assessing evidence to establish causality

Very difficult, but also very important.

- RCTs?

Yes, but ...

So rely on other things:

- Natural experiments
- Difference-in-difference
- Propensity matching