Evidence-based policy: Where is our theory of evidence?[1]

Nancy Cartwright with Andrew Goldfinch and Jeremy Howick
Professor of Philosophy, London School of Economics, UK, and University of California, San Diego, US

Address for correspondence
Professor Nancy Cartwright
Centre for Philosophy of Natural and Social Science
Lakatos Building
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
UK

Acknowledgements
The support of The Arts and Humanities Research Council (AHRC) is gratefully acknowledged. The work was part of the programme of the AHRC Contingency and Dissent in Science.

About the author

Nancy Cartwright is a Professor of Philosophy at the Department of Philosophy, Logic and Scientific Method at the London School of Economics and Political Science and a Professor of Philosophy at the University of California, San Diego. She is currently the President of the American Philosophical Association (Pacific Division) and was the President of the Philosophy of Science Association in 2008. Her research interests include philosophy and history of science (especially physics and economics), causal inference and objectivity and evidence, especially on evidence-based policy. She is a Fellow of the British Academy and a member of the American Philosophical Society and the American Academy of Arts and Sciences.

[1] Originally published under the same title as Technical Report 07/07 (ISSN 1750-7952 Print, ISSN 1750-7960 Online) by the Contingency And Dissent in Science Project, Centre for Philosophy of Natural and Social Science, The London School of Economics and Political Science, UK. The editors and publishers of the *Journal of Children's Services* gratefully acknowledge the author's permission to reproduce this version of the original paper, which has been updated and edited to fit the Journal's style guidelines.

Abstract

This article critically analyses the concept of evidence in evidence-based-policy, arguing that there is key problem: that there is no existing practicable theory of evidence, one which is philosophically grounded and yet applicable for evidence-based policy. The article critically considers both philosophical accounts of evidence and practical treatments of evidence in evidence-based-policy. It argues that both fail in different ways to provide a theory of evidence that is adequate for evidence-based-policy. The article contributes to the debate about how evidence can and should be used to reduce contingency in science and in policy based on science.

Keywords

Evidence-based policy; philosophy of science; levels of evidence; children's services

The rise of evidence-based policy

In both the UK and the US there is an increasing drive to use evidence to inform, develop and refine policy and practice. This push to improve how research and analysis informs policy and practice is increasingly being felt in a wide range of areas: in addition to evidence-based health and social care, we now hear of evidence-based housing policy, transport policy, education and criminal justice. Since the election of the Labour Government in 1997, the UK has been firmly committed to evidence-based policy as a way of developing social programmes. The UK Government signalled its commitment to evidence-based policy in the 1999 White Paper *Modernising Government*, which calls for the 'better use of evidence and research in policy-making and better focus on policies that will deliver long term goals' and stipulates evidence as a key principle for policy making (Cabinet Office, 1999: 16). A year later, the Cabinet Office's Performance and Innovation Unit (2000) called for a 'fundamental change in culture' in order to place good analysis at the centre of policy-making and recommended that training for new Ministers and senior civil servants 'should emphasise the importance of analysis for evidence-based policy' (p4). In response to this recommendation the UK's National School of Government, which provides training for the civil service, now runs regular courses on analytical skills and evaluation methods, including introductions to, and overviews of, evidence-based policy making.

An example of evidence-based approach to policy making is the UK Sure Start programme. Initiated in 2001, the aim of the programme is to break the cycle of poverty by providing children and families with childcare, health and educational support. The Sure Start programme has been evidence-based from the start, using extensive reviews of research findings on what approaches and early interventions are most likely to work; its execution and continuing evaluation and refinement have also been evidence-based (Hunter, 2003). Another notable example is the UK's National Institute for Clinical Excellence (NICE)[i], which provides regulatory guidelines for the National Health Service (NHS) on particular treatments. These guidelines are based on reviews on the effectiveness and cost-effectiveness of various treatments.

In the US, the Department of Education is actively committed to furthering evidence-based approaches to education policy and practice. The Department's Institute of Education Sciences established the What Works Clearinghouse in 2002 'to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education'[ii]. Furthermore, the Department in 2005 implemented a

recommendation by the Coalition for Evidence-Based Policy[iii] that projects that include a randomised evaluation should have priority in its grant process.

The commitment to evidence-based policy has been matched with funding. In June 2000, the UK Treasury established the Evidence-Based Policy Fund. With a budget of £4 million over two years, the aim of the fund was to support cross-cutting research and links between research institutes, universities, and government. Several government departments have also contributed funding to the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre)[iv]. Based at the Institute for Education, University of London, the EPPI-Centre collects, reviews and organises the results of evidence-based public policy and research in an accessible way for policy-makers and others. The Economic and Social Research Council (ESRC) funded the UK Centre for Evidence Based Policy, which was based at King's College London. The centre co-ordinated a network of research centres dedicated to promoting evidence-based policy and practice by contributing to the development of methods for evaluating and summarising research.

Not only are evidence-based approaches to policy making funded by governments but also some government funding is increasingly being tied to demands for evidence. For example, proposals to expand the Sure Start programme led to a £16 million research project to establish whether the programme was achieving results (see Belsky *et al*, 2007; Melhuish *et al*, 2008). In the US, the so-called No Child Left Behind Act 2001 enshrines in law the principle that federal funds should support educational activities that are based on 'scientifically-based research'. Title I funding is designed to help schools improve the achievement of disadvantaged students. Those schools that receive Title I funding are required by the Act to use effective methods and strategies grounded in scientifically-based research.

In addition to executives, legislatures too are beginning to take a strong interest in evidence-based approaches to policy making. In November 2005, the UK Parliament's Select Committee on Science and Technology agreed to establish an inquiry on 'Scientific Advice, Risk and Evidence: How Government Handles Them'. The inquiry examined the extent to which polices are evidence-based, what mechanisms are in place for the use of evidence, and the way in which guidelines relating to the use of advice are being applied. Issues addressed include 'sources and handling of advice' and 'the relationship between scientific advice and policy development'. Particular questions explored by the Select Committee

include: ‘What mechanisms are in place to ensure that policies are based on available evidence?’; ‘Are departments engaging effectively in horizon scanning activities and how are these influencing policy?’; and ‘Is Government managing scientific advice on cross-departmental issues effectively?’

Although the drive for evidence-based policy is strongest in the UK and US, the movement is picking up elsewhere in Europe. In its 2001 white paper on governance, the European Union acknowledged that,

> *‘Scientific and other experts play an increasingly significant role in preparing and monitoring decisions. From human and animal health to social legislation, the Institutions rely on specialist expertise to anticipate and identify the nature of the problems and uncertainties that the Union faces, to take decisions and to ensure that risks can be explained clearly and simply to the public.’* (Commission of the European Communities, 2001)

So in the UK, the US, and gradually in Europe, at the executive and legislative levels, and pushed by national and international organisations such as the Campbell and Cochrane Collaborations[v], institutions and regulations are increasingly attempting to ensure that evidence is appropriately considered at various levels of decision-making processes.

Evidence: the missing theory

Evidence-based policy is on the rise then, and all to the good we should suppose. *Except* that we do not have a theory of evidence that can be called upon in policy deliberations. We are supposed to base our policies on evidence but how exactly are we to proceed: what is to count as evidence and how shall we use it? My central thesis is that we lack a practicable theory of evidence – one that can be put to use for evidence-based policy. There are three essential ingredients missing. We do not have:

- A reasonable and practicable concept of evidence
- A reasonable and practicable account of what different pieces of evidence say about a hypothesis and with what strength they speak (see Hammersley, 2005 for a discussion of the variety of kinds of questions evidence can speak to)
- A reasonable and practicable account of how to evaluate a hypothesis in the light of all the candidate evidence.

Philosophical accounts of the concept of evidence

What is it in virtue of which a fact is evidence for a hypothesis? Our philosophical accounts fall into two categories. First are accounts based on some features of the probabilistic relations between the evidence and the hypothesis – for example, increase in probability or various functions of likelihoods (see Mayo, 1996 Chapter 3 for an overview of such positions). These are not useful for evidence-based policy. What we need is a concept of evidence that we can use to judge whether some fact should be taken into consideration – whether it should be 'on the table' for consideration. Then we would expect to look at all the evidence on the table to decide on the probability of the proposed policy claim. Concepts of evidence based on facts about probabilities put the cart before the horse. We need a concept that can give guidance about what is relevant to consider in deciding on the probability of the hypothesis not one that requires that we already know significant facts about the probability of the hypothesis on various pieces of evidence.

Second are those accounts that are based on facts about explanation – for example, versions of inference to the best explanation (Lipton, 2004) or explanatory connectedness (Achinstein, 2001). The problem here is the concept of explanation. A good many accounts end up explaining explanation by reference to probability relations between the 'explanans' [the means of making plain] and the 'explanandum' [that which is being made plain]. This simply recreates the previous problem. Also, it seems to me that the concept is too narrow. Suppose for example that we are considering a policy to combat segregation, perhaps making 'diversity training' mandatory in schools. But recall Thomas Schelling's (1978) game-theory model where checkers are moved on a checkerboard so as to avoid any one checker being the only one of its colour in a group. Eventually clumping occurs even though no moves are designed to put checkers in neighbourhoods that are predominately of their own colour. This is an important model to consider in judging the efficacy of the program for diversity training in reducing segregation. But it is far-fetched to see it as explanatorily connected with the claim that the policy will be efficacious.

Besides these problems, our accounts of evidence also tend to be accounts of *genuine evidence*. But we need an account of what makes something *candidate evidence*. I think I can convince you that you have such a concept by pointing out that we are often ready to blame people for failing to report facts that, though they may turn out *not* to be evidence, under some scenarios *could have been*. Mystery stories are rife with examples. In these

cases the aim is to evaluate a retrospective rather than a prospective causal claim but the point is the same.

Hypothesis: John Jones killed Roger Ackroyd. He could have done so by doing A, B and then C. Ah, but he couldn't because in that case he would have had to travel between Binsey and Summertown in 8 minutes and even the fastest car could not do that. But you are Jones's girlfriend and you know he keeps a fast cross-country motorcycle in his garage so he could have gotten there across Port Meadow in time. You are blameworthy if you do not speak up. Yet if it turns out that A, B and C did not occur, the fact that he owns a dirt bike is totally irrelevant to the hypothesis that Jones caused Ackroyd's death. The case would be exactly similar if you were on a commission and did not report some fact you knew that might be relevant to the efficacy of a policy under consideration but in the end turns out not to be.

What we are urged to do in practice

We also have philosophical accounts that provide the second and third components that I claim to be missing from our theory of evidence. The problem with these is the same as with our philosophical accounts of what evidence is: they are generally not very practicable. They are well reasoned and make sense. But they are usually either too abstract or too circular to provide useable advice about how to conduct evidence-based policy. By contrast, there are now available a host of far more usable schemes – *evidence-ranking schemes*. The problem is that these schemes are not well reasoned and sensible; many seem to me to be daft, indeed pernicious. Yet they are being pushed by a number of influential institutions, not the least of which are the UK and US governments.

These schemes provide all three of my 'missing' components in one fell swoop. Kinds of evidence are ranked according to their 'quality'. Then: (1) Evidence is all and only facts of the kind listed in the ranking. (2) All evidence is taken to speak for or against the *truth of an hypothesis* and the strength of its support is in line with its quality: top ranked evidence indicates that the hypothesis is very likely true, and as quality decreases, so does the strength of support for the truth of the hypothesis. (3) In general the recommendations associated with these schemes do not combine evidence at all. Very often the advice is: if you have top grade evidence, go with what that says. The US Department of Education, for instance, which requires evidence of efficacy in order for a school to receive Title 1 support, tells us that RCTs (randomised clinical trials) are needed to establish strong evidence and that 'Two or more typical school settings, including a setting similar to that of your

schools/classrooms' is the quantity of evidence needed.

There are a vast number of similar schemes available. I choose as an example one particularly thoughtful one, SIGN (Scottish Intercollegiate Guideline Network)[vi]. As their own document reports:

> *'SIGN formerly used the levels of evidence developed by the US Agency for Health Care Policy and Research (AHCPR, now the US Agency for Health Research and Quality, AHRQ). However as a number of limitations were becoming apparent in that system, a review was carried out and new levels of evidence and associated grades of recommendation were developed. Following extensive consultation and international peer review, the new grading system was introduced in Autumn 2000.'* (SIGN, 2008: 36)

The SIGN grading system is this:

SIGN (Scottish Intercollegiate Guideline Network) grading system

Levels of evidence:

1++ High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias

1+ Well conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias

1 - Meta-analyses, systematic reviews of RCTs, or RCTs with a high risk of bias

---

2++ High quality systematic reviews of case-control or cohort studies
High quality case-control or cohort studies with a very low risk of confounding, bias, or chance and a high probability that the relationship is causal

2+ Well conducted case control or cohort studies with a low risk of confounding, bias, or chance and a moderate probability that the relationship is causal

2 - Case control or cohort studies with a high risk of confounding, bias, or chance and a significant risk that the relationship is not causal

---

3 Non-analytic studies, eg. case reports, case series

---

4 Expert opinion

The grading scheme goes like this:

> Grades of recommendation:
>
> A At least one meta analysis, systematic review, or RCT rated as 1++, and directly applicable to the target population; or a systematic review of RCTs or a body of evidence consisting principally of studies rated as 1+, directly applicable to the target population, and demonstrating overall consistency of results
>
> B A body of evidence including studies rated as 2++, directly applicable to the target population, and demonstrating overall consistency of results; or extrapolated evidence from studies rated as 1++ or 1+
>
> C A body of evidence including studies rated as 2+, directly applicable to the target population and demonstrating overall consistency of results; or extrapolated evidence from studies rated as 2++
>
> D Evidence level 3 or 4; or extrapolated evidence from studies rated as 2+

Look now at samples of the kind of advice on offer about how to arrive at an overall judgment:

> Statements that one piece of 'level 1++' evidence is sufficient:
>
> GRADE Working Group: 'Once the results of high quality randomized trials are available, few people would argue for continuing to base recommendations on non-randomised studies with discrepant results' (Atkins *et al*, 2004: 2). [GRADE (Grades of Recommendation Assessment, Development and Evaluation) is an international project aimed at developing a methodologically sound system that can be applied across countries and cultures[vii].]
>
> SIGN: The following quote from the SIGN 50 document seems to imply that if there

are RCTs, the other evidence need not be considered: ‘It is also intended to allow more weight to be given to recommendations supported by good quality observational studies where RCTs are not available for practical or ethical reasons.’ (SIGN, 2008: 36)

EBM [Evidence-based medicine]: ‘If the study wasn’t randomized, we’d suggest that you stop reading it and go on to the next article in your search’ (Sackett *et al*, 2000: 108).

Cochrane Collaboration: In personal correspondence with Jeremy Howick [co-author], Julian Higgins of the Cochrane Collaboration replied to the question of whether evidence from RCTs is sufficient, with the following statement: ‘I'm sure there are very many people who subscribe to this view [that RCT evidence is sufficient] (if interpreted as further evidence on the same questions that the RCTs address). Indeed, one might infer this from the fact that the majority of Cochrane reviews include only RCTs. This strongly implies that the authors believe there is no need to look at other evidence (or believe that 'Cochrane' thinks they shouldn't look at other types of evidence). I have much sympathy with this, given the numerous unpredictable and largely poorly understood biases in observational studies.’

In answer to the question of whether a single well-done RCT trumps evidence from any number of observational studies, Julian Higgins states that ‘If the RCT was done well, then I would always claim this is either the right answer or the answer to a different question from the observational studies.’
[The Cochrane Collaboration is dedicated to encouraging RCTs]

So, what’s wrong with that?
Virtually everything.

1. The concept of evidence involved is too restrictive
Hardly anything gets on the table. This is bad for a number of different reasons. To start with, the type of evidence restricts the type of conclusion for which we can have evidence. These schemes are all for judging efficacy claims. But more, concepts in the study have to match exactly with those in the policy claim; especially they must be completely operationalisable and they must be operationalised in the same way. How could Oxford Council have used

evidence like this to decide whether to build a leisure centre in the new housing estate at Blackbird Leys? Certainly not in the way envisaged in the grading schemes.

Candidate evidence is not even in the ballpark.

The advantage of an RCT is that it can *clinch* results. If the ideal conditions for an RCT are met, positive results *deductively imply* that in at least some subpopulations of the experimental population the treatment causes the relevant effect. But other methods have this advantage as well and they are not in the list. These include various econometric modelling techniques, deduction from established theory and experiments in ideal model systems (Cartwright, 2007a: Section I.3).

A host of other methods that can *vouch for* results even if not clinch them are excluded. These include the hypothetico-deductive method when used for confirmation, qualitative comparative analysis, game-theory modelling, ethnographic methods, and so on. Moreover, any 'voucher' can be turned into a 'clincher' by adding some additional premises - premises that may be reasonable to entertain in particular cases. All methods presuppose other assumptions. These ranking schemes seem to presuppose that the background assumptions required by the methods listed are more likely to be true for all cases than those for methods omitted, which is highly implausible.

2. The claims about strength of evidence in the rankings are mistaken

Much is written about the pros and cons of the specific kinds of evidence that appear in these listings – fully randomised trials, partially randomised trials, observational studies, and so forth. I want to concentrate instead on the basic underlying ideas, which I think are way off base. I have already noted that the kinds of evidence permitted are only good for efficacy claims so I shall confine my attention to these, ignoring other policy issues such as claims about side effects (which nevertheless turn out to be an important issue in the example I will use), about implementation, about the effects of moral, cultural and political considerations, about estimates of costs and the like. I shall also concentrate on RCT evidence for concreteness but what I say can be carried over, mutatis mutandis, to other types of evidence that these rankings admit.

Consider: we wish to evaluate a proposal to do A in order to achieve R: say to treat African children who are HIV-infected prophylatically with an inexpensive antibiotic called

'cotrimoxazole' in order to reduce mortality and morbidity from opportunistic disease until they are old enough for retroviral treatment, as in the 2005 UNAIDS and UNICEF call to ensure that prophylaxis with cotrimoxazole reaches 80% of children in need by 2010 (UNAIDS, 2006: 165). An evidence ranking scheme tells us which kinds of evidence speak strongly for or against this proposal, which less strongly. In this case the justification for the policy is an RCT on children in Zambia published in the *Lancet* in 2004, which concluded that the antibiotic reduced mortality in HIV-infected children by more than 40% (Chintu *et al*, 2004: 1870).

What is the underlying logic that shows how a study like this – assuming even that it meets all the ideal requirements – can serve as strong evidence for the efficacy of the policy? As far as I can see the most plausible construction of the underlying justification assumes that actions are justified by principles. We suppose:

(a) There is a certain type of HIV-infected child population, *T*, for which the Zambian RCT establishes 'In T cotrimoxazole reduces average morbidity/mortality'.

(b) The target population – in this case HIV-infected children in resource-poor settings across Africa – is of type T.

(c) So administering cotrimoxazole in the target population will reduce average morbidity/mortality.

That is, we need some way to get from the evidence to the conclusion, and a way that shows how this evidence can speak so strongly for the conclusion. I think the only way it can work is via an intermediate principle. But this won't do since both the way up to the principle and the way back down to the policy are shaky, and for much the same reason: how to specify T. This is now explained in more detail.

As regards moving from principle to policy, what is wrong here is what is generally wrong in supposing you can read off conclusions about single cases from scientifically established principles: almost all principles are defeasible and those that are not (like 'All men are mortal') do not provide very detailed advice. We can all imagine a vast variety of happenings that can defeat the policy efforts even in the face of the principle. One may have the happy idea that if the target population is really of the right kind – kind T, whatever that is – the defeaters will be distributed the same in the target as in the trial population so the conclusion will still obtain. That has its own problems:

- We do not know what T is. This means that the guidelines may be able to

provide sound advice but it is not practicable advice: we do not know how to tell whether we are following it or not.

- Our target population may start out satisfying the characterisation 'T', whatever that is, but our efforts to implement the policy may change the distribution of defeating conditions or the underlying causal structure. This is a common worry about interventions in economics (Lucas, 1976, 1988) but not much discussed in the evidence-ranking and grading schemes.

In relation to moving from RCT to principle, a positive result in an ideal RCT can establish that in at least one subpopulation of the population involved in the trials the treatment causes the relevant effect. It can also establish that the average result in this population is improvement in the effect. The principle says the treatment causes the relevant effect, or produces an average improvement, in any population of type T. How do we get from the first to the second? Laying aside Hume's problem of induction[viii], we suppose that the positive result will hold in any population like the one in the trial. Hence the emphasis on identifying T: 'like' in what respects?

This is obviously not an unfamiliar problem. We do of course pay attention to what constitutes T. For instance, there were earlier RCTs in Cote d'Ivoire involving the treatment of adults with cotrimoxazole (Wiktor *et al*, 1999) These obviously were not good enough because a population of children can be very different from one of adults. Moreover, many African children live in areas with high rates of bacterial resistance. So the RCT that is used to justify the UNAIDS and UNICEF-proposed policy was performed on children and in Zambia where there are high rates of bacterial resistance to cotrimoxazole. But what else might be relevant?

The answer is a tough one. The only way to characterize T that works is – 'populations that have just the same causal structure and the same joint probability distribution across all relevant variables as the population in the study'. (This is clearly not really true. But this is the only characterisation that does not depend on details about what the probability distribution or causal structure are - see Cartwright, 2007b for a fuller discussion.) And this is clearly not a practicable description. We can try to sidestep the problem by insisting that the experimental population be a random sample from the target. How practicable is that, say for our cotrimoxazole policy? Moreover, random sampling procedures require a great deal of knowledge of the relevant structure of the population sampled if they are to be at all reliable.

Not only do we not in general have such knowledge – the guidelines generally do not take this much into account.

My basic point here is much the same as the one I made in discussing clinchers and vouchers. Nothing can count as evidence for anything except relative to a host of auxiliary assumptions; and the strength with which a body of evidence supports a hypothesis can never be higher than the credibility of these auxiliaries. The privileged items that tend to appear in evidence-ranking schemes have built-in methods for assuring that a few of the necessary auxiliaries are met – blinding in RCTs, for example, is good at ensuring that one source of confounding for the results is eliminated. But there are huge gaps left. And there is no reasonable promise that the gaps are in general smaller than with types of evidence that are commonly not even allowed on the table by these schemes.

3. The advice about how to combine evidence is dreadful

Grading schemes do not combine evidence at all – they go with what is on top. But it seems to me to be daft to throw away evidence. Why are we urged to do it? Because we do not have a good theory of exactly why and how different types of evidence are evidence and we do not have a good account of how to make an assessment on the basis of a total body of evidence. Since we lack a prescription for how to do it properly, we are urged not to do it at all. That seems daft too. But I think it is the chief reason that operates. That is why the philosophical task is so important.

Conclusion

We need to develop a practicable theory of evidence, a theory that will work for evidence-based policy. But it had better be a good theory, one that is both sound and usable: that is, a theory that is both practicable and philosophical.

References

Achinstein, P (2001) *The Book of Evidence*, Oxford: Oxford University Press.

Atkins D, Best D, Briss PA *et al* [GRADE Working Group] (2004) Grading quality of evidence and strength of recommendations, BMJ 328 (7454):1490 (19 June), doi:10.1136/bmj.328.7454.1490.

Belsky, J., Barnes, J. & Melhuish, E. (eds) (2007) *The National Evaluation of Sure Start: Does Area-based Early Intervention Work?*, Bristol: The Policy Press.

Cabinet Office (1999) *Modernising Government*, White Paper Cm 4310, London: HMSO.

Cabinet Office Performance and Innovation Unit (2000) *Adding It Up: Improving Analysis & Modelling in Central Government*, London: HMSO.

Cartwright, N. (2007a) *Hunting Causes and Using Them: Approaches in Philosophy and Economics*, Cambridge: Cambridge University Press.

Cartwright, N. (2007b) 'Are RCTs the gold standard?', *BioSocieties* 2, 11-20.

Chintu C, Bhat GJ, Walker AS, *et al* (2004) 'Co-trimoxazole as prophylaxis against opportunistic infections in HIV-infected Zambian children (CHAP): a double-blind randomized placebo-controlled trial.' *Lancet* 364, 1865-71.

Commission of the European Communities (2001) *European Governance: A White Paper*, COM(2001) 428 final, Brussels: Commission of the European Communities.

Hammersley M (2005) Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers' case for research-based policymaking and practice, *Evidence and Policy*, 1 (1) 85-100.

Hunter DJ (2003) 'Evidence-based policy and practice: riding for a fall?', *Journal of the Royal Society of Medicine* 96(4), 194–196.

Lipton, P (2004) *Inference to the Best Explanation*, London: Routledge.

Lucas, RE (1976) Econometric policy evaluation: a critique, in Lucas, RE (ed) (1981) *Studies in Business Cycle Theory*, Oxford: Basil Blackwell.

Lucas, RE (1988) 'On the Mechanics of Economic Development', *Journal of Monetary Economics*, 22, 3-32.

Mayo, D (1996) *Error and the Growth of Experimental Knowledge*, Chicago: University of Chicago Press.

Melhuish E, Belsky J, Leyland A *et al* (2008) 'A quasi-experimental study of effects of fully-established Sure Start Local Programmes on three-year-old children and their families', *The Lancet*, 372, 1641-1647.

Sackett DL, Straus SE, Richardson WS, Rosenberg & Haynes RB (2000) *Evidence-Based Medicine: How to Practice and Teach EBM* (Second Edition), Edinburgh: Churchill Livingstone.

Schelling TC (1978) *Micromotives and Macrobehaviour*, New York: Norton.

SIGN (Scottish Intercollegiate Guidelines Network) (2008) *SIGN 50: A Guideline Developer's Handbook (Revised edition, January 2008)*, Edinburgh; SIGN Executive.

UNAIDS (2006) *2006 Report on the Global AIDS Epidemic: Executive Summary*, / UNAIDS, available on-line at: http://data.unaids.org.

Wiktor SZ, Sassan-Morokro MD, Grant AD *et al* (1999) 'Efficacy of trimethoprim-sulphamethoxazole prophylaxis to decrease morbidity and mortality in HIV-1-infected patients with tuberculosis in Abidjan, Côte d'Ivoire: a randomised controlled trial.' *Lancet*,

May 1, 353 (9163): 1469-75.

[i] www.nice.org.uk/
[ii] www.whatworks.ed.gov/whoweare/overview.html
[iii] http://coalition4evidence.org/wordpress/
[iv] http://eppi.ioe.ac.uk/cms/
[v] www.cochrane.co.uk and www.capmbellcollaboration.org.
[vi] www.sign.ac.uk
[vii] www.gradeworkinggroup.org
[viii] The problem of induction is the philosophical question of whether inductive reasoning – ie. making a series of observations and inferring a new claim based on them – leads to knowledge.