What is Evidence in Evidence Based Policy?

Alessandra Tanesini
Cardiff University
May 2012

‘We philosophers of science are faced then with a hard job. Here as elsewhere in the natural and social sciences, in policy and technology, we can help. But to do so we need somehow to figure out how better to engage with scientific practice and not just with each other’ (Cartwright, 2010).

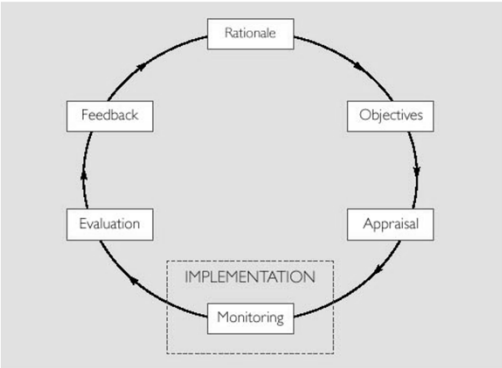
2

What's to come….

- **Main Thesis**: Constraints on what counts as admissible evidence for the effectiveness of policy interventions have several distorting consequences:
  - Measurable proxies are conflated with or selected instead of desired outcomes.
  - Accuracy is sacrificed for Generalisability.
  - Service deliverers professionalism and epistemic authority is undermined.
  - Loss in democratic accountability.
- **Sect 1**: Evidence Based Social Policy in the UK.
- **Sect 2:** What Evidence?
- **Sect 3**: Trust in numbers why worry? And what could help.
- **Sect 4:** Measuring Poverty: A Case study.

3

Sect. 1: The Policy Cycle or ROAMEF Cycle



- The Green Book (HM Treasury, 2003)

4

Rationale:

1. Need and Legitimate for government to intervene. Two possible 1. market failure broadly conceived 2 equity considerations (distributive Justice)

2.Objectives

3. Option Appraisal: consider different possible ways of achieving the objectives. Doing nothing must always be considered. Options are appraised using a cost benefit analysis (rather than cost effectiveness).

4. Implementation which includes monitoring and collection of data.

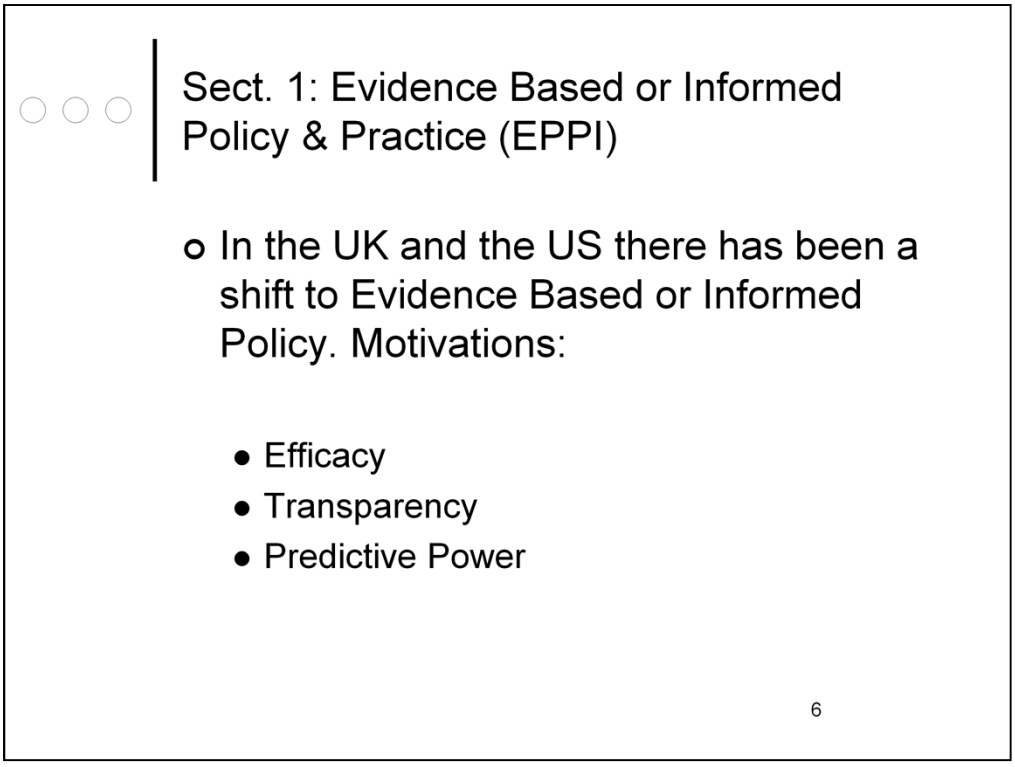
5. Evaluation: There are of three sorts. Process, impact and economic. Process is concerned with how well the policy was implemented. My concern is with impact evaluation which is concerned with how well the policy did vis a vis securing the original objectives. In particular it is concerned with the attribution problem, by establishing the causal contribution of the policy to the outcome.

6. Feedback: learn the lesson form a given intervention to inform future ones.

Sect. 1: The Green and the Magenta Books

- The Green Book (HM Treasury 2003)
  - Emphases that economic principles that should be applied to both appraisal and evaluation
- The Magenta Book (HM Treasury 2011)
  - Provides in-depth guidance on how evaluation should be designed and undertaken.

5



A brief history of evidence based or evidence informed policy in the UK.

Develops on the tail coat of evidence based medicine. EBM develops in the late 1980's driven by the realisation that expert medical judgement was often quite unreliable but also by the democratic concern with making available information about the latest treatments to ordinary practitioners.

Its main justifications in the field of social policy (in medicine, dealing with selection bias was extremely important):

1. Efficacy as measured in terms of cost benefit analyses

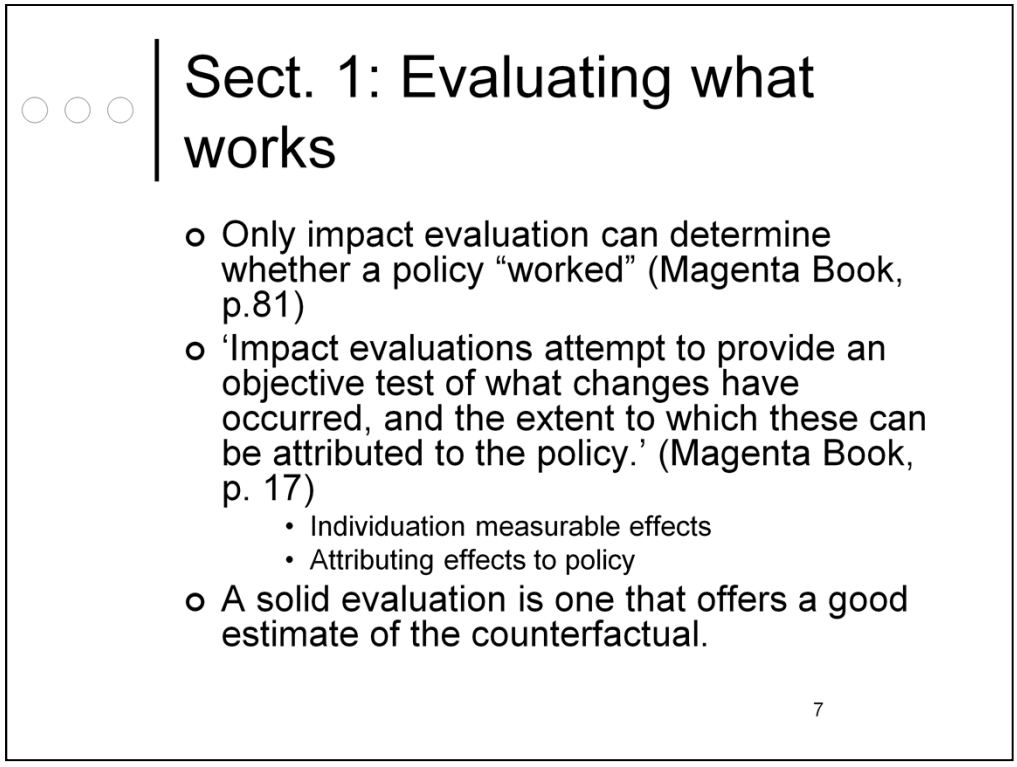
2. Transparency as it replaces expert opinion with evidence which can be analysed by all

3. Predictive power

Social Policy should be based on evidence so as to maximise the chances of effective intervention as measured in cost-benefit analyses

It is intended to replace expert opinion with more transparent and quantitative evidence.

It is meant to be predictive rather than retrospective



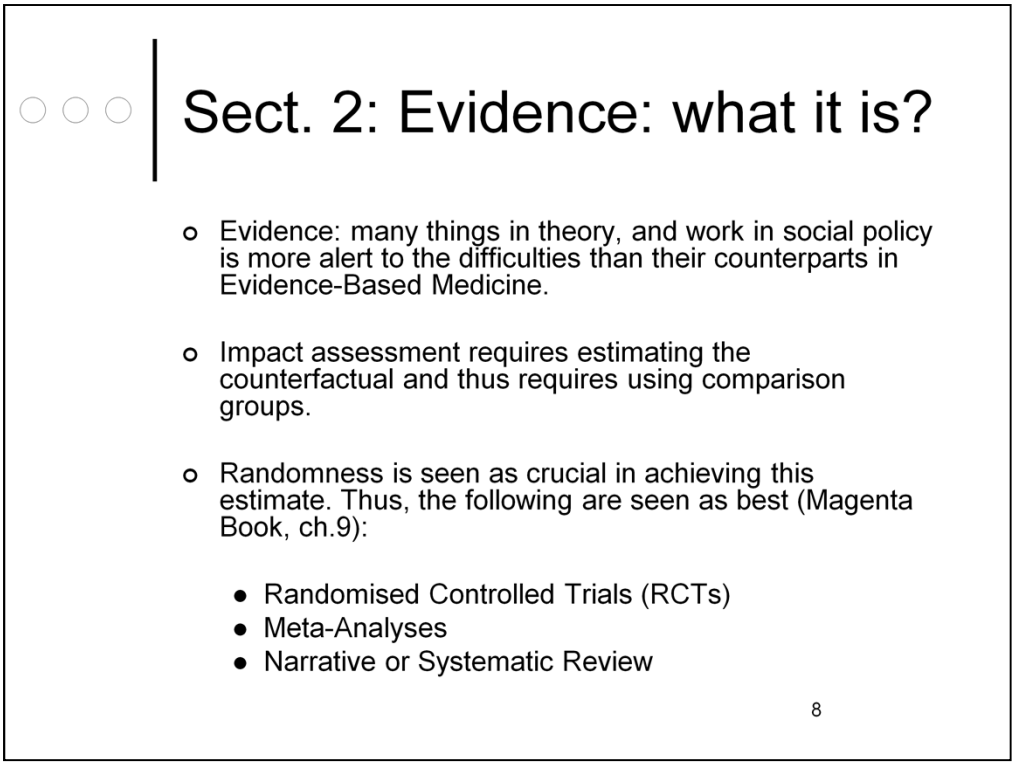
Process evaluations assess whether a policy is being implemented as intended and what, in practice, is felt to be working more or less well, and why. Impact evaluations attempt to provide an objective test of what changes have occurred, and the extent to which these can be attributed to the policy (Magenta Book, p. 17)

So evaluation is concerned exclusively with causal explanations. Presumably because they are taken to have predictive power.

Because of this they take it establishing the counterfactual is essential to conducting a proper evaluation.

Obviously in the social sciences this seems a nearly impossible thing to do.

.



What counts as evidence?

The three staples of EBP

1. Randomized Controlled Trails. They find their origin in medicine. This include allocating ramdomly (and in a way that is doubly blind) members of an homogenous group to one of two subgroups (called the control group and the treatment group) and then treat them in same way with the exception that one group is treated and the other is not.

1. Meta-analyses these are way of working out what works by dividing interventions by kind, then for each study within one group average out all its multiple effects, then do the average of all the studies in each sub-group. Then show this in a table showing the net mean effect of each subgroup. This is the crudest version they can be more sophisticated by making sure that they do not hide the heterogeneity of the multiple outcomes within a study and they also make more effort to put like with like within the same group (e.g., they do not ignore if the study is one in a series say).

3. Narrative Reviews: are meant to be more qualitative and compare large databases of outcomes but include more of a description of the processes by means of which these outcomes are obtained.

For example

From the Centre for Evidence-Based Medicine, Oxford

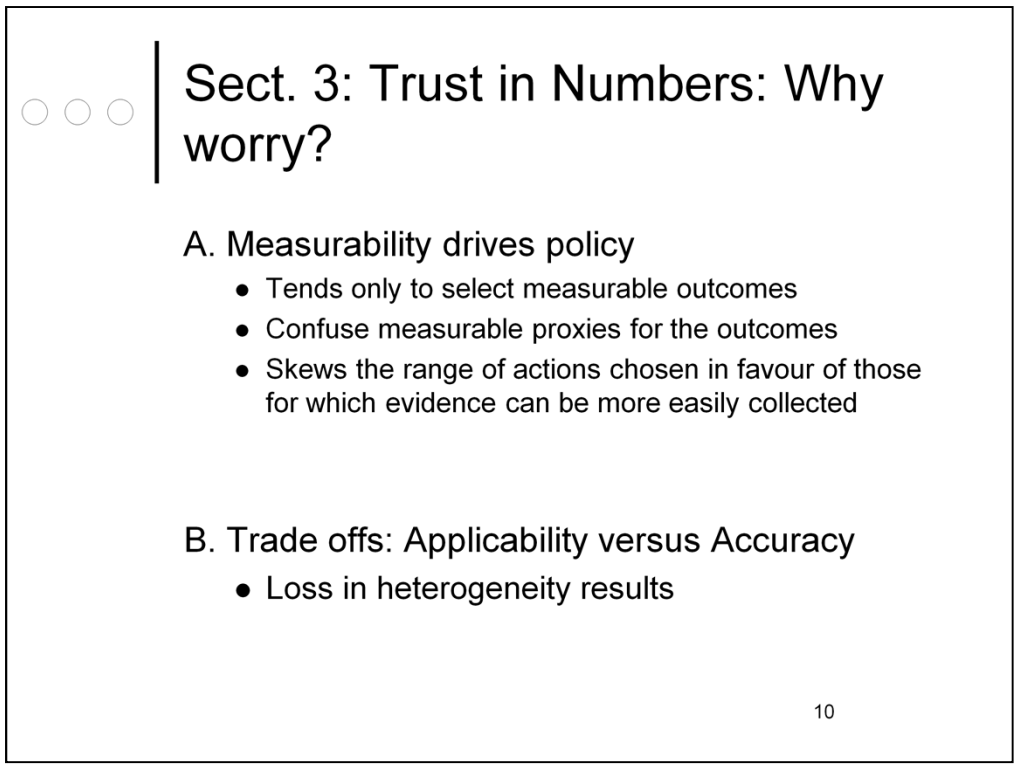
For the most up-to-date levels of evidence, see www.cebm.net/?o=1025

Therapy/Prevention/Etiology/Harm:

1a:Systematic reviews (with homogeneity) of randomized controlled trials1a-:Systematic review of randomized trials displaying worrisome heterogeneity1b:Individual randomized controlled trials (with narrow confidence interval)1b-:Individual randomized controlled trials (with a wide confidence interval)1c:All or none randomized controlled trials2a:Systematic reviews (with homogeneity) of cohort studies2a-:Systematic reviews of cohort studies displaying worrisome heterogeneity2b:Individual cohort study or low quality randomized controlled trials (<80% follow-up)2b-:Individual cohort study or low quality randomized controlled trials (<80% follow-up / wide confidence interval)2c:'Outcomes' Research; ecological studies3a:Systematic review (with homogeneity) of case-control studies3a-:Systematic review of case-control studies with worrisome heterogeneity3b:Individual case-control study4:Case-series (and poor quality cohort and case-control studies)5:Expert opinion without explicit critical appraisal, or based on physiology, bench research or 'first principles'

Sect. 2: Evidence: what it is?

‘An RCT is usually regarded as the strongest possible means of evaluating a policy, because of its ability to balance out the differences between the groups. As was pointed out above, policy allocation by its very nature is not usually random, so opportunities to use it in practice are limited. If the policy is by intention “experimental”, however, then randomised allocation might be more readily acceptable. In these instances the policy will usually begin with a pilot in a restricted number of areas only.’ (Magenta Book, p. 103)



1. As evaluations require questions that are answerable, and tend to occur in parallel with implementation, measurable outcomes are a necessity.

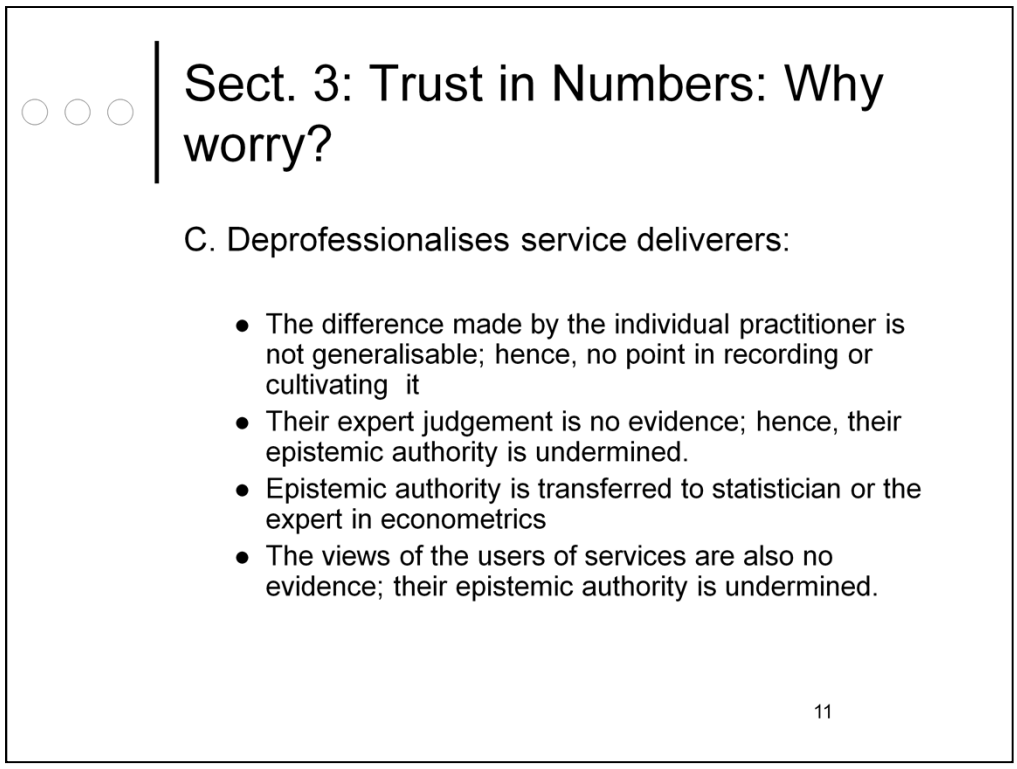
> Even when thought as mere proxies, they sometimes are confused with what they were a proxy for.

2. RCTs and Meta-Analyses especially suppress heterogeneity. Further, they lead to trade off between applicability to a wide range of situations and efficaciousness within a given context.

> Leads to prefer measurement of efficacy that can be standardised over accurate measurements that cannot be compared with results elsewhere. (Porter, 1995) Trades efficacy in a particular situation (accuracy) for some efficacy over a broader range of situations (numerical precision).

Meta-Analyses conceal of programme contexts. The contextual features of a programme (its actors and circumstances) might have been as significant to the outcome as the programme itself and yet this fact is not part of the picture. This is true both for the success and the failure of the programme. Meta-analysis only records and averages such successes and failures but cannot record information about what causes them (the programme, the actors, the programme when applied to these subjects, etc. As a result instead of focusing on clearly targeting programmes which might be appropriate for a specific context, there is an emphasis for preferring programmes that work most widely (that is in a wide range of contexts). Pawson, R. (2002) Evidence-based policy: in search of a method. *Evaluation,* 8**,** 157.

3. The opinion of service deliverers does not count as evidence, the way in which their individual qualities have contributed to the outcome must be discounted as they cannot be replicated.



Might tend to depersonalise the delivery of services by minimising the need for intimate knowledge between user and deliverer (Porter, 2003)

The community might be changed in undesirable ways by the very processes of evidence collection and service delivery. In particular it might undermine trust (Porter, 2003)

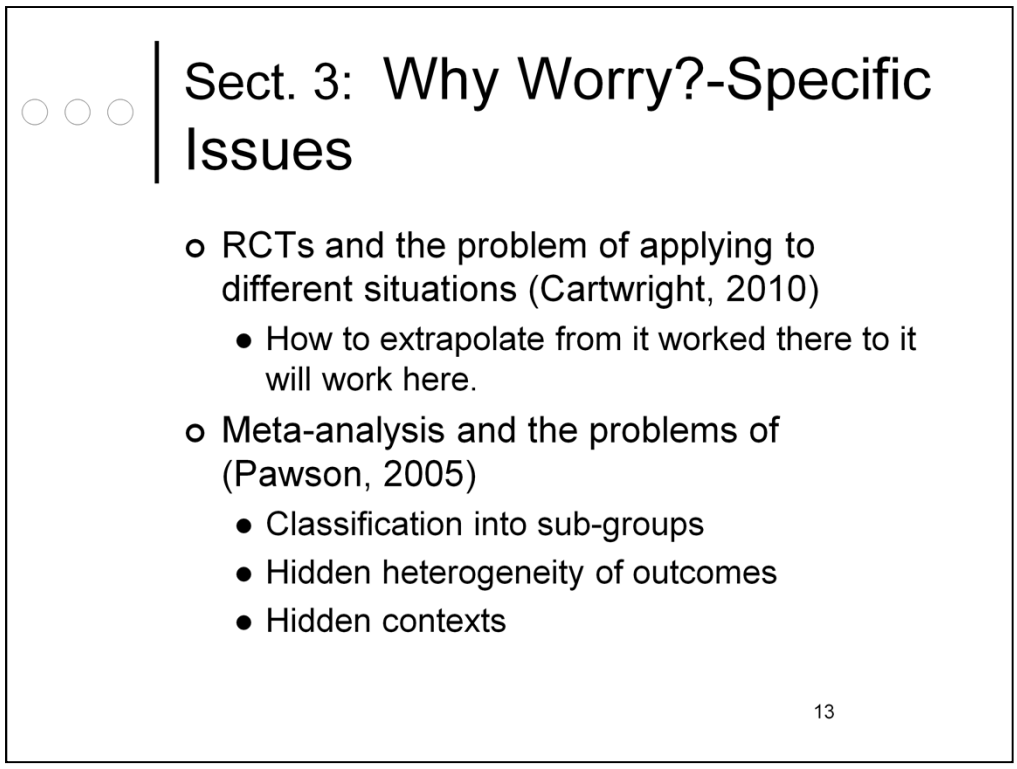
Skews the range of actions chosen in favour of those for which evidence can be more easily collected.- Particularly in evidence in the Magenta Book.

Sect. 3: Trust in Numbers: Why worry?

D. Makes policy decisions less accountable to the public

- Adopts the rhetoric of impartiality and value neutrality for what are value-laden appraisals and evaluations. Hence, prevents full scrutiny of the political will of governments and funders.
- Deprives deliverers and users of ownership of services.
- Might undermine trust

12



1. RCT

A. The problem of external validity

B. The problem of unknown confounders (problem for internal validity also)

2. Numerical Meta-analyses are a way of working out what works by dividing interventions by kind (classification typically done by mode of delivery of a programme e.g., change classroom set up, or change teaching approach) then for each study within one group calculate all of its outcomes (tally) and then average them out and then do the average of all the studies in each sub-group. Finally, offer a comparison of all these results in a table showing the net mean effect of each subgroup. This is the crudest version they can be more sophisticated by making sure that they do not hide the heterogeneity of the multiple outcomes within a study and they also make more effort to put like with like within the same group (e.g., they do not ignore if the study is one in a series say).

He identifies three main problems with meta-analyses

i) the melding of programme mechanisms: this is a criticism of the classification into sub-groups. The worry is that classification by mode of delivery often does not group like with like, because some programmes are set within a context and the mode of delivery has to be understood within it so for instance a programme might actually be misclassified as changing teaching approach when it is changing approach in a newly design classroom, for instance. How the efficacy of the programme is tested will also change depending on how its delivery is conceived but when the programme is grouped with others that are unlike it some of its effects which are actually important in understanding its degree of success are not considered because measuring these is irrelevant to the other studies in the subgroup (so a study where measuring both children's educational achievement and their emotional well-being might get grouped with stuff where measuring the latter does not matter and so the effect of this programme on well-being is not considered in the meta-analysis. So to judge the

programme by appropriate standards can pull in a different direction from the need to use a uniform measurement apparatus.

ii) the oversimplification of programme outcomes. this is a criticism of the averaging out of multiple outcomes. because outcomes are averaged for each one study that has multiple effects and then these averages are themselves averaged within the subgroup, the result is that heterogeneity is hidden. Also the complex relations between the multiple outcomes of a single programme are not studied as the effects are averaged (I take he means that there might say be inverse correlations between some of these oucomes for instance and these facts do not emerge because these relations are not studied). Also different studies within a subgroup might have used different indicators by which to measure change and this also is concealed by the averaging procedure

iii) the concealment of programme contexts. The contextual features of a programme (its actors and circumstances) might have been as significant to the outcome as the programme itself and yet this fact is not part of the picture. This is true both for the success and the failure of the programme. Meta-analysis only records and averages such successes and failures but cannot record information about what causes them (the programme, the actors, the programme when applied to these subjects, etc. As a result instead of focusing on clearly targeting programmes which might be appropriate for a specific context, there is an emphasis for preferring programmes that work most widely (that is in a wide range of contexts).

Sect. 3: How could feminist philosophy of science help?

- By investigating the role of values and of judgments about salience
- EPPI by focusing on evidence (true propositions in support of the policy intervention) ignores the ways in which values guide
  - which studies are conducted
  - the concepts in which the studies are formulated
  - the way studies are classified into sub-groups
  - the conceptualisation of the outcomes

14

Values contribute to determining standards of significance and adequacy to be met by theories.

Values and interests determine whether the knowledge produced by the theory is of practical value. Remember that what makes knowledge a goal of enquiry is its practical import, and whether it has such import (in human action) it will greatly depend on what our values and purposes might be.

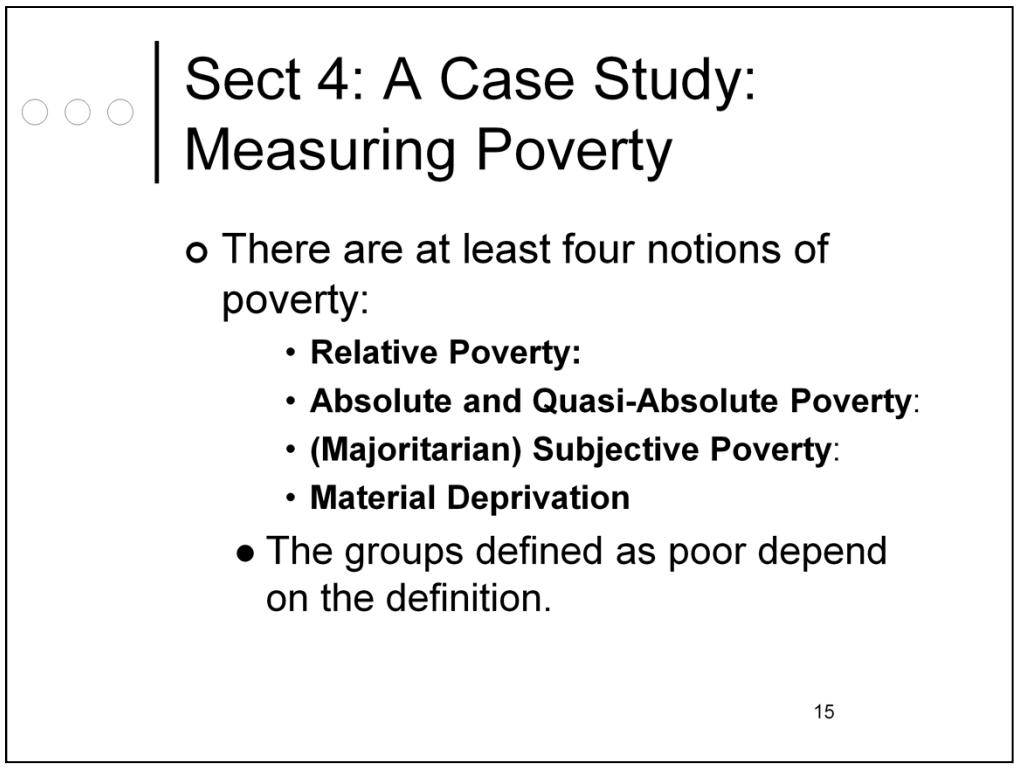
But values also determine the adequacy of a body of knowledge: they help to assess whether it contains enough of the truth so that not to mislead by only listing those items of knowledge that might lead others to make false inferences when contrary evidence is also known. Value can determine whether a theory is significant or true but trivial, they can also determine whether the theory is true but misleading, partial or inadequate by determining whether it contains all the truths that would be relevant.

Values contribute to the classification of objects and phenomena into kinds.

Many scientific theories in medicine and the social sciences especially include so-called thick concepts and other tems that make an implicit reference to human interest and values. In medicine, the term pathogen makes a reference to human helath, and the term health itself is a thick term. These have descriptive conditions of application but their extension is meant to track their normative point. They also have normative consequences.

Values contribute to choices of methods since some possibilities can only explored by the use of some methods rather than others.

The methods adopted will determine with sort of features can be discovered if they exist. So in theories of IQ if one does not even consider what people could achieve in non-oppresive situations (which cannot be measured) but focuses on quantifiable data then certain possibilities cannot show up.



There are at least four notions of poverty:

Relative Poverty: The poverty line in the UK is set at 60% of the equivalised median income.

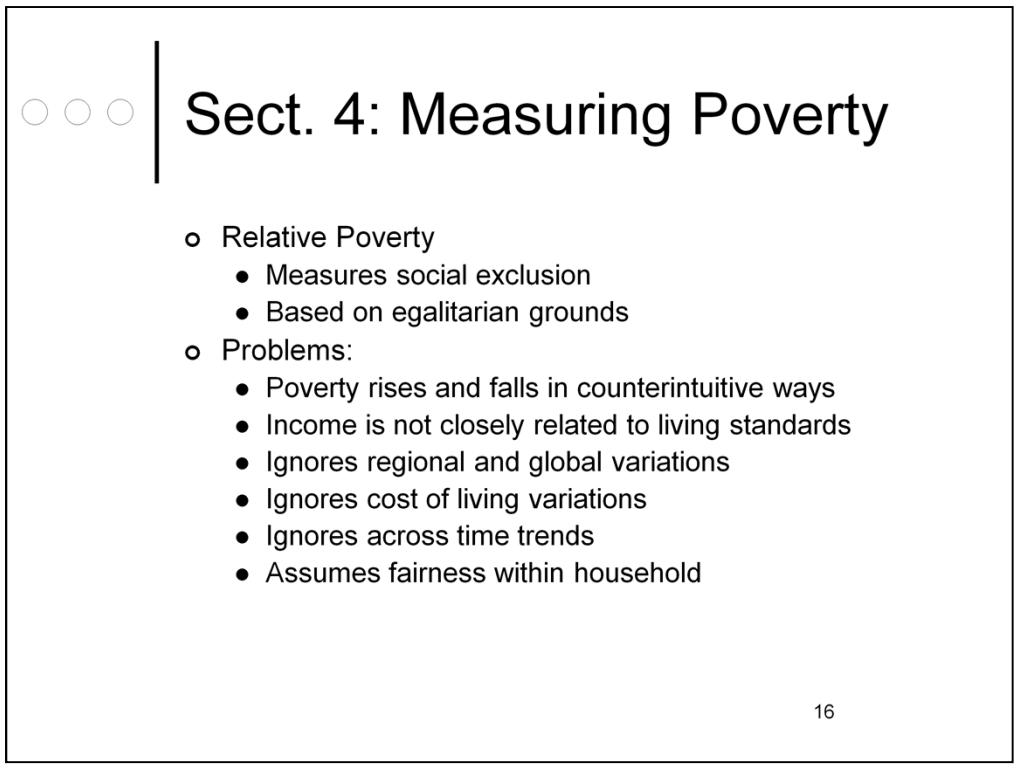
Absolute and Quasi-Absolute Poverty: Setting a minimum living standard (e.g., $1 a day as set by World Bank)

(Majoritarian) Subjective Poverty: The point were most people below it think that they are poor and above it think they are not.

Material Deprivation: Based on involuntary lack of essential goods.

The individuals defined as poor depend on the definition.

In the UK about 22%-25% households are in relative poverty and about the same percentage are materially deprived. But only 50% of either group falls under the other category also.



The UK uses the relative poverty measure and there are prima facia reasons in its favour:

Poverty is a relative business- it does not involve being able to afford what others around one take for granted. So social inclusion is pertinent.

In the Uk the measure chosen is one concerned with the egalitarian distribution of income. It basically says no household income when equivalised should fall below 60% of the median national income.

The definition has counterintuitive consequences

:Poverty rises and falls in counterintuitive ways

During a recession poverty tends to fall as median income fall faster than the lowest 4 10th percentiles

Growth and increase in standard of living is unrelated to poverty if the median income go up faster than the lower ones

Income is not closely related to living standards

Self-employed have volatile incomes (but save for a rainy day); students are poor but have loans; pensioners have assets.

Ignores regional and global variations

Both UK and Italy have large regional variations,Wales could 'address' its poverty problems by seceeding from the UK. In Europe the use of national standards hides the level of material deprivation in the accession countries. It also discourages international action as high poverty figures at home encourage national intervention

Ignores cost of living variations

London is much more expensive than the rest of the UK-to account for this after housing costs figure often used. But ignores higher transportation costs in rural communities, higher living costs faced by disabled people, etc.

Ignores across time trends

Takes a snapshot picture and thus does not capture persistent as opposed to temporary low standards of living

Assumes fairness within household Income is not closely related to living standards

Takes for granted each member of the household has the same fate

Sect 4: A Case Study: Measuring Poverty

- Why was it chosen?
  - Egalitarian considerations
  - Ease of measurement
- Income based measure also used in the US (although it is absolute)

Sect 4: A Case Study: Measuring Poverty

- Material Deprivation (inclusive of social goods)
- Subjective Poverty
- Both rejected because ‘subjective’.