Reducing errors in predicting effectiveness: (Way) beyond statistics

Nancy Cartwright
LSE and UCSD
Kent Conference, June 2009

EBP.

- Evidence-based policy has been the rage for over a decade and there are now a vast number of advice guides, all much of a muchness, teaching how to tell good evidence.
- The guides rank methods for the production of evidence for policy effectiveness.
- In general all the methods ranked are statistical methods.
- More narrowly they are all methods that strive to be as much like RCTs as possible. But this kind of statistical evidence can at best establish **it-works-somewhere** claims and the *somewhere* is never where we aim to implement policy.

- Usual label = ‘external validity’.
- Popular fix from philosophers and statisticians alike = invariance.
- This paper argues that external validity is the wrong way to express the problem and that invariance is a poor strategy for fixing it.
  - Statistical results are invariant under only the narrowest conditions, almost never met.
- What’s useful is to establish not the invariance of the statistical result but the invariance of the contribution the cause produces – a ‘tendency claim’.
  - Tendencies are the conduit by which ‘it-works-somewhere’ claims supply support for ‘it-will-work-for-me’ claims.

I shall argue:

1. we need lots more than statistics to establish tendency claims to begin with;
2. we need much different evidence than statistics provides to make tendency claims relevant to the 'it-will-work-for-us' claims we need to predict the effectiveness of our policies.

Two jobs for RCTs

1. They ensure that we can correctly infer a probabilistic difference in outcome between treatment and control wing on the basis of a frequency difference. Thereby a difference in *effect size* of O with and without T in the experimental population administered as in the experiment – i.e. the *efficacy of T for O, relative to the experimental population and set-up* .This is what statisticians are expert in and I do not belabour it.

1. My topic: They ensure that the treatment, administered as it is in the experiment, causes the outcome in some individuals in the experimental population, X, as well as ensuring the claim 'T causes O in some subpopulation, φ, of X that is causally homogeneous wrt O (relative to T)

Effect size

- When will the mean difference be the same between X and target population θ?
- ANS: When X and θ are the same wrt
  1. The causal laws affecting O
  2. The probability of all 'causally homogeneous' subclasses.
- Otherwise it is an accident of the numbers.

Effect size example

- AJS 114, 144-88. Ludwig et al (from Chicago, Harvard, Brookings,…): 'What Can We Learn about Neighborhood Effects from the Moving to Opportunity Experiment?'
- '*Internal versus External Validity…*MTO defined its eligible sample as…Thus MTO data…are strictly informative only about this population subset – people residing in high-rise public housing in the mid-1990's, who were at least somewhat interested in moving and sufficiently organized to take note of the opportunity and complete an application. The MTO results should only be extrapolated to other ppulations if the other families, their residential envioronmnets, and their motivations for moving are similar to those of the MTO population.'

Effect size example: whence external validity?

Especially problematic given the authors' insistence that RCTs are necessary because we don't know what the causally relevant factors are.

- 'The key problem facing nonexperimental approaches is classic omitted-variable bias'.
- 'A second problem … is our lack of knowledge of which neighborhood characteristics matter…Suppose it is the poverty rate in a person's apartment building, and not in the rest of the census tract…[BUT an experimental] mobility intervention changes an entire bundle of neighborhood characteristics, and the total impact of changing this entire bundle…can be estimated even if the researcher does not know which neighborhood variables matter.

Direction of effect size

- When will an increase in <O> given T in X be sufficient for an increase in θ?
- ANS: If T has same effect on every individual.
- ANS: If X and θ

1. Have the same causal laws
2. *Unanimity:* T acts in the same direction wrt O in all causally homogeneous subpopulations.

- ANS: If θ has 'the right' subpopulations.

Three kinds of causal claim

1. *It-works-somewhere claims:* T causes O somewhere under some conditions (e.g. in study population X, administered by method M).
2. *Tendency claims:* T has a (relatively) stable tendency to promote O.
3. *Effectiveness claims:* T would cause O in population θ administered as it would be administered in θ given policy P (i.e. it will work for us).

1. T causes O somewhere…

This is important information.

See, for example, Curtis Meinert, prominent expert on clinical trial methodology and outspoken opponent of the US NIH diversity act demanding studies of subgroups:

> ‘There is no point in worrying whether a treatment works the same or differently in men and women until it has been shown to work in someone.’

1. T causes O somewhere…

- This is the kind of claim that an RCT can **CLINCH.**
- But what makes it evidence for effectiveness claims: T would cause O in population θ administered as it would be administered in θ given policy P? (T will cause O for us.)
- Standard answer: external validity.
- NC answer: **tendency claims:** T has a (relatively) stable tendency to promote O.

2. T has a stable tendency to promote O

- Examples…
  - Masses have a stable tendency to attract other masses.
  - Aspirins have a relatively stable tendency to relieve headaches.
- We assume this frequently in reasoning about effectiveness; cf. California class-size reduction failure.

2. T has a stable tendency to promote O

- This is (generally) the kind of claim that needs to be established in order for it-works-somewhere claims to count as evidence for effectiveness claims.
- Problem: RCTs – even lots of them – can't establish these.
  - Ultimately **you need a theory**. And most advocates of RCTs like them because no theory is required to do what they do – i.e. establish an 'it-works-somewhere' claim.

External validity v tendncies

- Looking for ‘external validity is thus
  - Epistemically dicey
  - Limited. Recall, causally homogeneous subpopulations must have the same distribution in target and experimental populations
  - Wasteful – IF there are stable tendencies to be had.
- Recall NC: stable tendencies are the best conduit by which it-works-somewhere claims provide evidentiary support for it-will-work-for-us claims.

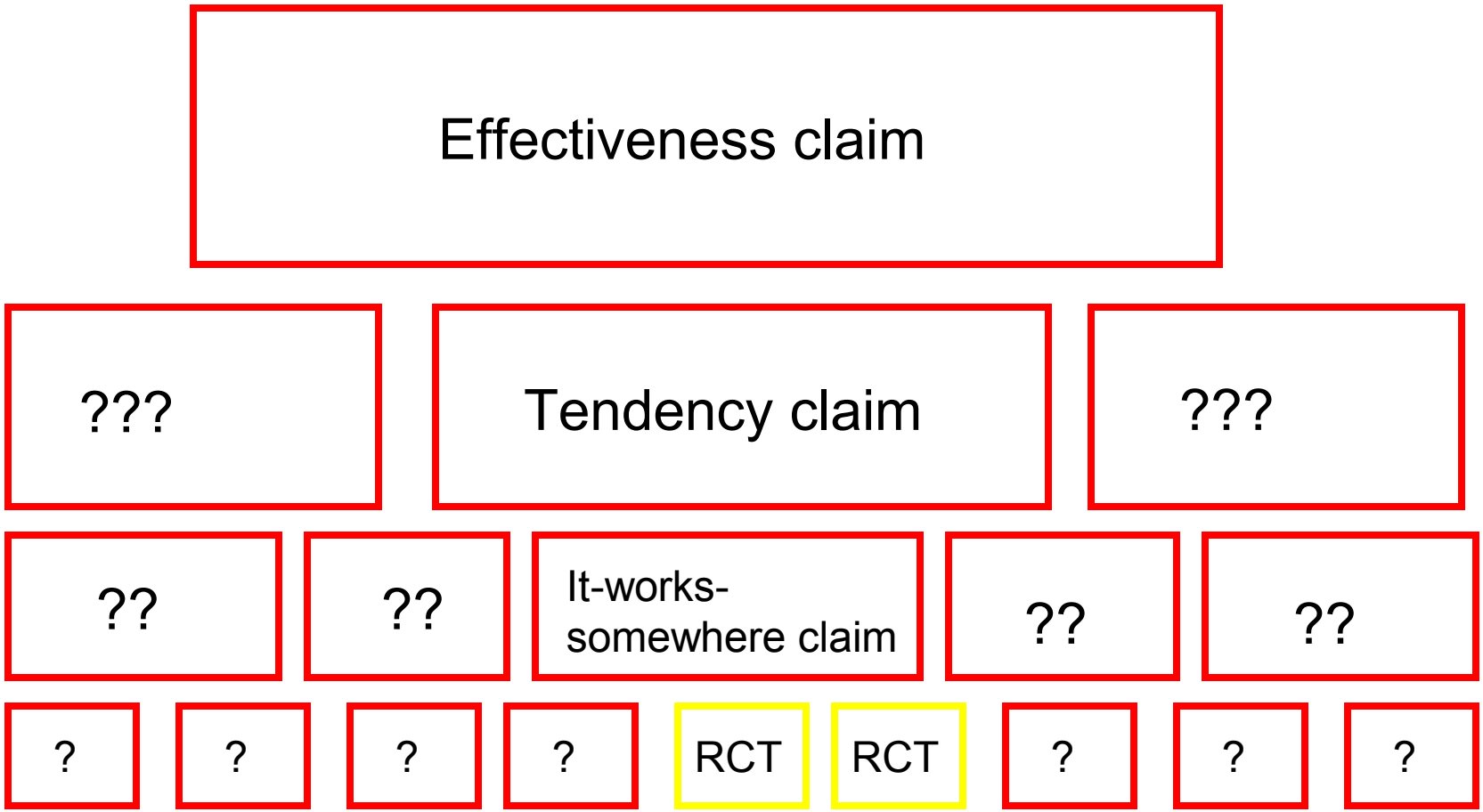
RCTs cannot
hand evidentiary support
directly
to effectiveness claims

Evidentiary support is passed through intermediaries: specifically, through it-works-somewhere claims and tendency claims

Return to stable tendencies

- Many causal relations are local and are not indicative of stable tendencies.
- So it takes a great deal of evidence beyond it-works-somewhere claims to establish tendencies.
- One gold brick cannot make a solid foundation.
- What other evidence is needed?

Hierarchy of evidential support

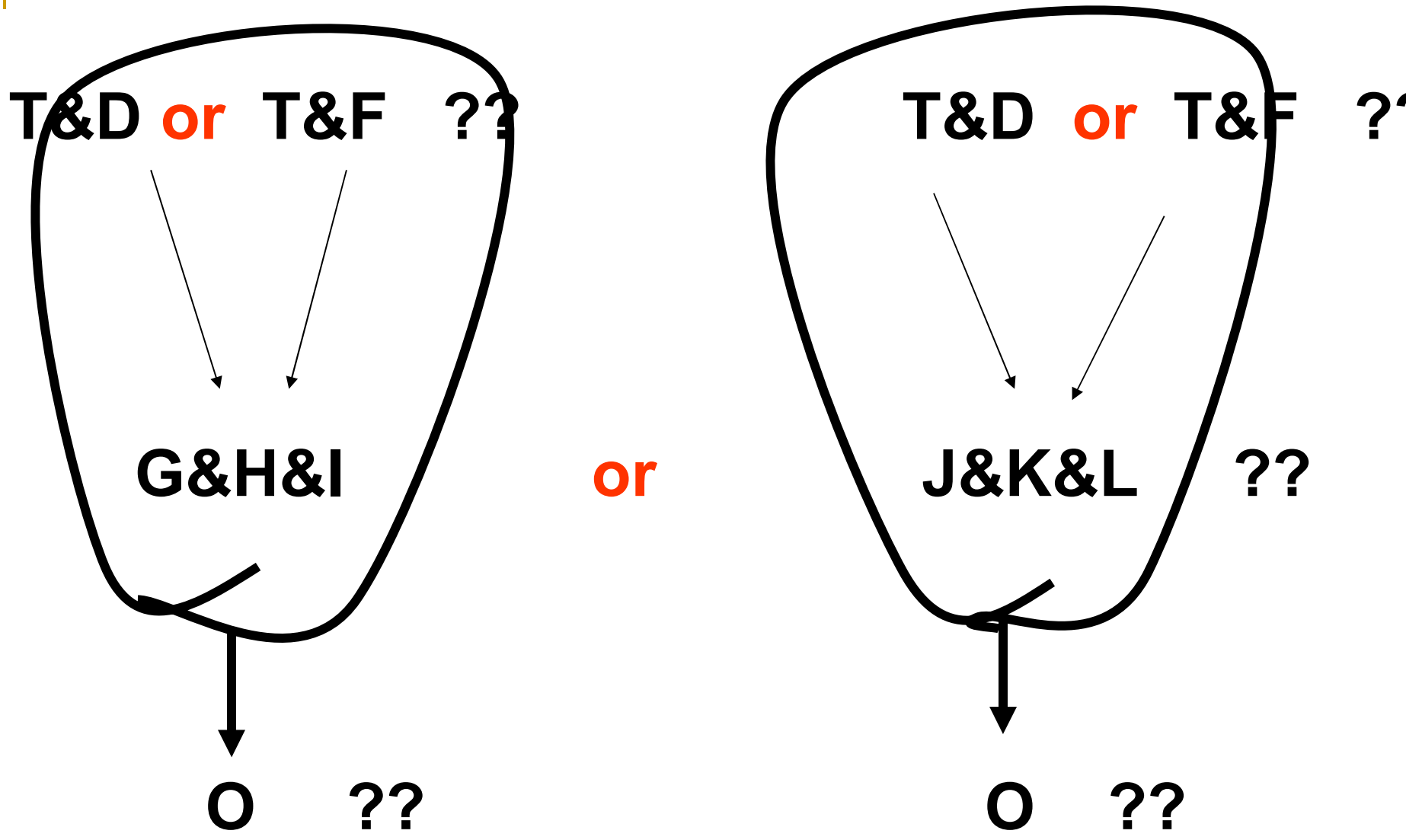


How should we predict effectiveness?

We should

1. **Construct a causal catalogue:**
   - A list of all the causes that can affect O that will be present in our circumstances once we have implemented T (not forgetting the ones we may introduce during implementation).
2. **'Add up' :** Estimate how much O will result when all these factors act together.
1. In adding up, recall **even stable tendencies have necessary helping factors.**

To predict outcomes, we need to figure out ...



Checklist: Questions that need answers – with evidence – to decide 'Will T produce O for you?'

1. Does T promote O somewhere?
2. Is that an accident of local circumstances or does T have a stable tendency to promote O?
3. What helping factors must be there to ensure T acts to promote O?
4. What confounders can act to retard O?
5. Which do you have in your circumstances – or might introduce in implementation?
6. With all that in place, how much of O will you get? Enough?

Conclusion

- It's a long road from
  - It works somewhere

to

  - It will work for me.

- And statistics can't

carry us enough of the way.