# UNDERSTANDING & MISUNDERSTANDING RANDOMIZED CONTROLLED TRIALS

Angus Deaton, Princeton:
Nancy Cartwright, Durham & UCSD

Stockholm October 2016

Durham University

CENTRE FOR HUMANITIES ENGAGING SCIENCE AND SOCIETY

K4U
Knowledge for Use

UC San Diego
DEPARTMENT OF PHILOSOPHY

erc
European Research Council
Established by the European Commission
Supporting top researchers from anywhere in the world

# Outline

1. Doing trials
   - Bias and precision: (perfectly conducted) RCTs are unbiased for ATE, but not necessarily precise
   - Precision is about balance: RCTs do not automatically balance or control anything
   - Standard errors: are much more difficult than you might think
   - Unblinded trials require exclusion restrictions, just like IVE: no automatic superiority

# Outline (2)

2. Using trials
   - Trials are clearly useful sources of evidence, like non-RCTs
   - What are RCT results good for? Sometimes enough in and of themselves
   - Usually need to be integrated into the broader body of experimental and non-experimental knowledge
   - Extrapolation, or simple replication, not the right way to think about this
   - Finding out “what works” unconditionally makes no sense

# 1. DOING RCTS

# What RCTs are good for

- Two groups, treatments and controls, randomly selected from the *study sample*
- Under minimal assumptions, the observable difference in means $\bar{Y}_1 - \bar{Y}_0$ is an *unbiased* estimate of the average treatment effect (ATE), the mean of individual treatment effects
- Individual treatment effects: the value of Y that i would have if treated – the value i would have if not treated
- This ATE might or might not be interesting

# Qualifications (all familiar)

- Study sample might be a random sample of a population of interest; then the ATE is unbiased for the relevant population mean
  - Happens but uncommon
  - More commonly, study sample is selected in some way
- Unbiasedness refers to an expectation taken over repeated randomizations within the study sample
  - We (usually) have only one realization/trial
- Unbiasedness proof requires that the expectation of the mean be the mean of the expectation
  - This is not true for other statistics, such as variance or the median (or other quantiles), which might be substantively interesting and statistically better behaved

# Precision must be earned

- Unbiasedness is good, but we would rather have the ATE be close to the truth (as in small MSE)
- MSE is the sum of variance and squared bias
  - So we would like to be able to trade in some unbiasedness for a reduction in variance
- Cannot lexicographically prefer an RCT just because it is unbiased

# Not always understood?

- JPAL website says that RCTs “are generally considered the most rigorous and, all else equal, produce the most accurate (i.e. unbiased) results”
- Shadish, Cook, and Campbell (2002) write “randomized experiments provide a *precise* answer about whether a treatment worked” and “The randomized experiment is often the preferred method of getting a precise and statistically unbiased estimate of the effects of an intervention”
- Earlier writers, e.g. Cronbach, are at pains to make clear that unbiasedness is close to useless by itself
- Of course, RCTs can be precise, but not inherent in design

# On balance

- Useful to think of a complete all-cause model, supposing causes are INUS conditions and generalising from dichotomous variables

$$Y_i = \beta_i T_i + \sum_{j=1}^{J} \gamma_j x_{ij}$$

- We construct treatment and control groups, not necessarily randomly

$$\overline{Y}^1 - \overline{Y}^0 = \overline{\beta} + \sum_{j=1}^{J} \gamma_j (\overline{x}_j^1 - \overline{x}_j^0)$$

- We get what we want if the groups are *balanced* on the net effect of other causes
- If there were perfect balance (what N calls ‘an *ideal* RCT’), we would have a perfectly precise estimate and know the truth.
- Closer to balance, closer to truth

# Balancing acts

- Laboratory experiment. Other causes excluded manually
- Matching. Choose groups with similar sample averages, e.g. by matching individuals pairwise
  - Cannot match on unobservables
- Randomization matches -- *in expectation*

$$\bar{x}_j^1 - \bar{x}_j^0 \neq 0; \quad \mu_j^1 - \mu_j^0 = 0$$

# Is imbalance problematic?

- Better balance gives more precision in a given trial
  - Balance on average over repeated trials does *nothing* for any one trial
  - Why is repetition relevant?
- Lack of balance does not corrupt confidence intervals
  - Imprecision ought to show up in the standard errors
  - Confidence intervals are potentially correct
  - Wider than they would be with better balance
- Yet people ascribe magical powers to RCTs
  - We suspect that these magical powers contribute a lot to the credibility granted to RCTs by trialists and the public
- Some quotes

# Magical balance

“We can be very confident that our estimated average impact, given as the difference between the outcome under treatment (the mean outcome of the randomly assigned treatment group) and our estimate of the counterfactual (the mean outcome of the randomly assigned comparison group) constitute the true impact of the program, since *by construction we have eliminated all observed and unobserved factors that might otherwise plausibly explain the difference in outcomes*” Gertler et al (2011) ( World Bank implementation manual.) (Italics added).

# More magic

- “complications that are difficult to understand and control represent key reasons *to conduct* experiments, not a point of skepticism. This is because randomization acts as an instrumental variable, balancing unobservables across control and treatment groups.” Al-Ubaydli and List (2013) (italics in the original.)
- “As in medical trials, we isolate the impact of an intervention by randomly assigning subjects to treatments and control groups. This makes it so that all those other factors which could influence the outcome are present in treatment and control, and thus any difference in outcome can be confidently attributed to the intervention” Karlan, Goldberg and Copestake (2009).

# From medicine

“The beauty of the randomized trial is that the researcher does not need to understand all of the factors that influence outcomes. Say that an undiscovered genetic variation makes some people unresponsive to medication. The randomizing process will ensure—or make it highly probable—that the arms of the trial contain equal numbers of subjects with that variation. The result will be a fair test.” (Peter D Kramer, *Ordinarily well*, 2016)

# Bigger samples, better balance?

- Yes, but there can be many causes
- If *J* is large, perhaps they average out, leaving balance
- After all, what we need is balance on the net effect of them all
- Maybe
- But suppose there is only one important cause, it is unknown, unobservable, and unbalanced

# What to do?

- We want to be precise
- Only way to do this is to collect and use baseline information
  - As in an ordinary regression
- Control for it, e.g. by adding covariates in treatment regression
- Or by stratification, or by re-randomizing with stratification
- Bayesians minimizing expected loss will *never* randomize
- This *does not* mean let people choose their own arm
- Match people you know things about, and for those you know nothing about, split them anyway you like
  - Randomization is OK for these units, but it doesn't yield extra precision – supposing one is importing no knowledge
  - Using a random number generator reassures folks about this

# Standard errors

- We need good standard errors!
- One simple case due to Angus
- Treatment effects over individuals have mean zero, and are distributed as a shifted lognormal distribution: asymmetric treatment effects
    - Something like a microfinance experiment
- Although PATE is zero, get significant effects too often
    - Problem persists at quite large sample sizes, though improves

# What is happening here?

- Outliers!
  - From lognormal: estimate of ATE depends on whether the outlier is a treatment or a control
- More generally, skewed treatment effects are problematic
  - Bahadur-Savage (1956) showed that, without some limitations, t-statistics don’t work for means
  - Median treatment effects would be fine, but medians are not recoverable from an RCT
  - RCTs lock us into statistics ill-behaved with skew

# Blinding & post baseline happenings

- Randomization guarantees orthogonality at base line
- Lots can happen after randomization: differential drop out, exposure to different external causes, placebo, Hawthorne, John Henry, Pygmalion effects
- Blinding can help with some of this and is essential: subjects, administrators, analysts
- But not all
- Some can be dealt with by statistical correction or proper monitoring
- But not without lots of other assumptions of just the kind the RCTs are meant to avoid
- Lesson again: Cannot lexicographically prefer an RCT just because it is unbiased

# 2. USING RCTS

# Using the results

- What c T result?
- Knowi ust as import get them
  - A chain of evidence is only as strong as its weakest link
  - A rigorously established result whose use elsewhere is justified by a loose declaration of simile is no stronger than a number pulled out of the air

# Deep parameters & INUS factors

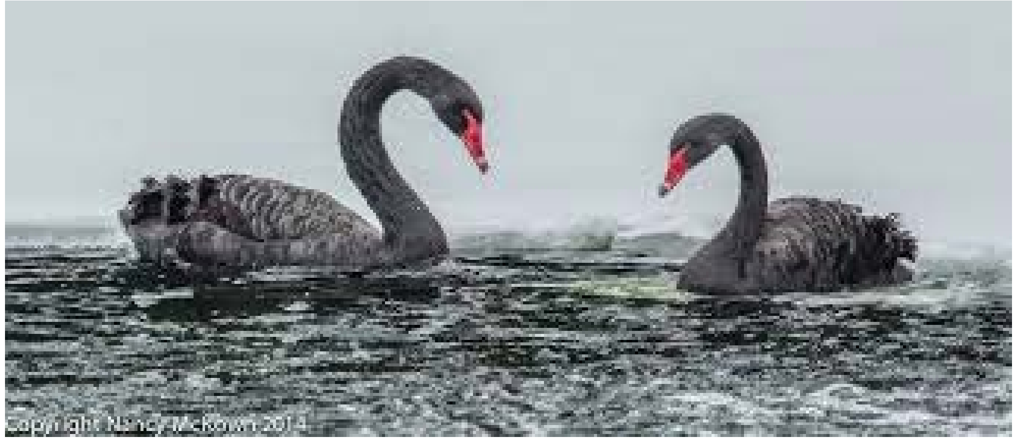
- Sometimes thought that β is a deep parameter and that careful technique will recover it
  - This supposes the cause has, by its very nature, a particular amount of oomph
  - Like the pull of gravity
  - Not likely for the causes we usually test in RCTS
- Recall, INUS causality $Y_i = \beta_i T_i + \sum_{j=1}^{J} \gamma_j x_{ij}$
  - $\overline{\beta}_i$ represents the net effect of all the support/ interactive factors that together are sufficient for a contribution to Y
  - So $\overline{\beta}_i$ depends on the distribution of these in the study population

# ~~External validity~~

- Concept of external validity is unhelpful
- External validity = ‘same’ result holds elsewhere
- Rare. Why should it given
  - Interpretation of β
  - Causal principles themselves depend on underlying structures?
- And if we have a handful of RCTS ‘pointing in the same direction’?
  - NB: very seldom get same ATE estimate and often not same sign!

# Study --> everywhere?

- Beware induction by simple enumeration.
- Swan 1 is white, swan 2 is white,… So



Copyright Nancy McKown 2014

# Have we forgotten Russell’s chicken?

Christmas……

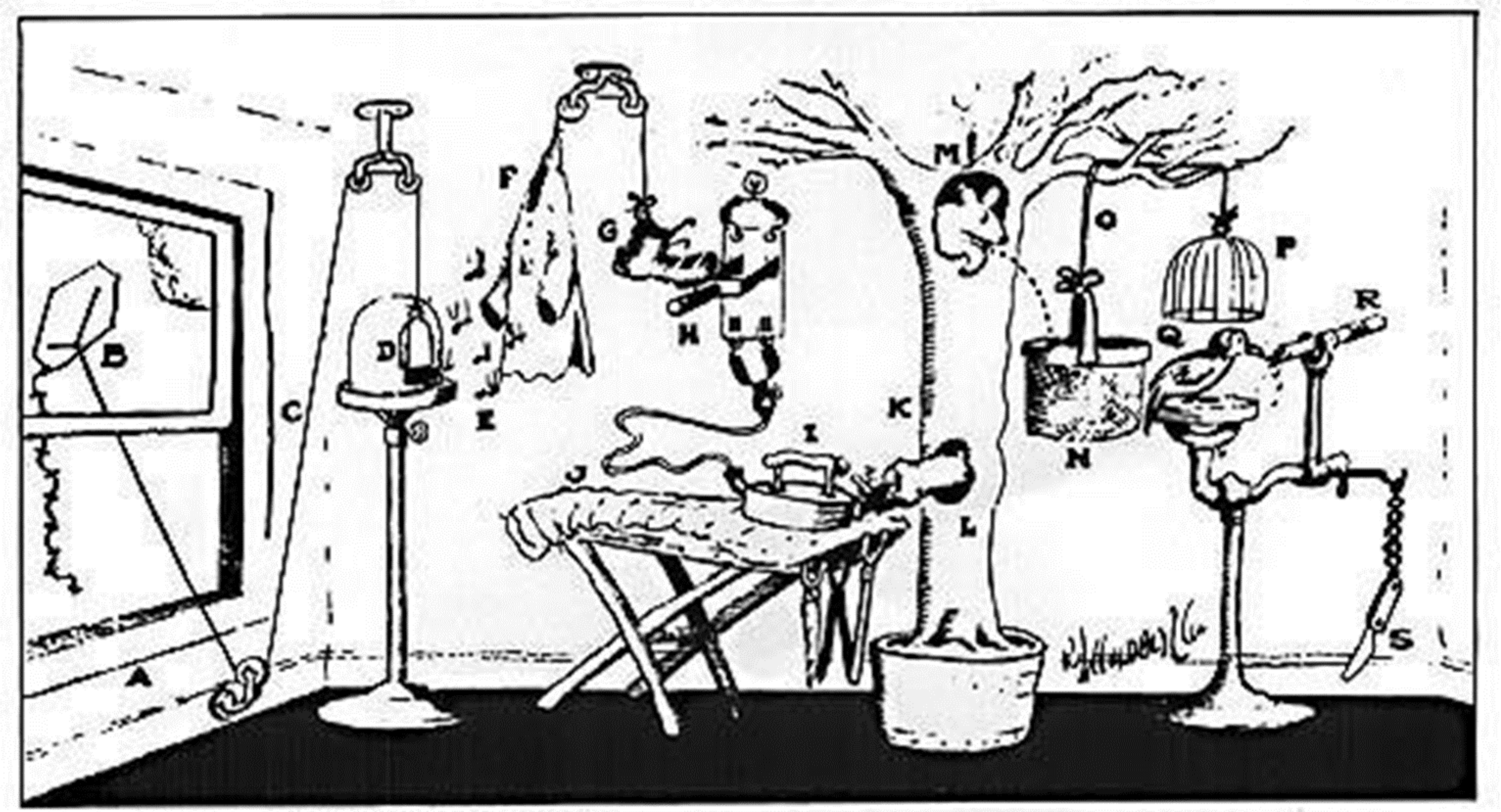
Her problem is not study design.

She doesn’t get the underlying socio-economic structure

# What is “causal” good for?

- Establishing causality does not help with generalization
- Support/interactive factors required, which may be present or absent
- Causality often local. As with the chicken or …

# Underlying structure fixes causal principles



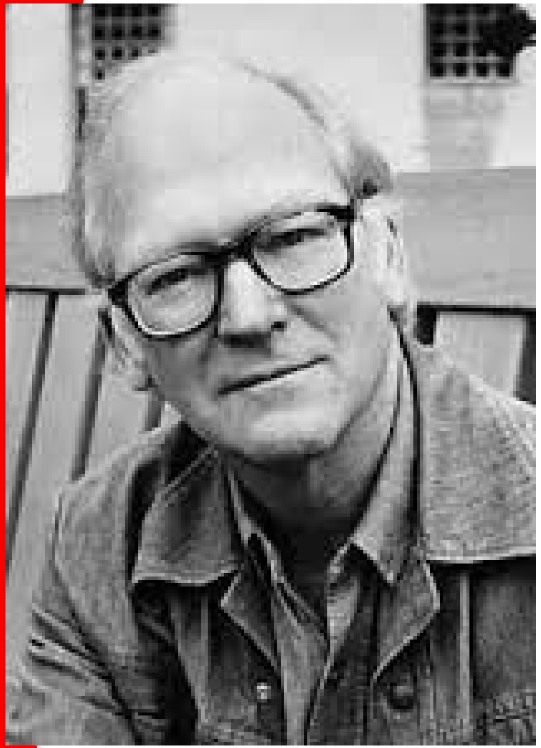
Pencil Sharpener RUBE GOLDBERG (tm) RGI 038

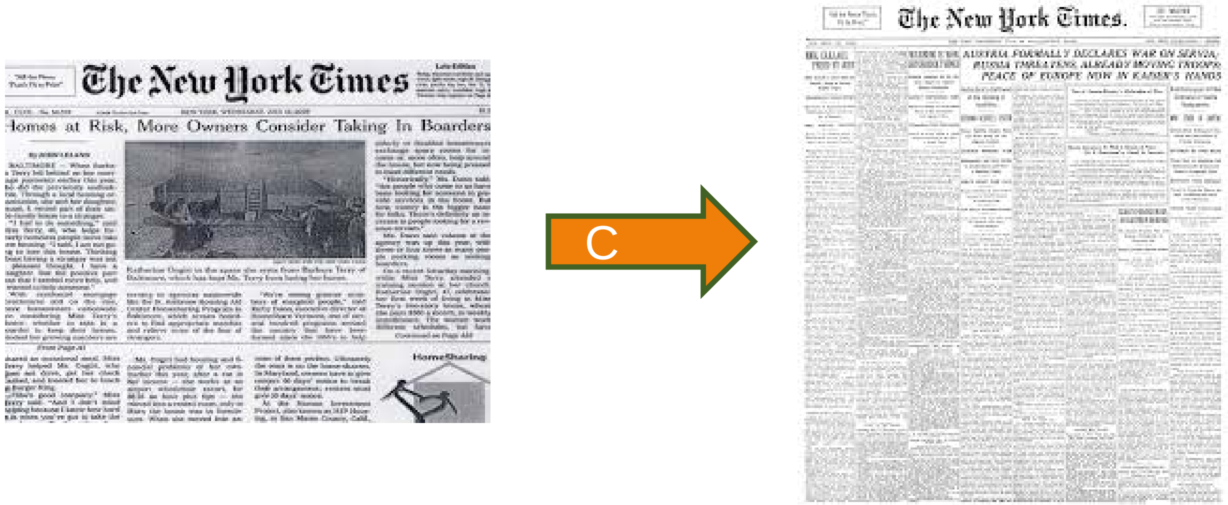
RCT result: Kite flying sharpens pencils

# Compare results of pressing a lever

# Philosophers collude in this mistake

Donald Davidson on concepts that carry causal oomph





The New York Times

Homes at Risk, More Owners Consider Taking In Boarders

The New York Times.

AUSTRIA FORMALLY DECLARES WAR ON SERVIA; RUSSIA THREATENS, ALREADY MOVING TROOPS; PEACE OF EUROPE NOW IN KAISER'S HANDS

# No generalization required

- An RCT that tests a theory
  - Refutation tests
  - Confirming initially unlikely propositions
  - Proof of concept, though concept not always clear
- Evaluation: fiduciary responsibility to funders
  - But evaluation is *not* a global public good
- ATE in a well-defined population is itself the object of interest
  - Public health where target is average health of population
  - “Pragmatic” trials in medicine
  - In economics, investigating revenue effects of a tax or change in welfare policy

# Scaling up?

- GE effects and the like are frequently recognized, but rarely dealt with
- Example: new fertilizer methods increase output for experimental cocoa farmers over controls
  - Scale up, price goes down, farmers worse off
  - Opposite sign: causal effect in the opposite direction
- This should *NOT* be seen as a failure of RCT
  - An opportunity to *use* RCT in a broader context
  - Require observational work and modeling
  - Not a *disadvantage*: just what it takes to do serious work!
- Going to scale generally requires this sort of process

# Drilling down

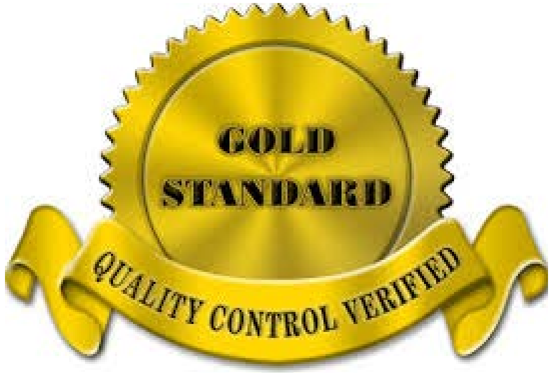
- An RCT gives an ATE that (obviously!) does not necessarily apply to everyone
  - Even those who were in the RCT!
  - "If the patient met the inclusion criteria, then results are applicable" *JAMA* guide to EBM, is nonsense
- Whether we force practitioners to use the mean is controversial and not obvious
  - Mean might be better than prejudice and false knowledge
  - Practitioners may have useful implicit knowledge about individual cases

# Tale of two schools

- New teaching method tested in RCT. It works
  - What should a particular school do?
  - Previous attempt at a neighboring school failed
  - Maybe the "anecdote" is more useful then the "average"
  - Go visit the neighboring school and try to figure out what is going on?
  - Interested in improvement, or even optimization, not just in finding out what works
- US Department of Education tells schools to adopt if RCTs have demonstrated effectiveness in more than one site, in school settings similar to yours.
  - Whatever "similar" means! That is the central issue.

# Conclusions

- RCTs are the ultimate in “credible” estimation of an ATE
- But do nothing to give precision
- And there are difficulties with inference, esp when the target distribution is skewed
- Irony that RCTs deliver means, which are so hard to make inferences about
- Transportation requires all the stuff we have done away with
- RCTs need to be integrated with a network of other knowledge and studies

RCTs ≠ 

# Thank you

The K4U project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 667526 K4U) The above content reflects only the author's view and that the ERC is not responsible for any use that may be made of the information it contains