George Tyler

CAN WE TEST SOCIAL POLICIES EXPERIMENTALLY?

JS Mill once argued that his inductive methods could not be used in economics: Mill's view was that economics was an 'inexact and separate science' (Hausman, 1992), and as such his method of difference could not be used. Economic phenomena were too complex to be described exactly, and so he called for a purely deductive method of formulating economic laws with implicit *ceteris paribus* statements (Hausman, 1981, p. 198). This approach – which strongly influenced traditional consensus on economic policymaking – has been challenged by the use of randomised controlled trials (RCTs), an experiment which probabilistically implements Mill's Method of Difference, to shape and justify social policies. After setting out the method's rebuttal to Mill, I challenge the view of RCT practitioners that legitimate social policies must have their effectiveness verified experimentally. I do so as follows. For advocates, RCTs stand at the top of an evidence hierarchy: due to certain qualities, they are the most 'objective'. However, following Deaton & Cartwright (2018), I assert that there are internal and external validity problems with RCTs: for these reasons, the ordering of the evidence hierarchy qua objectivity collapses. In sum, we can test social policies experimentally, but only with important qualifications and in conjunction with other forms of evidence.

RCTs vs. Mill's argument against social method-of-difference experiment

For Mill, economics cannot be inductive because economic phenomena are complex (Reiss, 2013, pp. 169). Consider his inductive method of difference: comparing two situations that are exactly alike apart from the independent variable of interest. Social situations cannot be compared with reference to any single economic phenomenon: we cannot fully establish the determinants of this phenomenon, and so we cannot isolate a specific determinant to act as the independent variable. Mill's economics is instead an inexact, deductive science (Hausman, 1981), with social phenomena determined by a compound effect of social laws. Economics could not use the inductive method, he argued, because it could not fully describe

all the laws at play in a particular social event. For Hausman, Mill's methods of induction provided a definite proof of a hypothesis already predicted by theory. Since an economic instance was unclear in laws, it could not provide definite results, nor could the results be fully hypothesised before the fact: in sum, the link between theory and observation was incomplete. Instead, economics proposes tendencies: laws that work as long as there are no other laws which affect the result at play.

Mill says that we cannot use his method of difference for causal inference from a social experiment, because we cannot have two situations that differ in one causal factor as the factors are too complexly arranged. This argument as it stands therefore casts doubt on the usefulness of empirical results in justifying economics: we cannot infer causal relationships between variables from empirical data, since the underlying mechanisms cannot be isolated. Structural econometrics takes a different approach: researchers propose a model of a causal interaction and estimate the size of its parameters. The key element is that this model is taken from *deduced* economic theory, which is then tested against empirical data. Causes, in this model, *are not inferred from the data:* the data identifies the extent to which variables are related, but does not identify the underlying causal mechanism, which is posited by deductive theory (Hoover, 2008).

Social experiments, however, assert that causal inference is possible from empirical data. This is achievable through randomisation: by applying a treatment to a randomised subset of a population, all causal factors are evenly distributed over the distribution. In this way, a social experiment is an application of Mill's method of difference. Take the set-up above, and assume that there is a statistically significant difference in outcome over treated and non-treated groups. If we stipulate that the probability of an outcome is fixed by the state of its underlying causes, it follows that the treatment causes an effect for the subpopulation of individuals in the study who happen to have the required arrangement of causal factors: the

treatment works in the *causally homogenous*[1] *subpopulation* of the test sample (Cartwright & Munro, 2010). The stipulation is called causal fixing, and is the standard assumption made to connect difference in probability with actual causation. It is the cornerstone of the counter to Mill's assertion: randomisation and causal fixing can indeed enable us to unravel and identify causal factors.

Using RCTs in social policymaking: the 'What Works' epistemic paradigm

In the last 10 years, the UK government has established institutions in various areas to conduct inductive social experiments for policymaking. The white paper setting out this aim is called 'Test, Learn, Adapt' (Haynes et al., 2012); similar documents and policymaking praxis form an 'epistemic paradigm' (Reiss, 2017) which argues that experimentally testing a policy's effectiveness leads to its legitimacy. There is no unified epistomology for RCTs, but we can identify a kernel of key values to the movement: methodological rigour, precision, unbiasedness, and ability to obtain causal conclusions (Khosrowi, 2019). These values are premised on their role of estimating treatment effects. We can therefore reconstruct the basic argument of the paradigm: policies act as treatments for problems. These treatments should be 'effective', and meeting the above kernel of values satisfies this requirement. RCTs meet the kernel the best, so are the most effective, and therefore should be used to test policies. Considering internal and external validity, I shall now show how the RCTs do not meet these standards, and so discredit the ordering of the evidence hierarchy, invalidating the argument.

RCTs and efficacy: issues of internal validity

The question being answered in this section is: does an ideal RCT discover a true causal relation in its sample? In an ideal trial, it is guaranteed, the main attraction of an RCT. Given perfect randomisation with no influence post-randomisation factors, the other causal factors

[1] (with regard to the outcome)

are the same or equally distributed in the treatment and control group. We can infer from this that the average treatment effect (ATE) is probabilistically dependent on the treatment, since we have eliminated all other confounders (Cartwright, 2009b): we discover a true causal relation. Randomisation equally distributes the confounders, which means that the expected value of the net effect of other causes between the two groups is zero – which means the RCT is *balanced;* this balance leads to the isolation of the mean treatment effect in the experiment. Through a typical single-instance randomisation an ATE is guaranteed to be an *unbiased*[2] estimator of the true ATE in the sample, but crucially it is not necessarily *precise*[3]. It may happen that in any one trial there is an imbalance, or confounding error: an unequal distribution of causal factors over the groups means that the ATE is explained by a covariate as well as the treatment. Randomisation does not automatically balance an experiment, meaning that RCTs alone do not guarantee precise estimates – as precision is only guaranteed if the RCT is balanced (Deaton & Cartwright, 2018). Re-randomising over the same population would solve the problem, allowing us to calculate the standard error of the ATE, but in practice this is infeasible and onerous, especially for a social experiment.

Imprecision can therefore be controlled for by adjusting the ATE with regression analysis. This brings theory into the picture, however, which undermines the 'methodological rigour' aspect of objectivity. For practitioners, a major benefit of RCTs is that they can identify precise treatment effects without theory: if imbalance must always be controlled for using theory-laden methods, we lose the advantage of not needing prior knowledge of possible covariates. However, Senn (2013) asserts that in medical practice the control procedure allows for imbalance due to unobserved covariates. While this may be true in a medical context, it relies on being able to probabilistically model the distribution of all covariates, which is unlikely in the 'inexact science' of economics: in a social RCT it is more difficult to control for covariates, which are more numerous and complex.

[2] (over repetition the average effect is close to the truth and thus not affected by a consistent error)
[3] (the distribution of ATEs over repeated sampling is narrow and hence close to the truth in any one instance)

Another rejoinder is that ideal RCTs could still dominate other methods at producing precise treatment effects. Also, an ideal RCT does indeed lead to a valid causal relation, although this is not unique among econometric methods (Cartwright, 2009a). Despite this, the thrust of the Deaton & Cartwright argument stands; in practice, randomisation does not guarantee an ATE's precision. The consequences of imbalance are poorly understood and conclude that theory is always required in order to interpret a study's ATE. This result therefore brings into question the superiority of RCTs compared to other, more theory-laden methods: indeed, some observational studies have been shown to offer negligible difference in performance in estimating treatment effects (Benson, 2000). This undermines a characteristic value of RCTs, which face problems regarding controlling for precision, contrary to the agreed epistemic paradigm.

RCTs and effectiveness: issues of external validity

While Deaton and Cartwright show that there is overconfidence in the epistemic paradigm regarding the precision of potentially imbalanced single-sample ATEs, RCTs remain a powerful method of uncovering a true causal inference in a sample population. However, generalisation is more relevant for policy and presents a large problem. RCTs do not generalise well and must be combined with other forms of evidence when formulating social policy.

Regardless of the strength of an RCT's identification of efficacy, what matters for social policy is effectiveness: whether the result stands over the whole population. An RCT's results 'clinch' efficacy: done right they deductively establish a true causal relation (Cartwright, 2009a). However, other claims are required for effectiveness, which RCTs do not provide: this is the problem of transportation of results. In order to transport the ATE onto a different context, we must explain the necessary assumption that the background of covariate causes is very similar to that in the sample or adjust the ATE with reference to the general

population. Both require further assumptions and use of theory. In sum, a simple RCT holds no information on its pure generalisability; if we want to extrapolate the result, we must make a case for doing so based on prior knowledge of the underlying mechanisms. For a simple translation, the causal relation uncovered must be a fundamental mechanism, but this is in no way guaranteed by an RCT.

We might like to weaken our claim and say that the treatment has a capacity or *tendency to promote* the desired effect. However, as Cartwright & Munro (2010) point out, method-of-difference techniques cannot achieve this because of Mill's argument against it seen above: due to social complexity we cannot establish a stable tendency inductively from experiment without also using deductive intuition from theory. So again, we can test social policies experimentally only with background knowledge.

There is, then, a lack of information about the similarity of the sample population and the target population, and we have seen that RCTs themselves do not give results that help with this without relying on theory, something practitioners try to avoid. Pearl & Bareinboim (2014) use 'do-calculus' to formalise a model which gives information about which causal effects will hold in a target population. This procedure identifies what information matters to obtain transportability. This is an improvement on simple extrapolation, as it gives a consistent framework for the use of theoretical knowledge, but still requires us to hold some theory-driven assumptions about the similarity of the underlying structure (Cartwright, 2018). These assumptions may hold in evidence-based medicine, but again complexity presents a problem in the social context: the structure may shift over time and space in ways that are much more difficult to state probabilistically.

Another problem which may occur for generalisability in RCTs is the Hawthorne effect: participating in an experiment may change the behaviour of both control and treatment group (Banerjee & Duflo, 2009). Additionally, the act of randomisation may itself lead to

bias in people's behaviour during it (Heckman, 1992). Both of these are serious problems for generalising from RCTs.

Conclusion

We have established that RCTs do not automatically yield precise estimates of efficacy, and even if they did it is very difficult to show that this efficacy translates to policy effectiveness. Hence, background knowledge is always required to interpret RCTs. This suggests that we should use a problem-first approach, rather than the strict hierarchy of the epistemic paradigm: RCTs are not guaranteed to always be the most objective, nor the superior method of obtaining causal conclusions. RCTs are not always unbiased and precise, and in a social context they mandate complex control procedures. Ultimately, RCTs are not specially privileged in studying causation, meaning there is no hierarchy (Worrall, 2002); evidence should be evaluated by the appropriate method. RCTs are certainly a very powerful method of inference but are ultimately another tool: we can evaluate social policies experimentally, but not without reference to other methods of inference.

2496 words

Bibliography

Banerjee, A. V., & Duflo, E. (2009). The Experimental Approach to Development Economics. *Annual Review of Economics*, *1*, 151–178.

Benson, K. (2000). A Comparison of Observational Studies and Randomized, Controlled Trials. *The New England Journal of Medicine*, 9.

Cartwright, N. (2009a). Evidence-based policy: what's to be done about relevance? *Philosophical Studies*, *143*(1), 127–136.

Cartwright, N. (2009b). What are randomised controlled trials good for? *Philosophical Studies*, *147*(1), 59.

Cartwright, N. (2018). What evidence should guidelines take note of? *Journal of Evaluation in Clinical Practice*, *24*(5), 1139–1144.

Cartwright, N., & Munro, E. (2010). The limitations of randomized controlled trials in predicting effectiveness: Limitations of RCTs for predicting effectiveness. *Journal of Evaluation in Clinical Practice*, *16*(2), 260–266.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21.

Hausman, D. M. (1981). John Stuart Mill's Philosophy of Economics. *Philosophy of Science*, *48*(3), 363–385.

Hausman, D. M. (1992). *The inexact and separate science of economics*. Cambridge: CUP.

Haynes, L., Service, O., Goldacre, B., & Torgerson, D. (2012). Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. *SSRN Electronic Journal*.

Heckman, J. J. (1992). 5 Randomization and Social Policy Evaluation. *Evaluating Welfare and Training Programs*, 201.

Hoover, K. D. (2008). Causality in economics and econometrics. *The New Palgrave Dictionary of Economics: Volume 1–8*, 719–728.

Khosrowi, D. (2019). TRADE-OFFS BETWEEN EPISTEMIC AND MORAL VALUES IN EVIDENCE-BASED POLICY. *Economics and Philosophy*, *35*(1), 49–78.

Pearl, J., & Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, *29*(4), 579–595.

Reiss, J. (2013). *Philosophy of economics: a contemporary introduction*. New York: Routledge.

Reiss, J. (2017). On the Causal Wars. In H.-K. Chao & J. Reiss (Eds.), *Philosophy of Science in Practice* (pp. 45–66). Springer.

Senn, S. (2013). Seven myths of randomisation in clinical trials. *Statistics in Medicine*, *32*(9), 1439–1450.

Worrall, J. (2002). What Evidence in Evidence-Based Medicine? *Philosophy of Science*, *69*(S3), S316–S330.