Reassessing Quasi-Experiments: Policy Evaluation, Induction, and SUTVA

Tom Boesche

Abstract

This paper defends the use of quasi-experiments for causal estimation in economics against the widespread objection that quasi-experimental estimates lack external validity. The defence is that quasi-experimental replication of estimates can yield defeasible evidence for external validity. The paper then develops a different objection. The stable unit treatment value assumption (SUTVA), on which quasi-experiments rely, is argued to be implausible due to the influence of social interaction effects on economic outcomes. A more plausible stable marginal unit treatment value assumption (SMUTVA) is proposed, but it is demonstrated to severely limit the usefulness of quasi-experiments for economic policy evaluation.

1 Introduction

The estimation of causal effects is at the heart of any economic evaluation of policy. Although the evaluation of any policy is ultimately a normative question, evaluations will normally depend on which consequences a policy is thought to have.

For example, the public debate around minimum wages is often dominated by controversy about the causal effect of increases in the minimum wage on employment. Traditionally, economists have argued that increases of the minimum wage above the equilibrium level will lead to reductions in employment.[1] This purported causal effect of minimum wages has long been a staple argument of those objecting to minimum wage increases. However, empirical studies have not been able to unequivocally confirm such a negative employment effect.[2] If, in fact, this negative employment effect did not exist, this would weaken the case against minimum wage increases, potentially leading current opponents of such increases to reconsider their position.

The estimation of causal effects is therefore of particular significance to empirical economics and economic policy evaluation. Over the last two decades, the traditional approach to such estimations — theory-based causal modelling — has been challenged by the so-called quasi-experimental approach. Its proponents, namely Joshua Angrist, Guido Imbens, and Jörn-Steffen Pischke, argue that the quasi-experimental approach has led to a 'credibility revolution' in empirical economics because it makes less demanding assumptions.[3]

Notable economists and philosophers, such as Angus Deaton, Nancy Cartwright, and James Heckman, have argued that this quasi-experimental approach cannot be used for most policy evaluation because its results lack external validity.[4] Without external validity, the causal

[1] Compare (Robbins [1932]).

[2] For example, Card and Krueger ([1993], [2000]) find that a modest increase in New Jersey's minimum wage had either a small positive or no employment effect, while Neumark and Wascher ([1995]) reach the opposite conclusion with different data.

[3] See, for instance, (Angrist and Pischke [2009]; [2010]; Imbens [2010], and Imbens and Angrist [1994]).

[4] Compare (Deaton [2010]; Deaton and Cartwright [2017]; Heckman [2008]).

effects estimated by a quasi-experiment could not be extrapolated beyond the observed situation used for the estimation. This would severely limit the usefulness of the quasi-experimental approach for policy evaluation because most policy decisions require knowledge of the consequences of policies when applied to novel populations.

This paper will argue that this objection to the quasi-experimental approach is based on a misconstrual of the approach as purely deductive, but that its usefulness is nonetheless severely restricted by its innate reliance on the stable unit treatment value assumption (SUTVA).[5] It will conclude that the thus limited results of quasi-experiments may still be useful as input for more traditional econometric models.

In section 2, the causal effect of a policy will be defined in counterfactual terms. This definition entails two problems: an epistemological problem and a conceptual problem.

In section 3, the epistemological problem — also known as the fundamental problem of causal inference — will motivate a short introduction to the traditional and quasi-experimental approaches to causal estimation. Then the external validity objection to the quasi-experimental approach will be examined. It will be argued that replication of quasi-experimental estimates and consequent inductive reasoning can be used to endow the results of quasi-experiments with external validity.

Section 4 will introduce the conceptual problem of causal estimation and show how it is commonly dealt with by appeal to SUTVA. The quasi-experimental approach's reliance on SUTVA will be shown to be problematic due to the prevalence of social interaction effects in economic policy evaluations. The stable marginal unit treatment value assumption (SMUTVA) will be proposed as a more plausible alternative. This limited version will be demonstrated to reduce the approach's usefulness considerably. Moreover, it will be argued that inductive reasoning provides no relief against this objection.

Section 5 concludes.

[5] An assumption developed in (Rubin [1978]; [1980]) and first named in (Rubin [1986]).

2 Causal Effects in Econometrics

Both the proponents and critics of the quasi-experimental approach either explicitly or implicitly use Rubin's model of causal inference — a definition of causal effects in terms of counterfactual outcomes.[6] Let X_i be a binary random variable denoting whether individual i is subject to the policy of interest or not. If individual i is subject to the policy, then $X_i = 1$, and, otherwise, $X_i = 0$. Then define Y_{i,x_i} to be the outcome variable of individual i when the policy variable $X_i = x_i$. The causal effect of the policy on individual i's outcome is consequently defined as the individual treatment effect:

$$Y_{i,1} - Y_{i,0} \tag{2.1}$$

For any individual i, either $X_i = 1$ or $X_i = 0$ is observed, but not both. Hence, either $Y_{i,1}$ or $Y_{i,0}$ is a counterfactual outcome. The causal effect of the policy on i is thus the difference between the actual outcome and the counterfactual outcome which would have been observed if the policy status of i had been different.

To illustrate, suppose a policymaker's nation has expanded its military a decade ago and the policymaker is now interested in the causal effect of this expansion on the civilian earnings of the additional veterans. She may be interested in this for a variety of policy reasons. For instance, if she thought military service might have had a negative effect on veterans' civilian earnings, she may desire to compensate them for the financial loss their service to her nation has caused them. Alternatively, the policymaker may be interested in the long-run impact which a future expansion would have on aggregate demand in her national economy. Either way, the policymaker thus tries to find the causal effect which military service has had on the civilian earnings of veterans.[7]

[6] The name of the model was coined by Holland ([1986]), as a label for a model developed in (Rubin [1974], [1977], [1978], and [1980]). For work by proponents using this definition, see (Angrist [1991], Imbens and Angrist [1994]; Angrist and Pischke [2009]). For a critic's explicit use of this definition, see (Heckman [2005]; [2008]).

[7] Strictly speaking, the policymaker would only be interested in the causal effect on the additional veterans who just entered due to the expansion, see (Heckman [2008]; Heckman and Urzua [2009]). For the purposes of this exposition, it is assumed that these two effects are identical.

Then, for each veteran, the causal effect, or individual treatment effect, of military service is the difference between what they earned after they served and what they would have earned if they had not served. If a veteran actually earned less than she would have earned without serving, then military service has had a negative causal effect on her civilian earnings.

There are two obvious problem with this definition of the causal effect: an epistemological problem and a conceptual problem. Conventionally, the epistemological problem receives more attention in the literature and will be discussed in the following section. The conceptual problem will be introduced in section 4 and it will motivate the development of this paper's objection to the use of quasi-experiments for causal estimation.

3 The Epistemological Problem

The epistemological problem — also known as the fundamental problem of causal inference — is that, if the causal effect is defined counterfactually, knowledge of the counterfactual outcome $Y_{i,0}$ is necessary to determine the causal effect. But, by definition, this counterfactual outcome cannot be measured.

In other fields, such as medicine, randomized controlled trials (RCT) are standardly employed to address this problem by estimating counterfactual outcomes. RCTs estimate counterfactual outcomes by comparing the average outcome of those who received the treatment $E[Y_{i,1}]$ with the average outcome of those who did not receive the treatment $E[Y_{j,0}]$:

$$E[Y_{i,1}] - E[Y_{j,0}] \tag{3.1}$$

Given large enough sample sizes and that the individuals have been assigned their treatment status randomly, it can then be assumed that the group of the treated $I = \{i : X_i = 1\}$ and untreated group $J = \{j : X_j = 0\}$ are, on average, equal to each other with respect to everything other than their treatment status. Given the stable unit treatment assumption (SUTVA), the average counterfactual outcome $E[Y_{i\in I,0}]$ is then equal to the observed average outcome

$E[Y_{j\in J,0}]$ and, so, from equation 3.1:[8]

$$E[Y_{i,1}] - E[Y_{i,0}]. \tag{3.2}$$

By the mathematical properties of the expectation operator, the above is an unbiased estimator of the so-called mean treatment effect (MTE):

$$E[Y_{i,1} - Y_{i,0}]. \tag{3.3}$$

Although this is not the individual treatment effect, the MTE — the average of the individual treatment effects — is nonetheless an important causal effect in the evaluation of policy.[9] Randomization thus offers an avenue to estimating the average causal effect of a policy.

However, despite some noteworthy exceptions in behavioural and development economics, economists have to mostly rely on non-randomized observational data. This is a problem because, without randomized assignment, the treated $i \in I$ and the untreated $j \in J$ may differ with regard to so-called confounding factors. A confounding factor — or confounder — is a variable C_i which directly influences both the policy status X_i and the outcome variable Y_i. If such a confounder is present, any covariation of X_i and Y_i might be partly or entirely due to C_i — obscuring any causal effect X_i might have on Y_i.

For instance, Jackson et al ([2012]) suggests that the personality traits of those who choose to serve in the military may, on average, be different from the personality traits of those who do not serve. Moreover, these same personality traits are argued by Judge et al ([1999]) to have effects on career success. Therefore, even if military service had no causal effect on civilian earnings, the average civilian earnings $E[Y_{i\in I,0}]$ may have been different from the average earnings $E[Y_{j\in J,0}]$ of non-veterans because veterans and non-veterans differ in personality traits which affect their earnings. $E[Y_{j\in J,0}]$ can hence not be used as an unbiased estimator for $E[Y_{i\in I,0}]$.

Economic policy is commonly assigned according to factors which also have an influence

[8]SUTVA will be discussed in more detail in section 3.

[9]This is accepted by critics such as Deaton ([2010]), but the MTE's usefulness is scrutinized by Heckman ([2008]).

on the outcome variable or people self-assign themselves according to such properties, like in the above example. Confounders are therefore normally present in economic observational data.

To overcome the problem of confounders in their data, econometricians have developed a plethora of statistical methods for estimating counterfactual outcomes. How these methods are best applied has, however, been a matter of controversy.

3.1 The structural approach

Traditionally, economists have approached this epistemological problem by building 'structural models' — that is, systems of equations — which, informed by economic theory, model the present causal mechanisms.[10] Such models yield an estimate $\widehat{Y_{i,0}}$ of the counterfactual outcome $Y_{i,0}$. The individual treatment effect is then estimated by:

$$Y_{i,1} - \widehat{Y_{i,0}}. \tag{3.4}$$

However, this 'structural' approach faces considerable criticism because economic models are unlikely to capture all relevant causal channels as many aspects of individuals' choice behaviour are not captured by economic data.

For example, Judge et al. ([1999]) provide evidence that some personality traits are linked to career success. Although this is not conclusive evidence that these personality traits have a significant causal effect on civilian earnings, it certainly seems plausible to hypothesize that such a causal channel exists.

However, if an economic model were used to estimate the counterfactual civilian earnings of the veteran υ, υ's personality traits would typically not be included because such psychological data is often not collected at large scales. Similarly, there will be many other factors which would have significantly affected υ's counterfactual civilian earnings but are not measured. Therefore, unless these factors can be accurately estimated themselves, they are not

[10]Note that econometric models are not only built to estimate the causal effect of changes in individual policies, but also to facilitate a multidimensional understanding of the causal interactions within a system and the magnitude of these interactions, see (Heckman [2008]). However, such features are not the subject of this paper's discussion.

included in these economic models.

Given such omissions, it seems unlikely that the estimate $\widehat{Y_{i,0}}$ is an accurate reflection of the true counterfactual outcome $Y_{i,0}$. If $\widehat{Y_{i,0}}$ is not an accurate reflection of the true counterfactual outcome $Y_{i,0}$, the estimated causal effect will not measure the true causal effect.

3.2 The quasi-experimental approach

Due to the above difficulties with building explicit models of all relevant causal factors, the proponents of the quasi-experimental approach argue for the use of statistical methods which require less extensive models. These range from instrumental variable estimation over difference-in-differences to regression discontinuity designs and propensity score matching.[11] What these 'quasi-experimental' estimation methods have in common is that they draw inspiration from randomized experiments.

Quasi-experimental methods overcome the problem of confounders in observational data by using mostly informal economic reasoning to discern and compare subsets $I^* \subseteq I$ of the treated and $J^* \subseteq J$ of the untreated which are, on average, the same with regard to all confounders:

$$E[Y_{i^*,0}] = E[Y_{j^*,0}], \text{ where } i^* \in I^*,\ j^* \in J^*. \tag{3.5}$$

The causal effect on people in these subsets can then be estimated by comparing their means:

$$E[Y_{i^*,1}] - E[Y_{j^*,0}], \text{ where } i^* \in I^*,\ j^* \in J^*. \tag{3.6}$$

If $I^* = I$, this estimates the MTE. If $I^* \subset I$, this is the causal effect on those individuals in the subsets which the quasi-experimental method has discerned. This causal effect will subsequently be referred to as the generalized local average treatment effect (GLATE).[12] It is

[11] See (Angrist [1990]) for an example of instrumental variable estimation, (Card and Krueger [1993]) for difference-in-differences, (Angrist and Lavy [1999]) for regression discontinuity design and (Jackson et al. [2012]) for propensity score matching.

[12] The local average treatment effect (LATE) is a treatment effect estimate specific to instrumental variable estimation, first introduced in (Imbens and Angrist [1994]). LATE is 'local' or dependent on the choice of instrumental variable because different instruments affect different subpopulations. While other quasi-experimental methods, such as the ones mentioned above, require other formal assumptions, they also discern an average effect

important to note that none of the above requires the members of each subset $i^* \in I^*$, $j^* \in J^*$ to be known. Rather, all that is needed is a good argument for why the subsets should be equivalent with regard to all confounders (see equation 3.5) and the ability to determine the observed mean outcomes for each subset (see equation 3.6).

To illustrate, Angrist ([1990]) uses an instrumental variable design to estimate the effect of military service in the Vietnam War on civilian earnings of American veterans in the early 1980s. In the early 1970s, young men born between 1950 and 1953 were conscripted using draft lotteries. For these draft lotteries, every day of the year was randomly assigned a Random Selection Number (RSN) from 1 to 365 and all 19- and 20-year-old men who were born on days with RSNs below a certain threshold number were eligible to be drafted.[13] Given a certain set of informally justified assumptions, the average effect of conscription on those who only served because of the lottery can be estimated because the confounding properties of this subset of the drafted is, on average, the same as those who were not draft-eligible and did not serve.[14]

3.3 The external validity objection

Heckman ([2008]), Heckman and Urzua ([2009]), Deaton ([2010]), Deaton and Cartwright ([2017]) argue that the above quasi-experimental approach is of little interest to policy evaluation because it cannot be used to address many important policy evaluation problems. This is justified by the approach's alleged exclusive focus on 'narrow' local average treatment effects.

To this end, the critics of the approach contend that knowledge of the causal effect on specific subsets of the population is not sufficient for addressing many policy evaluation problems. Further, they hold that extrapolation from the local average treatment estimates of quasi-experiments to 'broader' causal effects for larger populations is not possible because the

'local' to the discerned subpopulations I^* and J^*. Thus, GLATE is an informal generalisation of LATE from IV estimation to the underlying quasi-experimental approach.

[13] The RSN lottery was only the first step in the conscription process and was followed by additional screening steps, like physical examinations.

[14] See (Imbens and Angrist [1994]) for the required assumptions in the case of instrumental variable estimation.

approach lacks 'external validity'. Therefore, the critics conclude that the quasi-experimental approach cannot be used to address policy evaluation problems which concern more than small, quasi-experimentally discernible subpopulations.

3.3.1 'Narrow' and 'broad' causal effects

In the previous section, it has been shown that quasi-experimental methods generally only estimate the GLATE. Although it is theoretically possible for quasi-experiments to estimate the MTE, this is unlikely in practice as it would require the quasi-experimental discerning of a subset of the untreated $J^* \subseteq J$ which is on average equal to the whole population of the treated $I = I^*$. The GLATE is a 'narrow' causal effect because it is not the causal effect of the policy on every possible individual but only the effect the policy had on those treated individuals $i^* \in I^*$ who were discerned by the applied quasi-experimental method. Without extrapolation, the GLATE can therefore only be used in the evaluation of a policy's effects on the discerned subset of the population I^*.

For example, Angrist ([1990]) provides the causal effect of the involuntary military service on 'reluctant' conscripts, that is, on those conscripts who would not have served without the draft lottery. On its own, this causal effect has little implications for the evaluation of any realistic policy because, in practice, it is very difficult to distinguish between 'reluctant' and 'willing' conscripts. Moreover, even if we were able to make the distinction between 'reluctant' and 'willing', (Angrist [1990])'s estimate would only be of interest if policymakers wanted to design a compensation scheme for 'reluctant' veterans of the Vietnam War.

Like in this example, these 'narrow' effects of a policy may address a small subset of the relevant policy evaluation problems, but, oftentimes, they will not shed light on the policy's effects on larger subpopulations. For instance, (Angrist [1990])'s estimate is not the effect of conscription on Vietnam conscripts as a whole because (Angrist [1990])'s instrumental variable estimation does not consider the outcomes of 'willing' conscripts. Even if the latter group's outcomes happened to be numerically identical to the outcomes of the 'reluctant' conscripts and, thus, they happened to make (Angrist [1990])'s estimate numerically correct for all conscripts, this would be a mere 'accident of the numbers' (Cartwright and Munro

[2010], p. 261).

Therefore, on its own, the estimate can neither be used to evaluate the causal effect of conscription in the Vietnam War on the civilian earnings of conscripts nor can it evaluate the causal effect of military service on the civilian earnings of all Vietnam veterans. Yet both these causal effects would be important if policymakers wanted to adequately compensate all conscripts or all veterans for their service to their country. These effects would also be relevant if policymakers wanted to evaluate the long-run effect which conscription has had on aggregate demand, compared to voluntary military service.

Moreover, without further assumptions, these 'narrow' effects only allow for the retrospective evaluation of policies because these effects describe the consequences of an already implemented policy on a specific subpopulation I^*.[15] Here, too, although the quasi-experimental approach evaluation of implemented policies can be useful, the evaluation of prospective policies is also important. It is important to be able to adequately compensate the conscripts of the Vietnam War in hindsight but it would be even better if the policymakers knew the causal effect of conscription on the civilian earnings of prospective conscripts before a new war starts. Unless the latter is known, the design of future conscription compensation schemes is void of empirical evidence until after the conscripts' post-war civilian incomes are measured.

3.3.2 Extrapolation and external validity

The results of quasi-experiments are only limited in the above way, if additional assumptions cannot be used to extrapolate the 'narrow' results to the 'broader' population. However, in many cases, such extrapolation seems natural. Indeed, Angrist ([1990]) does just this and claims to have 'measured the long-term consequences of military service during the Vietnam era' (Angrist [1990], p. 313) — that is, of all military service.

The implicit assumption here seems to be that, if conscription had an effect on civilian earnings, then this must have been exclusively due to it involving military service and, therefore, any effect which conscripts have suffered will have, on average, been suffered in the

[15] Compare (Heckman [2008])'s three problems of policy evaluation. For a more accessible summary of (Heckman [2008])'s discussion, see (Reiss [2015], p. 378.)

same way by voluntary servicemen. In other words, the 'narrow' effect on 'reluctant' conscripts is extrapolated to the whole population based on the assumption that, on average, all sufficiently large subsets of the population are impacted the same way by military service. This assumption that all sufficiently large subpopulations are on average affected equally by a policy will here be called the assumption of causal homogeneity.[16] Formally, this can be expressed as the average causal effect on an individual in the quasi-experimentally discerned subpopulation $\kappa^* \in I^* \cup J^*$ being equal to the average causal effect on an individual in a non-discerned subpopulation $\gamma \in C$, where $C \subseteq \{I \cup J\} \setminus \{I^* \cup J^*\}$:

$$E[Y_{\kappa^*,1} - Y_{\kappa^*,0}] = E[Y_{\gamma,1} - Y_{\gamma,0}] \tag{3.7}$$

Note that this causal homogeneity does not require that the expected outcomes $E[Y_{i,x_i}]$ need to be equal within or across subpopulations $I^* \cup J^*$ and C. Rather, the assumption is a statement about the true average causal effect of the policy on different subpopulations. As stated above, in the case of (Angrist [1990]), this is the assumption that military service had the same causal effect on the 'reluctant' conscripts, which are discerned by the instrumental variable, as it had on those who would have volunteered for military service anyway.

Cartwright ([2007]) argues that this assumption is not supported by quasi-experiments themselves because they are so-called 'clinchers'.[17] 'Clinchers' are deductive estimation methods for causal effects. The truth of these methods' estimates is guaranteed '*if* they [the 'clinchers'] are correctly applied *and* their assumptions are met' (Cartwright [2007], p. 12).

[16] The condition that the subpopulations be 'sufficiently large' is necessary to distinguish this assumption from the untenable assumption that every individual is affected identically. Subpopulations are thus 'sufficiently large' when the averages of their outcomes no longer reflect idiosyncratic differences between individuals.

[17] Cartwright ([2007]) never explicitly states this because her discussion focuses on randomized controlled trials (RCT). However, she mentions that this label also applies to 'certain econometric methods' (Cartwright [2007], p. 12) and Deaton ([2010]) and Deaton and Cartwright ([2017]) criticize both RCTs and instrumental variable estimation in this way. Furthermore, it can be shown that the results of RCTs can be interpreted as instrumental variable estimations with randomization as instrument, see (Heckman [1995]). For an independent argument for instrumental variable estimation being a 'clincher' method, see (Reiss [2007], pp. 126–145). In this paper, this criticism is generalized to all quasi-experimental methods.

In (Angrist [1990]), these assumptions are the informally justified assumptions which allow the application of the draft lottery as an instrumental variable. As long as these assumptions are sufficiently well justified to believe their truth, one is also justified to believe the instrumental variable estimate to be true. However, given that the assumptions' justifications only apply to the subpopulations I^* and J^*, these assumptions may not hold for other subpopulations and, for these, the same causal inference cannot be made.

Further, the discerned subpopulations I^* and J^* are often not only equal with respect to all confounders, but they also intuitively seem more similar to each other than to other subpopulations in other ways. This provides *prima facie* evidence for potential causal heterogeneity between the discerned I^* and J^* and other subpopulations. To illustrate, in (Angrist [1990]), it seems intuitively plausible that conscription could have a different effect on the 'reluctant' conscripts than on those conscripts who would have been interested in serving voluntarily because of, for example, their personality traits. Hypothetically, involuntary military service could have additional psychological effects which may then have a greater impact on 'reluctant' conscripts' prospects in the labour market.

Thus, in the absence of support for the homogeneity assumption and the presence of such intuitions for causal heterogeneity, the results of quasi-experiments cannot be extrapolated to other subpopulations or other populations. Put differently, the estimates lack external validity. Without such external validity, the quasi-experimental approach remains restricted to 'narrow' effects and cannot be applied to the evaluation of policies with 'broad' effects or unimplemented policy options.

3.4 Replication and induction

What this external validity objection fails to consider is that quasi-experiments can and are used to support assumptions of causal homogeneity. While it is true that a single quasi-experiment is insufficient to justify such assumptions, a multitude of consistent quasi-experimental results may provide ampliative evidence for causal homogeneity across the whole population. Even though the differences across subpopulations may still be seen as *prima facie* evidence for causal heterogeneity, quasi-experiments can provide sufficient

evidence to establish defeasible, but strong support for causal homogeneity assumptions. In the context of medical RCTs, this response to the objection has been developed by Backmann ([2017]); Angrist and Pischke ([2010], pp. 22–5) outline a similar argument in less detail.

If a number M of quasi-experimental methods are used to estimate the causal effect of a policy, then they measure the policy's effect for a series of pairs of subpopulations I_k^* and J_k^*, each corresponding to a quasi-experiment k. Even if the estimates for all these subpopulations are very similar, the average causal effect for the whole population cannot be deductively inferred from the results of the quasi-experiments, unless the union of these subpopulations $\bigcup_{k=1}^{M}\{I_k^* \cup J_k^*\}$ is equal to the whole population I. As long as there remains a subpopulation for which the effect of the policy has not been estimated, it is still possible for this remaining subpopulation to be impacted differently by the policy. Put differently, unless $\bigcup_{k=1}^{M} J_k^* = I$, the similarity of the quasi-experimental results could still be a mere coincidence, not due to causal homogeneity across the population.

However, each successful replication of a causal estimate for a different subpopulation offers some inductive evidence that some assumption of causal homogeneity is justified. Each such replication gives more support to the conclusion that the causal effect does probably not only hold for the discerned subpopulation, I_k^* and J_k^*. For this purpose, different replications may offer support for other homogeneity assumptions. For instance, on the one hand, Angrist and Krueger ([1994]) show that conscription in World War II had a similar negative effect on conscripts' earnings as in (Angrist [1990]). Similarly, Imbens and van der Klaauw ([1995]) establish the same result for conscription in the Netherlands. Given that these papers all concern themselves with conscription but vary with regard to the country and time period, their congruent results offer some support for the homogeneity assumption that the average causal effect of conscription is negative across countries and across time. This would allow for extrapolation of the quasi-experimental result to future cases of conscription and, thus, overcomes the external validity objection, including Heckman ([2008])'s criticism that quasi-experiments can only evaluate retrospectively.

On the other hand, Angrist ([1998]) estimates the effect of voluntary military service quasi-experimentally and finds a moderately positive impact. This undermines any

assumption about the causal homogeneity between conscription and voluntary military service. Furthermore, Angrist ([1998]) employs three different quasi-experimental designs, each discerning different subpopulations, which makes it more likely that this positive causal effect of voluntary military service is shared by most or all subpopulations of the population in question. Similarly, Angrist et al ([2010]) use multiple quasi-experimental estimation methods to explicitly address 'concerns about the external validity of IV estimates' (Angrist et al [2010], p. 776).

It is true that, unlike the 'narrow' estimates on which they are based, these more applicable inductive extrapolations are not indefeasible, even if all the made assumptions are true. However, Cartwright, Deaton and Heckman are wrong to assume that this sacrifice of indefeasibility undermines the advantage which quasi-experiments have over traditional econometric models.

In section 2.3, it has been argued that the main advantage which quasi- experiments have over structural models is that the former do not rely on comprehensive models of whole causal systems. This is an advantage because such explicit models are unlikely to capture all causally relevant factors because some of these factors, like personality traits, will likely be unobserved. Because structural models rely on observed data, the causally relevant unobserved variables are either ignored or modelled using observed variables. If they are ignored, the model will only estimate the correct MTE by accident, if the effects of unobserved confounders happen to cancel each other out. If the relevant unobserved variables are modelled using observed variables, these models will be based on economic or behavioural theories.

Whether one believes such models to be more or less credible than quasi- experiments depends on one's trust in the ability of economists to comprehensively list all confounding factors and their theories' ability to accurately model each unobserved confounding factor using the available data.[18] A comprehensive list of confounding factors is unlikely to be achieved. But, even if this is ignored, current economic theory is often based on intuition and simplifying assumptions which are either not justified by empirical data or outright refuted by

[18]Even the proponents of the quasi-experimental approach admit this, see (Angrist and Pischke [2010], p. 22).

it. For instance, much of economics relies on expected utility theory whose assumptions have been widely criticized, even by prominent economists.[19] While the assumptions which allow the extrapolation of 'narrow' quasi-experimental estimates are based on defeasible inductive inferences, these inferences are at least built on deductively inferred 'narrow' causal estimates for subpopulations. Provided the 'narrow' estimates have been replicated, these extrapolations may therefore still be more credible than the 'broad' theory-based estimates of the traditional approach.

Therefore, replication and inductive inferences offer the quasi-experimental approach a path to overcome concerns about the external validity of its estimates. However, the next section will argue that there is a more fundamental concern related to the conceptual problem and SUTVA.

4 The Conceptual Problem

The conceptual problem arises in the definition of the counterfactual outcome $Y_{i,0}$. So far, counterfactual outcomes have only been defined as the outcomes which would occur if a treated individual i was not subjected to the policy.[20] The conceptual problem is that this definition does not yield a unique counterfactual value because there are many vastly dissimilar counterfactual scenarios in which i remains untreated.

To illustrate, there are counterfactual scenarios ('A-scenarios') in which veteran υ did not serve because the expansion of the military has never been implemented. Alternatively, there are many different counterfactual scenarios ('B-scenarios') in which the expansion occurred but υ did not join, for reasons unique to υ.

To represent these differences formally, it is necessary to expand the notation by defining the outcomes with respect to possible worlds — or hypotheticals — $\omega^k \in \Omega$, where Ω is the set of all possible worlds. Let ω^a be the actual world. Consequently, counterfactual worlds are defined as worlds ω^k for all $k \neq a$. The outcome of individual i in world ω^h will be written as Y^h_{i,x_i^h} and i's policy status will be $X_i^h = x_i^h$.

[19] See, for example, (Rabin and Thaler [2001]).

[20] In the following, the focus will, for simplicity, be on those individuals who were actually subjected to the policy — that is, the treated. In the above example, this is the veteran υ.

The two sets of counterfactual outcomes for υ can then be expressed as follows:

'A-scenario' outcomes = $\{Y^k_{\upsilon,0} : k \neq a \text{ and } X^k_i = 0 \text{ for } \forall i\}$

'B-scenario' outcomes = $\{Y^k_{\upsilon,0} : k \neq a \text{ and } X^k_i = 1 \text{ for } \forall i \text{ s.t. } X^a_i = 1 \text{ and } i \neq \upsilon\}$

However, the definition of the individual treatment effect requires that a choice be made between these different counterfactual outcomes because the effect is only uniquely determined with a unique value for the counterfactual outcome $Y^k_{i,0}$.

Thus, the conceptual problem necessitates a more substantial definition of the hypothetical which yields $Y^k_{i,0}$.

4.1 Counterfactual interventions and SUTVA

One such definition can be obtained by appeal to counterfactual 'interventions'.[21] Following the interventionist framework of causation, the Rubin's model of causal inferences can be extended thus:

Instead of simply defining the causal effect in terms of the counterfactual outcome after a change of policy status of one individual x_i, it will be defined in terms of the outcome which would have occurred if there had been an exogenous intervention which changed the policy status of any number of individuals.[22]

Formally, let $n \in \mathbb{N}$ be the number of individuals in the population and define the 'policy assignment' $\mathbf{P^h}$ in world ω^h as the n-dimensional vector of all individuals' policy statuses x^h_i:

$$\mathbf{P^h} = (x^h_1, x^h_2, ..., x^h_{n-1}, x^h_n) \tag{4.1}$$

With this notation, the causal effect from equation 3.1 can be restated as the difference

[21] A different solution would be the use of Lewis' framework of counterfactual causation, see (Lewis [1973]; [1979]). Indeed, Heckman ([2008]) references this approach. However, the problem with this framework is that it introduces the concept of 'closeness' between hypotheticals but fails to provide a measure of this 'closeness', see (Heckman [2008], p. 9).

[22] This paper limits itself to the elements of the interventionist framework of causation which are required for the purposes of this discussion. For a general and in-depth treatment, the reader is referred to (Pearl [2009]) or (Woodward [1997]). Alternatively, for a more succinct but still insightful introduction, see (Hitchcock 2001).

between the average outcomes Y_i given the actual 'policy assignment' $\mathbf{P^a}$, where $X_i^a = 1$, and under the counterfactual 'policy assignment' $\mathbf{P^k}$, where $X_i^k = 0$:

$$E[Y_i \mid \mathbf{P^a}] - E[Y_i \mid \mathbf{P^k}], \ \textit{where } X_i^a = 1 \textit{ and } X_i^k = 0. \tag{4.2}$$

The causal effect of the policy is thus not simply the effect of the policy as it was actually assigned. As noted by Holland ([1986]), 'the effect of a cause is always relative to another cause' (Holland [1986], p. 946) and, in this case, the other cause is the baseline of a counterfactual policy assignment $\mathbf{P^k}$. In other words, the effect of the policy is the causal effect of the change from a hypothetical policy assignment $\mathbf{P^k}$ to the actual policy assignment $\mathbf{P^a}$. In the parlance of the interventionist framework of causation, the causal effect of Rubin's model thus completes the consequent of the interventionist counterfactual 'If $\mathbf{P^k}$ was changed to $\mathbf{P^a}$, then Y_i would change by . . .'.

Further, the counterfactual 'policy assignment' $\mathbf{P^k}$ is not simply the 'policy assignment' which would have actually occurred if policymakers had not enacted the policy. Rather, which counterfactual 'policy assignment' $\mathbf{P^k}$ is the right baseline depends on the policy evaluation question at hand.

To illustrate, if a policymaker wanted to compensate a veteran υ for the financial losses their service status has caused them individually, then the appropriate counterfactual 'policy assignment' would be one under which veteran υ did not serve but the military expansion still took place, that is the 'policy assignment' $\mathbf{P^B}$ of 'B-scenarios'. On the other hand, if the evaluation question asked for the impact of these financial losses on aggregate demand because the policymaker wanted to estimate the effect on aggregate demand of a future military expansion, then the actual 'policy assignment' would be contrasted with a 'policy assignment' under which the military expansion had not taken place, that is the 'policy assignment' $\mathbf{P^A}$ of 'A-scenarios'.

The above examples illustrate how the same policy, in this case the military expansion, can have different effects depending on which counterfactual 'policy assignment' $\mathbf{P^k}$ it is contrasted with. And this is more than a mere interpretative difference — the magnitude and size of the causal effect may be contingent on this choice of counterfactual 'policy

assignment' $\mathbf{P^k}$.

This is due to social interaction effects, such as general equilibrium effects. Social interaction effects are the changes in individual i's outcome Y^h_{i,x_i} which are indirectly caused by the treatment statuses X^h_j of other individuals via the behavioural changes caused by these other individuals' treatment statuses.

For instance, the counterfactual outcomes of veteran υ in 'A-scenarios' and 'B-scenarios' may differ considerably. In 'A-scenarios', υ may have needed to compete with more young people for civilian jobs than she would have in 'B-scenarios', because, without the military expansion, more young people would remain on the civilian job market. This increased competition could mean that, on average, υ remains unemployed for longer in 'A-scenarios' than υ in 'B-scenarios' in which there is less competition on the civilian labour market. If this was the case, then, ceteris paribus, υ's civilian earnings under the 'A-scenario' assignment $\mathbf{P^A}$ would be lower than her counterfactual income under the 'B-scenario' assignment $\mathbf{P^B}$. Consequently, the causal effect of the service status would be more positive or less negative when using $\mathbf{P^A}$ as the baseline of comparison, rather than $\mathbf{P^B}$, because the actual earnings Y^a_υ would be compared to the lower 'A-scenario' earnings.

Although more than this informal labour market story would be required to determine whether the social interaction effects are, in aggregate, positive or negative in this case, it is clear that the estimated causal effect depends on the choice of policy question and, thus, of counterfactual intervention.

In the econometric literature, this conceptual problem is only rarely explicitly discussed.[23] If social interaction effects are mentioned, the stable unit treatment value assumption (SUTVA) is commonly invoked to assume away such effects.[24] This assumption postulates that, across hypotheticals, the outcome of any individual Y_i is invariant to changes in the treatment values of other individuals X_j across hypotheticals. Using the present notation, this

[23]For example, (Angrist [1991]; Imbens and Angrist [1994]; Angrist and Pischke [2009]) do not mention general equilibrium effects or other types of social interactions.

[24]SUTVA was first named in (Rubin [1986]). For statistical literature on SUTVA, see (Rubin [1978], [1980]; Holland [1986]). For econometric literature which expresses a cautious stance on the assumption, see (Heckman [2005], assumption (A-1), p. 11, [2008]; Heckman and Smith [1998], assumption A-1, p. 12; Imbens [2010], p. 401, [2014], footnote 3; Athey and Imbens [2017], p. 19–23).

can be expressed formally as:

$$Y^{h}_{i,x_i} = Y^{l}_{i,x_i} \text{ for any } \omega^h, \omega^l \text{ which differ only by } \bigcup_{j \neq i}^{n-1} \{X^h_j = 1\} \neq \bigcup_{j \neq i}^{n-1} \{X^l_j = 1\}. \tag{4.3}$$

In the remainder of this paper, this assumption will be argued to be implausible and that this poses a severe problem to the quasi-experimental approach.

4.2 The social interaction objection

Like randomized controlled trials, quasi-experiments are based on the assumption that the observed average outcome $E[Y^{a}_{j^*,0}]$ of the untreated is equal to the counterfactual average outcome $E[Y^{k}_{i^*,0}]$ of the treated if the treated had not been treated. It has been shown above that it is not clear in terms of which counterfactual scenario ω^k — and, thus, 'policy assignment' $\mathbf{P^k}$ — the average counterfactual outcome $E[Y^{k}_{i^*,0}]$ is defined and that the appropriate definition of the counterfactual intervention depends on the exact policy question which is to be answered. Further, a different definition of the counterfactual 'intervention' could lead to different estimates because of general equilibrium effects and other social interaction effects.

In the following sections of this paper, it will be argued that SUTVA, which is standardly invoked to avoid clearly defining a counterfactual, is implausible in the presence of social interaction effects. The novel stable marginal unit treatment assumption (SMUTVA) will be proposed as the foundation of a more plausible interpretation of quasi-experimental estimates in the presence of social interaction effects. However, adoption of SMUTVA will be shown to severely limit the type of policy evaluation questions which quasi-experiments can address. In analogy to section 3.4, an inductive attempt to overcome these restrictions will be considered. Finally, the conclusion will be that the applicability of quasi-experiments as a standalone approach to causal estimation remains severely limited, although an auxiliary role in economic policy evaluation remains a possibility.

4.2.1 From SUTVA to SMUTVA

For the evaluation of many types of interventions, SUTVA is a sensible assumption. In the case of most medical treatments, it is reasonable to assume that the health outcomes of any individual receiving a treatment will not be affected by the number of other people who receive the same treatment. This is the case because, for most health outcomes, the only significant causal mechanisms are those which directly affect the individual who receives the treatment, such as the bio-mechanical mechanisms of pharmaceuticals. The effect of other individuals' health outcomes are mostly negligible.

One exception might intuitively be vaccines for contagious diseases. To illustrate, imagine an experimental vaccine which reduces by 50% the chance of severe symptoms and contagion if infected by a specific virus, but does not completely eliminate the risk. In this case, if a sufficiently high proportion of the population within a certain region receives the vaccine, this will have a significant impact on the health outcome of any individual receiving the vaccine because the higher the number of vaccinated individuals the lower the risk of any individual being infected.

Unless it can be ensured that all individuals face the same risk of being infected, the health outcomes of any vaccinated individual could thus be affected by the number of other subjects receiving the vaccine.[25] In this case, SUTVA is no longer reasonable because it assumes away just this social effect of the number of treated people.

By the same reasoning, SUTVA is rarely realistic for the evaluation of economic or social policy. Economic and other social outcomes are typically affected by the number of treated via social interaction effects because economic outcomes are, by their nature, the result of social interactions between economic agents. A change in the behaviour of the members of a sufficiently large subpopulation will typically have a non-negligible effect on the behavioural responses and outcomes of the other participants in these social interactions.

For instance, conscription during the Vietnam War may have decreased the supply of young

[25] In practice, randomized controlled trials for vaccines are often designed to equalize the risk of infection across treatment and control groups. This is achieved by controlled infection of all subjects. For a review of such trials for influenza vaccines, see (Balasingam and Wilder-Smith [2016]). As described in section 3, this extent of control over a policy's subject is not common in economics or other social sciences.

men in the US economy, thus increasing the chances for non-conscripts to gain civilian work experience during the war. Alternatively, in developmental economics, the provision of income support by an nongovernmental organisation in one village may cause the local government to divert resources to other towns, causing a so-called spillover effect.[26] Similarly, in education, the reduction of class sizes in some schools may reduce the number of good teachers available to other schools in the area. This could diminish the quality of teachers at the remaining schools and, hence, their educational outcomes.[27]

Therefore, unless the evaluated policy only concerns a minuscule or completely isolated subpopulation, the observed average outcome of those who are actually untreated $E[Y^a_{j^*,0}]$ cannot be assumed equal to the counterfactual average outcome $E[Y^k_{i^*,0} \mid X^k_i = 0 \text{ for } \forall i \in I]$ which the treated would have experienced if no one had been treated.

This is not a problem for the structural approach because this approach does not directly equate an observed outcome with a counterfactual outcome. Instead, the observed outcome is only an input into a model which yields the counterfactual outcome. The model can be adjusted to account for interaction effects, such as general equilibrium or spillover effects.

However, the quasi-experimental approach directly equates an observed outcome $E[Y^a_{j^*,0}]$ with a counterfactual outcome $E[Y^k_{i^*,0} \mid X^k_i = 0 \text{ for } \forall i \in I]$. This is problematic because the observed outcome is influenced by social interactions effects which would not have been present in the hypothetical ω^k which yields the relevant counterfactual outcome.

The solution to this problem is to replace SUTVA with a more plausible assumption which only allows for equating observed and counterfactual outcomes if both are influenced by sufficiently similar social interaction effects. In most circumstances, this is reasonable if two scenarios only differ with regard to a small subpopulation's or individual's treatment status because most markets in which the economic outcomes are determined can be assumed to be large enough for differences in the behaviour of a marginal number of participants to have a negligible effect on the market's outcomes.

To illustrate, the observed average outcome of the untreated $E[Y^a_{j^*,0}]$ can reasonably be

[26] For concerns about spillover effects in developmental economics, see (Rodrik [2009], p. 19; Ravallion [2009], p. 53).
[27] See (Sims [2010], pp. 65–7).

assumed to be equal to the counterfactual average outcome
$E[Y^k_{i^*,0} \mid X^k_{i^*} = 0 \text{ and } X^k_i = 1 \text{ for } \forall i \in I^*, \text{ where } i \neq i^*]$ which the treated would have experienced if they had remained untreated in small groups or as individuals while the policy was implemented for everyone else. In the example of the military expansion, this would be the assumption that in 'B-scenarios' — that is, if any individual veteran or any small group had not served but the military expansion had still occurred — the veterans would have, on average, earned the same as what actual non-veterans earned. This is a sensible assumption as long as the regional civilian labour markets are large enough for such a change in the number of participants to not matter.

This assumption will be called the stable marginal unit treatment value assumption (SMUTVA) and can be expressed as follows:

$$Y^h_{i,x_i} = Y^l_{i,x_i} \text{ for any two worlds } \omega^h, \omega^l \text{ which at most differ by } \bigcup_{j\neq i}^{\epsilon}\{X^h_j = 1\} \neq \bigcup_{j\neq i}^{\epsilon}\{X^l_j = 1\} \text{ for a small enough } \epsilon \in \mathbb{N}. \tag{4.4}$$

Here the size of ϵ depends on the number and size of participants in the markets in which the relevant outcome Y is determined.

4.2.2 The implications of SMUTVA

However, based on SMUTVA, the quasi-experiment's estimate of the causal effect from equation 3.6 becomes the policy's direct or marginal causal effect:

$$E[Y^a_{i^*,1}] - E[Y^k_{i^*,0} \mid X^k_{i^*} = 0,\ X^k_i = 1 \text{ for } \forall i \in I^*, \text{ where } i \neq i^*] \tag{4.5}$$

This more restrictive definition has practical implications which are similarly limiting to the approach as a lack of external validity objection would have been. Defined in this way, the causal effect only allows for the economic evaluation of small policies which affect so few people that they do not cause social interaction effects. This excludes most interesting policy evaluation problems, such as most of the examples in this paper. For large-scale policies, quasi-experiments can be used to estimate the direct effect of the policy on the marginal

individual or small group. Such estimates could then be used as input into structural econometric models, but are not sufficient to evaluate large-scale policies.[28]

For example, Angrist [1990]'s estimate needs to be interpreted as the effect of conscription on veteran υ, given that conscription was implemented for everyone else. On the basis of this marginal estimate, a compensation scheme could be designed if the intention of this scheme was to compensate conscripts only for being drafted, and not for the introduction of conscription. Hypothetically, if the social interaction effects of conscription on non-conscripts were in aggregate positive, such a compensation scheme would compensate the conscripts by less than a scheme which also compensated for the introduction of conscription. This is because if the conscripts had not been drafted, their counterfactual civilian earnings as a non-conscript would have been increased by the social interaction effects. However, if a country's policymakers were contemplating whether or not to introduce conscription and there were concerns about the long-term impact of such a policy on aggregate demand, (Angrist [1990])'s estimate could not be directly used to address these concerns.

If the goal is to evaluate more than the effect of minuscule policies or of the addition of individuals to a large policy's population of subjects, the inclusion of social interaction effects in causal estimation is crucial. The quasi-experimental approach is unable to do so on its own because it is based on the comparison of observed outcomes which, at least partially, are themselves affected by these social interaction effects.[29]

4.3 Induction again?

Although both this argument and the external validity objection conclude that the applicability of the quasi-experimental approach is restricted to relatively unappealing policy evaluation problems, the response from section 3 does not transfer to this section. Replication and consequent induction are not a feasible response to the social interaction objection.

[28] For an econometric model which attempts to supplement treatment effect estimator by a simple general equilibrium model, see (Heckman and Smith [1998]).

[29] In principle, quasi-experimental estimates could be adjusted by social interaction effects either through formal models, such as noted above, or by informal argument. Note, however, that informal arguments are highly unlikely to establish the exact sign and size of non-zero aggregate social interaction effects because of the variety of interactions involved.

The difference between these two objections is that, while the external validity objection contends that quasi-experiments' 'narrow effects' cannot be extrapolated, the conceptual problem which the indirect social interaction objection is based on reveals that, even for the discerned subpopulations I^* and J^*, quasi-experiments fail to estimate the overall causal effect of the policy. Instead, their estimates' validity are themselves limited to marginal effects.

This is unsurprising if we recall that replication and induction were meant to support causal homogeneity assumptions, such as equation 3.7. While the truth of such assumptions would overcome the external validity objection, it is obvious that such assumptions do not help against the social interaction objection. The latter objection would hold even if the average causal effect of a policy were the same for every significant subset of the population because such homogeneity would not alter the omission of any potential social interaction effect from quasi-experimental estimation methods.[30]

In other words, replication and induction cannot overcome the social interaction objection because all quasi-experimental replications are identically ignorant of social interaction effects. Unlike in section 3.4, replication cannot provide inductive evidence about social interaction effects because quasi-experimental replications of the original estimate in different populations cannot reflect variations in social interaction effects.

For example, if a policy was implemented at a small enough scale to assume that no significant social interaction effects are present, a quasi-experiment's estimate of the marginal effect of the policy may be assumed to be equal to the overall effect in this setting. Even if this estimate is replicated by a quasi-experiment on a larger-scaled version of the same policy, this does not show that the overall effect is the same but only that the marginal effect is constant. All that is established by such scale-invariant effect sizes is the scale invariance of the policy's marginal effect.

[30]Once again, 'significant' here is meant to indicate that considered subpopulations need to be large enough to 'average out' idiosyncratic differences in the causal effects on individuals.

5 Conclusion

To summarize, the quasi-experimental approach can be used to evaluate the economic consequences of small-scaled policies with negligible social interaction effects. For such policies, quasi-experiments cannot only estimate the causal effects of an implemented policy on a small subpopulation, but also provide support for the inductive extrapolation of these estimates to implementations of the same small-scaled policy in other populations. Thus, such inductive extrapolation extends the approach's applicability beyond the retrospective evaluation of policies.

Yet, for policies whose social interaction effects are non-negligible, quasi-experiments' usefulness is limited. Given SMUTVA, they can only be plausibly used to estimate the direct causal effects of such policies on marginal subjects. By itself, the quasi-experimental approach cannot estimate the overall economic effects of large-scale policies, such as conscription's effect on aggregate demand or minimum wage increases' effect on employment. In the evaluation of such large-scale policies, quasi-experiments could still be used as input for more traditional econometric models, whose theoretical assumptions allow the estimation of social interaction effects.

Acknowledgements

I owe my deepest gratitude to Johanna Thoma, London School of Economics. Without her encouraging supervision and practical advice, this paper would not have been possible. This paper has also greatly benefited from the comments of the two anonymous referees. Their constructive comments helped with tying up the intellectual loose ends of the paper's argument. Finally, my special thanks go to Sarasvati Spaur for her patience and enthusiasm.

Tom Boesche

Philosophy, Logic and Scientific Method

London School of Economics

Houghton Street, London, WC2A 2AE

t.boesche@lse.ac.uk

References

Angrist, J. [1990]: 'Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records', *The American Economic Review*, **80(3)**: pp. 313-336.

Angrist, J. [1991]: 'Instrumental Variable Estimation of Average Treatment Effect in Econometrics and Epidemiology', *NBER Technical Working Paper No. 115*.

Angrist, J. [1998]: 'Estimating the Labor Market Impact of Voluntary Military Service Using Social Security', *Econometrica*, **66(2)**: pp. 249-288.

Angrist, J. and Lavy, V. [1999]: 'Using Maimonides' rule to estimate the effect of class size on scholastic achievement', *The Quarterly Journal of Economics*, **114(2)**: pp. 533-575.

Angrist, J., Lavy, V. and Schlosser, A. [2010]: 'Multiple Experiments for the Causal Link between the Quantity and Quality of Children', *Journal of Labor Economics*, **28(4)**: pp. 773-824.

Angrist, J. and Krueger, A. [1994]: 'Why Do World War II Veterans Earn More than Nonveterans?', *Journal of Labor Economics*, **12(1)**: pp. 74-97.

Angrist, J. and Pischke, J. [2009]: *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, NJ: Princeton University Press.

Angrist, J. and Pischke, J. [2010]: 'The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking The Con Out Of Econometrics', *Journal of Economic Perspectives*, **24(2)**: pp. 3-30.

Athey, S. and Imbens, G. [2017]: 'The State of Applied Econometrics: Causality and Policy Evaluation', *Journal of Economic Perspectives*, **31(2)**: pp. 3-32.

Backmann, M. [2017]: 'What's in a gold standard? In defence of randomised controlled trials', *Medicine, Health Care and Philosophy*, **20(4)**: pp. 513-523.

Balasingam, S. and Wilder-Smith, A. [2016]: 'Randomized controlled trials for influenza drugs and vaccines: a review of controlled human infection studies', *International Journal of Infectious Diseases*, **49**: pp. 18-29.

Card, D. and Krueger, A. [1993]: 'Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania', *NBER Working Paper No. 4509*.

Card, D. and Krueger, A. [2000]: 'Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply', *The American Economic Review*, **90(5)**: pp. 1397-1420.

Cartwright, N. [2007]: 'Are RCTs the gold standard?', *BioSocieties*, **2(1)**: pp. 11-20.

Cartwright, N. and Munro, E. [2010]: 'The limitations of randomized controlled trials in predicting effectiveness', *Journal of Evaluation in Clinical Practice*, **16(2)**: pp. 260-266.

Deaton, A. [2010]: 'Instruments, Randomization, and Learning about Development', *Journal of Economic Literature*, **48**: pp. 424-455.

Deaton, A. and Cartwright, N. [2017]: 'Understanding and misunderstanding randomized controlled trials', *NBER Working Paper No. 22595*, revised.

Heckman, J. [2005]: 'The scientific model of causality', *Sociological Methodology*, **35**: pp. 1-97.

Heckman, J. [2008]: 'Econometric Causality', *International Statistical Review / Revue Internationale de Statistique*, **76(1)**: pp. 1-27.

Heckman, J. and Smith, J. [1998]: 'Evaluating the Welfare State', *NBER Working Paper No. 6542*.

Heckman, J. and Urzua, S. [2009]: 'Comparing IV With Structural Models: What IV Can and Cannot Identify', *NBER Working Paper No. 14706*.

Hitchcock, C. [2001]: 'The Intransitivity of Causation Revealed in Equations and Graphs', *The Journal of Philosophy*, **98(6)**: pp. 273-99.

Holland, P. [1986]: 'Statistics and Causal Inference', *Journal of the American Statistical Association*, **81(396)**: pp. 945-960.

Imbens, G. [2010]: 'Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)', *Journal of Economic Literature*, **48**: pp. 399-423.

Imbens, G. [2014]: 'Instrumental Variables: An Econometrician's Perspective', *Statistical Science*, **29(3)**: pp. 323-358.

Imbens, G. and Angrist, J. [1994]: 'Identification and Estimation of Local Average Treatment Effects', *Econometrica*, **62(2)**: pp. 467-475.

Imbens, G. and van der Klaauw, W. [1995]: 'Evaluating the Cost of Conscription in the Netherlands', *Journal of Business & Economic Statistics*, **13(2)**: pp. 207-215.

Judge, T., Higgins, C., Thoresen, C., and Barrick, M. [1999]: 'The Big Five Personality Traits, General Mental Ability, And Career Success Across The Life Span', *Personnel Psychology*, **52**: pp. 621-652.

Jackson, J., Thoemmes, F., Jonkmann, K., LÃijdtke, O., and Trautwein, U. [2012]: 'Military Training and Personality Trait Development: Does the Military Make the Man, or Does the Man Make the Military?', *Psychological Science*, **23(3)**: pp. 270-277.

Lewis, D. [1973]: *Counterfactuals*, Oxford: Blackwell Publishers.

Pearl, J. [2009]: *Causality*, New York: Cambridge University Press.

Ravallion, M. [2009]: 'Comment', in Cohen, J. and Easterly, W. (eds), *What Works In Development? Thinking Big and Thinking Small*, Washington, D.C.: Brookings Institution Press.

Reiss, J. [2015]: 'Two approaches to reasoning from evidence or what econometrics can learn from biomedical research', *Journal of Econometric Methodology*, **22(3)**: pp. 373-390.

Robbins, L. [1932]: 'Chapter 6: The significance of economic science', *An Essay on the Nature and Significance of Economic Science*, London: Macmillan, p. 120-141.

Rodrik, D. [2009]: 'The New Development Economics: We Shall Experiment, but How Shall We Learn?', in Cohen, J. and Easterly, W. (eds), *What Works In Development? Thinking Big and Thinking Small*, Washington, D.C.: Brookings Institution Press.

Rubin, D. [1974]: 'Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies', *Journal of Educational Psychology*, **66**: pp. 688-701.

Rubin, D. [1977]: 'Assignment of Treatment Group on the Basis of a Covariate', *Journal of Educational Statistics*, **2**: pp. 1-26.

Rubin, D. [1978]: 'Bayesian Inference for Causal Effects: The Role of Randomization', *The Annals of Statistics*, **6**: pp. 34-58.

Rubin, D. [1980]: 'Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment', *Journal of the American Statistical Association*, **75**: pp. 591-593.

Rubin, D. [1986]: 'Comment: Which Ifs Have Causal Answers', *Journal of the American Statistical Association*, **8**: pp. 961-962.

Rabin, M. and Thaler, R. [2001]: 'Anomalies: Risk aversion', *The Journal of Economic Perspectives*, **15(1)**: pp. 219-232.

Sims, C. [2010]: 'But Economics Is Not an Experimental Science', *The Journal of Economic Perspectives*, **24(2)**: pp. 59-68.

Woodward, J. [2003]: *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press.